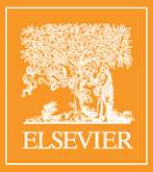


THE PSYCHOLOGY OF LEARNING AND MOTIVATION

volume 40

Douglas L. Medin

THE PSYCHOLOGY OF LEARNING AND MOTIVATION

Advances in Research and Theory

VOLUME 40

This Page Intentionally Left Blank

THE PSYCHOLOGY OF LEARNING AND MOTIVATION

Advances in Research and Theory

EDITED BY DOUGLAS L. MEDIN

DEPARTMENT OF PSYCHOLOGY
NORTHWESTERN UNIVERSITY
EVANSTON, ILLINOIS

Volume 40

ACADEMIC PRESS

A Harcourt Science and Technology Company

San Diego San Francisco New York Boston London Sydney Tokyo

This Page Intentionally Left Blank

CONTENTS

Contributors ix

DIFFERENT ORGANIZATION OF CONCEPTS AND MEANING SYSTEMS IN THE TWO CEREBRAL HEMISPHERES

Dahlia W. Zaidel

I. Introduction 1
II. Concepts, Schemas, and Categories in the Hemispheric Long-Term Memory 3
III. Constraining Influences in the Interaction between Hemispheric Specialization and Long-Term Memory 16
IV. Possible Answers to the Biological Logic of Hemispheric Long-Term Memory 17
V. Conclusions 18
References 19

THE CAUSAL STATUS EFFECT IN CATEGORIZATION: AN OVERVIEW

Woo-kyoung Ahn and Nancy S. Kim

I. Different Approaches to Understanding Feature Centrality 23
II. Causal Status Hypothesis: General Introduction 27
III. Essentialism and the Causal Status Hypothesis 29
IV. Main Empirical Results Supporting the Causal Status Hypothesis 31

V. Related Phenomena Accounted for by the Causal Status Hypothesis: Natural Kinds versus Artifacts 33
VI. Computational Modeling 37
VII. Moderating Factors 38
VIII. Causal versus Isolated Features 49
IX. Application 52
X. Other Types of Dependency Relations 57
XI. Conclusion 61
References 61

REMEMBERING AS A SOCIAL PROCESS

Mary Susan Weldon

I. Introduction 67
II. Arguments for the Social Nature of Memory 68
III. How Can Remembering Be Studied as a Social Process? 77
IV. Collaborative Remembering and Collective Memory 91
V. Conclusions 115
References 116

NEUROCOGNITIVE FOUNDATIONS OF HUMAN MEMORY

Ken A. Paller

I. What Evidence from Neuroscience Is Relevant for Understanding Cognition? 122
II. The Neural Implementation of Declarative Memory 124
III. Electrophysiological Measures of Recollection 131
IV. Electrophysiological Measures of Perceptual Priming 137
V. Conclusions: Cognitive Neuroscience and Human Memory 138
References 140

STRUCTURAL INFLUENCES ON IMPLICIT AND EXPLICIT SEQUENCE LEARNING

Tim Curran, Michael D. Smith, Joseph M. DiFranco, and Aaron T. Daggy

I. Introduction 147
II. Group Differences Revealed by Structural Manipulations 149

III. Effects of Reversal Frequency on Implicit and Explicit Sequence Learning (Experiments 3 to 5) 162
IV. General Discussion 173
V. Conclusions 178
References 178

RECALL PROCESSES IN RECOGNITION MEMORY

Caren M. Rotello

I. Introduction 183
II. Evidence on the Use of Recall in Recognition 184
III. Evidence on the Specific Nature of the Recall Process 186
IV. Factors That Influence the Use of Recall Processes in Recognition 193
V. Other Measures of Recall-to-Reject Processing 194
VI. The Receiver Operating Characteristic Curve 195
VII. Conjoint Recognition 201
VIII. Experiment 204
IX. Conjoint Recognition Estimate of Recall-to-Reject Processing 212
X. Familiarity-Based Explanations of the Data 216
XI. Conclusions 217
References 218

REWARD LEARNING: REINFORCEMENT, INCENTIVES, AND EXPECTATIONS

Kent C. Berridge

I. Reinforcement Theories 223
II. Bases of Modern Incentive Theory: Bolles–Bindra–Toates 234
III. Consequences for the Incentive Model of Motivation and Reward 239
IV. Do Conditioned Incentive Stimuli Take on the Motivational Properties of Primary Incentive Stimuli? 246
V. Splitting Pavlovian Incentives: "Liking" versus "Wanting" 252
VI. From Pavlovian Incentives to Incentive Expectations: Dickinson–Balleine 257
VII. Brain Mechanisms of Reward Learning 264
VIII. Conclusion 271
References 272

SPATIAL DIAGRAMS: KEY INSTRUMENTS IN THE TOOLBOX FOR THOUGHT

Laura R. Novick

I. The Toolbox for Thought 279
II. The Importance of Selecting the Appropriate Representational Tool 281
III. Spatial Diagram Tools 282
IV. Overview of the Chapter 282
V. The Use of Exemplar Knowledge in Choosing Diagrammatic Tools 285
VI. The Existence in Memory of Abstract Tool Schemas 290
VII. A Structural Analysis of the Three Spatial Diagram Tools 296
VIII. Empirical Investigation of the Structural Analysis and Students' Diagram Knowledge 301
IX. The Structure of Students' Knowledge about the Three Spatial Diagrams 312
X. Constructing Spatial Diagrams 317
XI. Discussion 318
References 322

REINFORCEMENT AND PUNISHMENT IN THE PRISONER'S DILEMMA GAME

Howard Rachlin, Jay Brown, and Forest Baker

I. Introduction 327
II. Playing Card Paradigm (Single Trials) 332
III. Playing Card Paradigm (Grouped Trials) 336
IV. Game Board Paradigm (Single Trials) 340
V. Game Board Paradigm (Single Trials, Increased Reward for Cooperation) 345
VI. Game Board Paradigm (Patterned Trials) 347
VII. Computer Paradigm (Probability of Reciprocation, Spinners) 348
VIII. Computer Paradigm (Probability of Reciprocation, No Spinners) 353
IX. Computer Paradigm (Probability of Reciprocation, Other Player) 354
X. Computer Paradigm (Probability of Reinforcement and Punishment Varied Separately) 356
XI. Computer Paradigm (Tit-for-Tat versus Pavlov) 358
XII. Conclusions 362
References 363

Index 365
Contents of Recent Volumes 377

CONTRIBUTORS

Numbers in parentheses indicate the pages on which the authors' contributions begin.

Woo-Kyoung Ahn (23), Department of Psychology, Vanderbilt University, Nashville, Tennessee 37240

Forest Baker (327), Department of Psychology, State University of New York at Stony Brook, Stony Brook, New York 11795

Kent C. Berridge (223), Department of Psychology, University of Michigan, Ann Arbor, Michigan 48109

Jay Brown (327), Department of Psychology, State University of New York at Stony Brook, Stony Brook, New York 11795

Tim Curran (147), Department of Psychology, Case Western Reserve University, Cleveland, Ohio 44106

Aaron T. Daggy (147), Department of Psychology, Case Western Reserve University, Cleveland, Ohio 44106

Joseph M. DiFranco (147), Department of Psychology, Case Western Reserve University, Cleveland, Ohio 44106

Nancy S. Kim (23), Department of Psychology, Yale University, New Haven, Connecticut 06520

Laura R. Novick (279), Department of Psychology, Vanderbilt University, Nashville, Tennessee 37203

Ken A. Paller (121), Department of Psychology and Institute for Neuroscience, Northwestern University, Evanston, Illinois 60208

Howard Rachlin (327), Department of Psychology, State University of New York at Stony Brook, Stony Brook, New York 11795

Caren M. Rotello (183), Department of Psychology, University of Massachusetts, Amherst, Massachusetts 01003

Michael D. Smith (147), Department of Psychology, Case Western Reserve University, Cleveland, Ohio 44106

Mary Susan Weldon (67), Department of Psychology, University of California, Santa Cruz, Santa Cruz, California 95065

Dahlia W. Zaidel (1), Department of Psychology, University of California, Los Angeles, Los Angeles, California 90095

DIFFERENT ORGANIZATION OF CONCEPTS AND MEANING SYSTEMS IN THE TWO CEREBRAL HEMISPHERES

Dahlia W. Zaidel

I. Introduction

The left and right sides of the brain, the cerebral hemispheres, are both anatomically (Galaburda, 1995) and functionally (D. Zaidel, 1993) asymmetrical, a neuroarrangement unique to humans in extent and scope (Sperry, 1974). Its adaptive evolutionary development can explain a great deal about the human experience, including speaking, writing, and reading. My research into brain organization has been driven by the assumption that in order to understand the human mind, as well as personality, emotions, and social behavior, it is critical to gain insight into the logic of hemispheric specialization and functional asymmetry. Thus, my findings emphasized left–right asymmetry in long-term memory (LTM) ascribing storage of common, prototypical, stereotypical, and cultural metaphors to the right hemisphere (RH) and the reverse of that to the left hemisphere (LH), that is, to storage of new, noncultural metaphors, creative, and modified familiar concepts whether reflected in artworks, single objects, or in pictorial semantics (D. Zaidel, 1986, 1987, 1988, 1990a; D. Zaidel & Kasher, 1989).

Ever since hemispheric asymmetry was accepted as a major principle of brain organization, starting in the early 1960s with the fascinating results from 'split-brain' patients, many researchers in neuropsychology focused

THE PSYCHOLOGY OF LEARNING AND MOTIVATION, VOL. 40

Copyright © 2001 by Academic Press.
All rights of reproduction in any form reserved.

0079-7421/00 $35.00

on which mental functions are lateralized to one or the other hemisphere (Bogen & Vogel, 1962a; Bogen & Vogel, 1975; Bogen, Schultz, & Vogel, 1988; Finger, 1994; Sperry, 1974). Earlier, neurologists and neuropsychologists learned about functional asymmetry from patients with unilateral focal brain damage (McCarthy & Warrington, 1990). In such patients, damage in a localized brain region produces a profile of perceptual and cognitive deficits not seen following damage in other cortical regions. Data from split-brain patients confirmed and augmented those findings. The functional laterality picture has not changed much over the years, with the LH still considered to have the major control over language functions, including speech, writing, and comprehension, and the RH to have the major control over facial memory, visuospatial cognition, and musical appreciation (McCarthy & Warrington, 1990; Hiscock & Kinsbourne, 1995). However, various modifications have been added. For example, the fact that the main language centers are in the LH does not mean that the RH cannot extract meaning from sensations or cannot process abstract concepts. A good example is the findings of Cronin-Golomb (1986), who used a completely nonverbal test of picture associations in a study of three complete commissurotomy patients and found 82% correct responses controlled by the RH compared to 90% correct by the LH. So although the RH lacks a highly developed language system, it is still capable of abstract cognitive processing. The associative process need not necessarily be language-mediated in either hemisphere.

Confirmation for the functional specialization of each disconnected hemisphere soon followed in healthy subjects with the use of the hemi-field technique (Sperry, 1974; D. Zaidel, 1985). What has changed somewhat throughout the years are the experimental details, emerging mostly from hundreds of experiments in which one or two facets are altered. The most intriguing question still remains: how do the two hemispheres work together? Connected by over 200 million fibers in the corpus callosum and with additional interhemispheric fibers in the anterior and hippocampal commissures, the cerebral hemispheres communicate selectively, and together they represent asymmetric yet complementary mental functions (D. Zaidel, 1993).

A. The Puzzle of Functional Brain Asymmetry

At the same time, even as the understanding of the pattern of lateralized functions was developing, there was no attempt to understand the biological logic behind the particular hemispheric assemblages, that is, the grand logic behind the separation between speaking and visuospatial cognition. We still do not fully understand why these two functions are mutually exclusive

and are specialized in opposite sides of the neocortex. Why do we even need functional asymmetry, besides efficiency in cognitive information processing?, I asked myself. Avoiding cognitive interference is an obvious answer. An additional clue must lie in the hemispheric meaning systems in LTM and in the nature of concept organization, I reasoned. Thinking, problem solving, and remembering proceed relative to what was experienced and stored previously (D. Zaidel, 1994), but even as we capitalize on what we have learned previously, we are also capable of learning new things. Thus, determining the nature of LTM or of knowledge of the world in each cerebral hemisphere could provide a glimpse into the biological logic behind functional hemispheric asymmetry.

B. Assumptions about Long-Term Memory

The mind in the brain creates order from experience with the aid of organized knowledge systems stored in LTM. Forms of knowledge organizations include serial organization, (e.g., alphabet, calendar, categorical organization [taxonomy]) (Medin & Coley, 1998; Medin, Lynch, Coley, & Atran, 1997) and schema organization (parts that together form a theme or a concept) consisting of some of our knowledge about objects, events, scenes, stories, or motor programs. Both organizations must be studied in parallel. LTM provides the framework for the meaning of what is sensed and perceived, as was shown long ago by Bartlett's famous experiments (Bartlett, 1932), regardless of whether we hear a verbal sentence and need to comprehend its meaning, when we navigate in a topographical terrain and need to know why we are there, or when we identify someone's face. The assumption here is that the same external experiences are available to both hemispheres but that each uses its own specialized knowledge system and its own strategy to store and retrieve these experiences. If LTM is the ink that colors sensations and perceptions, then I propose that the ink's shade is somewhat different in the two hemispheres. The ink's molecules and combinations could explain the particular assemblages in the left and right hemispheres.

II. Concepts, Schemas, and Categories in the Hemispheric Long-Term Memory

A. "Split-brain" (Commissurotomy) Patients: Findings on Schemas

The clue to hemispheric asymmetry in knowledge of the world emerged in testing commissurotomy patients ("split-brain"). The surgery disconnects

the cerebral hemispheres by severing the interhemispheric commissures. The patients in my studies were from the Bogen–Vogel series (also known as the Caltech series). They underwent a procedure in which the neuronal fibers connecting the hemispheres are sectioned in a single-stage operation; the corpus callosum was sectioned in its entirety as was the anterior and hippocampal commissures (Bogen, 1992; Bogen & Vogel, 1962b). Apart from a nonamnesic yet persistent memory impairment (D. Zaidel, 1990b; D. Zaidel, 1995), the surgery alleviated the debilitating seizures and the patients function essentially normally in everyday life (D. Zaidel, 1993). The full effects of hemispheric disconnection on perceptual, cognitive, and mnemonic asymmetry can be shown only under special laboratory conditions when stimuli are presented to one or the other cerebral hemisphere (Sperry, 1974; Sperry, Zaidel, & Zaidel, 1979; E. Zaidel, Zaidel, & Bogen, 1996). With this technique, I was able to uncover some of the hemispheric meaning systems (D. Zaidel, 1990a, 1994).

1. Distorting a Familiar Schema to Obtain the First Clue

The first clue to the hemispheric LTM was a serendipitous finding. I designed tests in the 1970s to uncover the causes of unilateral ideomotor apraxia (the inability to execute verbal commands) in the complete commissurotomy patients (D. Zaidel & Sperry, 1977), consciously wondering whether or not the unilateral left arm and leg errors that I observed could be explained by the notion that each disconnected hemisphere ''sees'' the world differently. Some of the verbal commands demanded knowledge of face parts in conjunction with arm movements. In the tests designed to uncover the causes, line drawings of a face showing a head-on perspective as well a face profile were lateralized to each hemisphere at a time through the use of the Z-lens, a technique which permits long lateralized visual exposures (as opposed to quick tachistoscopic presentations) (E. Zaidel, 1975). One would expect that with exposure of the stimuli to the RH, pointing to facial parts upon verbal commands would produce errorless performance, because the RH specializes in facial processing and memory (DeRenzi, Faglioni, & Spinnler, 1968). Consequently, I was puzzled by errors with a face profile in contradistinction to no errors with the head-on face. The errors suggested that a mental representation of a head-on face controlled the responses on the face profile. Dramatically, no errors whatsoever in pointing to a face profile were observed during LH performance. Thus, a clue was obtained that different mental concepts represent the same external reality in each hemisphere (D. Zaidel, 1984, 1990a, 1994).

2. Distortion in a Famous Painting

Using Magritte's painting, ''The Rape,'' I gained further support for the hypothesis that each hemisphere has a different view of the world (D.

Zaidel, 1984, 1990a, 1994). In this painting, the artist substitutes a female's nude torso for facial parts. What particularly fools the eye is the presence of a rich mane of hair and a normal-looking neck. Is it a face, then? In an experiment carried out in the 1970s, when I exposed a picture of this painting for a prolonged period to each disconnected hemisphere separately, with the Z-lens, the patients' RH failed to recognize the breasts, naval, and pubic for what they were but rather "acted" as if they were eyes, nose, and mouth, respectively. The LH, in contrast, was not "fooled" by the substitution, rather it recognized correctly the nude body parts even when asked to point to eyes, nose, and mouth (D. Zaidel, 1984, 1990a, 1994). The LH could view a highly familiar image, the face, and "accept" modification made to it, whereas the RH could "not accept" the deviations. Together, these results pointed to different concepts of reality in each hemisphere.

B. Natural Superordinate Categories

Several paradigms and methodologies developed by cognitive psychologists, particularly those pertaining to the category meaning system, are valid ways of measuring how knowledge is stored in each hemisphere.

1. Synopsis of Previous Findings: Central Tendency in Natural Categories

The study of the category knowledge system examines the relationships among concepts with reaction time (RT) paradigms (Medin & Coley, 1998; Medin et al., 1997). The RTs gauge "mental distances" among category members (Collins & Quillian, 1969; Rips, Shoben, & Smith, 1973; Rosch, 1975a, 1975b). Verification latencies in natural category membership tasks reflect the degree of similarity between items and the prototype invoked by the category concept (Rosch, 1975a). Prototypicality as the main organizing principle in LTM has been questioned (e.g., Barsalou, 1983), but for natural concepts there is ample evidence that it exists, and I have used the method of category membership decisions successfully to study the nature of the hemispheric LTM (D. Zaidel, 1987).

a. Participants In this experiment and in all subsequent experiments described in this paper, normal, right-handed participants were tested. They were all undergraduate students in introductory psychology courses at the University of California, Los Angeles, (UCLA), who volunteered in exchange for partial course credit.

b. Stimulus Presentation and the Hemi-Field Technique With this technique, each participant fixates the visual gaze on a red dot placed in the middle of the viewing screen while single images are flashed tachistoscopi-

cally, either to the left or right of the dot, in a pseudorandom order. Because of neuroanatomical arrangement of the visual system, this procedure effectively lateralizes the input to the left (LVF) or right (RVF) visual half-fields, from whence the visual information reaches the RH or LH, respectively.

In the central tendency in natural superordinate categories experiment, participants had to make speeded category membership decisions for natural superordinate categories (furniture, weapon, vehicle, fruit, vegetable) (D. Zaidel, 1987). As the statistical results were already published in full, a summary and graphic illustrations will be provided here in the interest of making a complete story. The typicality level of the exemplar was obtained from Rosch's (1975) norms. The accuracy rate was very high in both visual half-fields, indicating the existence of knowledge of the world in both hemispheres. The results for the RTs revealed a significant difference between high and low typicality exemplars in the RH but no difference in the LH (see Figure 1). If there were no statistical interaction, we would have concluded that hemispheric concepts are organized in highly similar ways. Rather, this outcome illustrated the asymmetry in the LTM category meaning system (D. Zaidel, 1987).

Subsequently, after testing three complete commissurotomy patients in two experiments, Cronin-Golomb confirmed presence of two qualitatively different semantic networks in both hemispheres (Cronin-Golomb, 1995). The task for the patients was to decide which member of a pair of pictures was more closely related to the target picture. Her findings suggested the presence of similar knowledge systems in both disconnected hemispheres but each operating with qualitatively different semantic organization.

2. *Instances of High Typicality Exemplars of Natural Categories*

a. Effect of Different Instances Given high typicality exemplars only, there may be a further hierarchy within this knowledge domain. To answer the question, pictures of instances from the categories furniture, weapon, and vehicle, with two different instances representing each of the categories, were used as stimuli. Thus, in this study there were six different pictures of chairs, drawers, guns, swords, cars, and trucks.

Sixteen subjects were tested individually in the hemi-field technique. Each subject viewed single pictures in the LVF or RVF for an exposure duration of 100 msec each. The subject's task was to decide with a button press whether or not the picture was a representation of a specific category named by the experimenter (i.e., of furniture, weapon, or vehicle).

The results revealed a very high accuracy rate in both visual half-fields, as in the previous experiment. Figure 2 is a graphic illustration of the mean RT data. The RTs were analyzed with a repeated measures analysis of

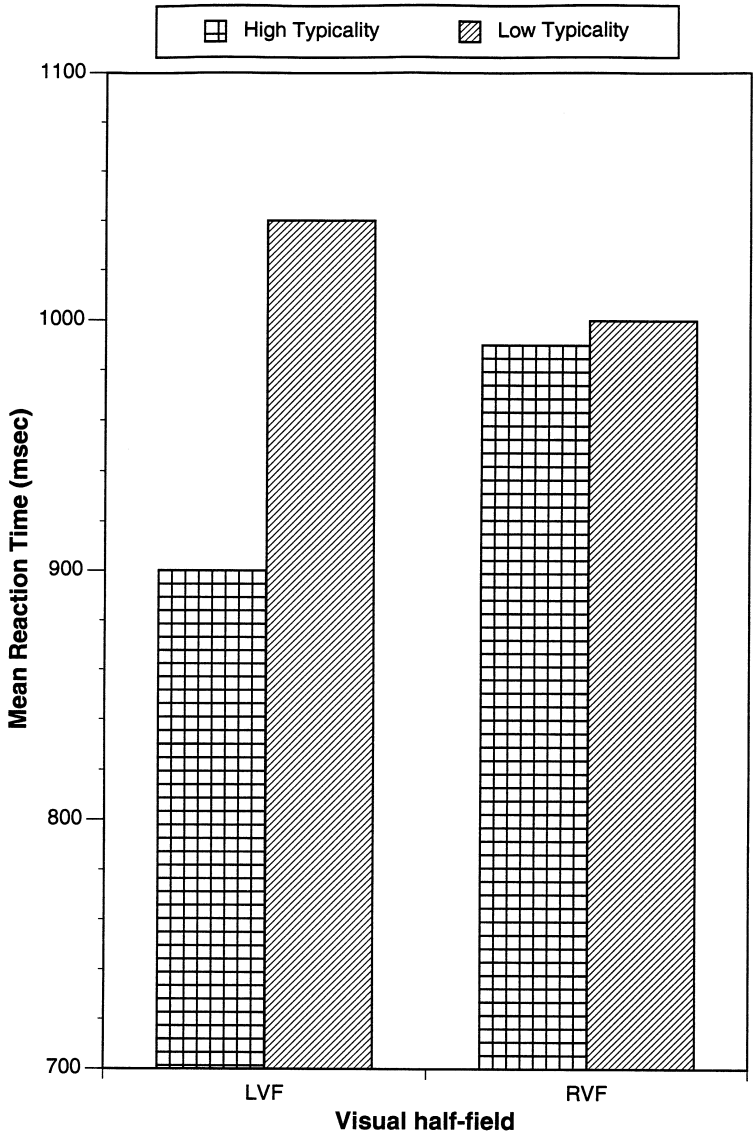


Fig. 1. Previously published results on central-tendency category decisions in the left and right visual half-fields (LVF and RVF, respectively). The stimuli were pictures representing high versus low typicality exemplars of natural superordinate categories (D. Zaidel, 1987). The dependent measure was reaction time. There were no differences in accuracy.

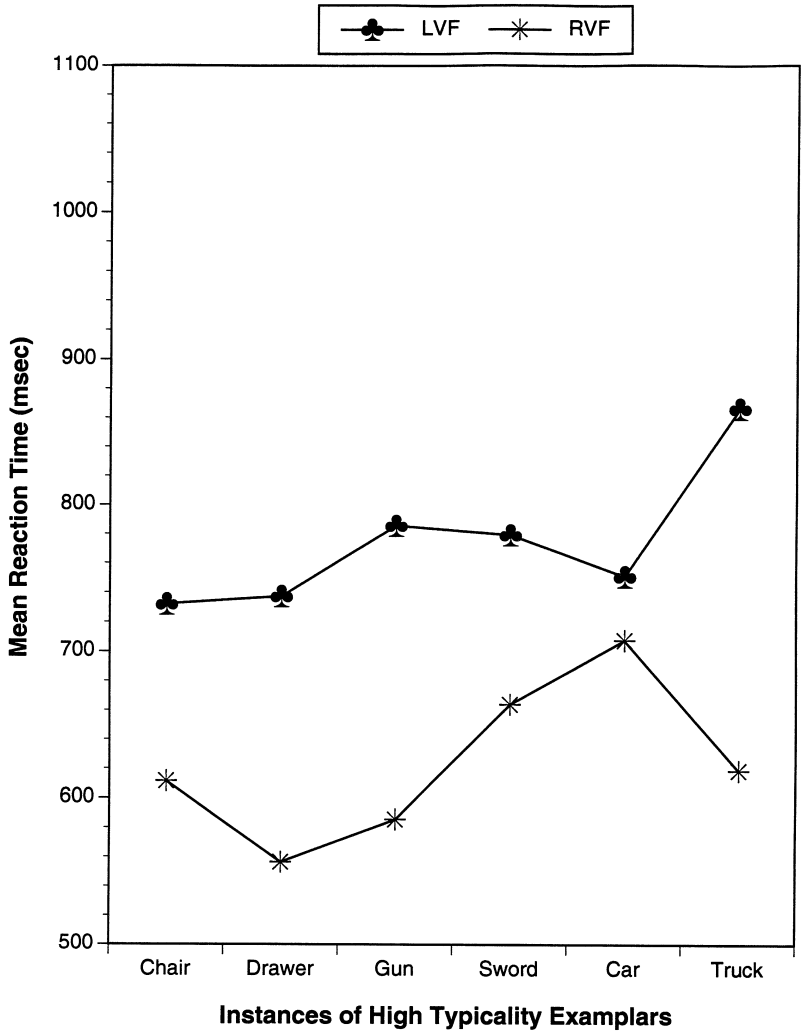


Fig. 2. This graphical summary shows mean reaction time to the effects of different instances of high typicality exemplars of natural superordinate categories, left and right visual half-fields in the (LVF and RVF, respectively).

variance (ANOVA) with a within-subjects factor of Visual Field (LVF, RVF) and a within-subjects factor for Instance (chair, drawer, gun, sword, car, truck). Only RTs to correct "yes" responses were analyzed. The results revealed a significant main effect for Visual Field $F(1, 15) = 37.77$, $p < .00001$, reflecting the fact that RT was faster in the RVF (mean =

623.98 msec) than in the LVF (mean = 775.26), and a significant interaction of Visual Field × Instance, $F(5, 11) = 14.75$, $p < .0001$. Analyses of simple effects revealed significant left–right differences for each of the instances, except for car. Indeed, with the exception of vehicle in the LVF, RTs to the two instances for each of the remaining categories were highly similar. This is contrasted with graded RTs in the RVF between each of the two exemplars for a given category. The RT pattern clearly shows hemispheric asymmetry in the organization of levels of category concepts.

b. ''Representativeness Effect'' for Instances of High Typicality Exemplars Given that some exemplars of natural categories are more typical than others, the question is whether or not different instances of an exemplar are equally ''representative'' for both hemispheres. A chair is a highly typical exemplar of the category furniture. Are some chairs more representative of ''chairness'' than others? The next question was whether or not some instances are more representative than others, even when they all represent a high typicality exemplar. Thus, the representativeness level for each picture was determined from the responses of 33 subjects who rated each instance on a 7-point scale, with 1 being a ''very good'' representation and 7 being a ''very poor'' representation. These typicality ratings were used to reanalyze the responses of the 16 subjects tested above, in order to determine the interaction of high versus low typicality with the responses. The cutoff for high typicality was a mean of 2.3 or less, whereas for low typicality, the cutoff was 5.75 or higher.

Figure 3 illustrates the RTs in the LVF and RVF. Again, we see a clear hemispheric asymmetry as a function of typicality level. A repeated measures ANOVA with a within-subjects factor for Visual Field (LVF, RVF), Typicality (High, Low), and Instance (chair, drawer, gun, sword, car, truck) was applied to the RTs in correct ''yes'' responses. The results revealed a significant main effect for Visual Field, $F(1, 15) = 41.88$, $p < .00001$, reflecting faster responses in the RVF than in the LVF, as well as a significant main effect for Typicality, $F(1, 15) = 24.80$, $p < .0002$, reflecting the fact that responses to high typicality instances were faster (mean = 608.92) than to low typicality instances (mean = 718.20). The interaction of Typicality × Instance was significant, $F(5, 11) = 5.30$, $p < .01$. The triple interaction of Visual Field × Typicality × Instance was significant as well, $F(5, 11) = 3.23$, $p < .01$. There were no other significant main effects or interactions. Subsequent analyses for simple effects revealed that Visual Field × Typicality for car was significant, $F(1, 15) = 6.97$, $p < .01$, and similarly a significant interaction for drawer dresser, $F(1, 15) = 8.52$, $p < .01$). This can be seen clearly in Figure 3. The fact that there was a significant interaction between instances and visual half-field suggests

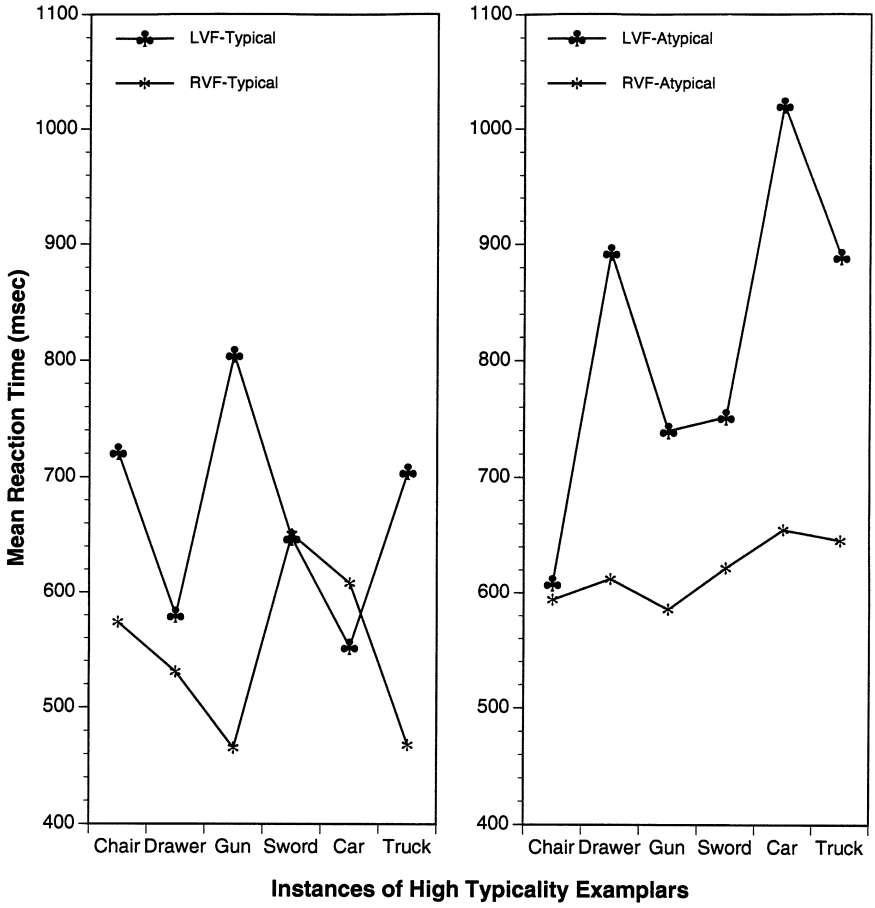


Fig. 3. Summary of the mean reaction time (RT) results investigating "representativeness effect" for instances of high typicality exemplars of natural superordinate categories, in the left and right visual half-fields (LVF and RVF, respectively).

that some categories are stored/retrieved preferentially in one versus the other hemisphere. And, as we saw earlier, representativeness had a greater effect in the LVF than in the RVF.

C. Logical Categories

1. Family Resemblance versus Logical Membership

Wittgenstein (Wittgenstein, 1953) is credited with alerting us to the complex structure of a meaning in a category name. Many objects fit under the

natural category of furniture, all looking different than each other and yet unified in the concept we have of furniture. Perhaps they share only 1 feature, perhaps 20 features (e.g., their physical properties, functionality, spatial location). Wittgenstein offered that a family resemblance is what unites the features into a single concept. However, in contrast to natural categories whose boundaries are ''fuzzy,'' logical categories are defined by a limited number of features that satisfy necessary and sufficient inclusion criteria (D. Zaidel & Frederick, 1988). The main interest here is whether or not hemispheric category meaning systems can be further understood and distinguished according to fuzzy versus strict inclusion criteria.

2. *Parity Category Membership*

I used the logical category of parity, odd and even numbers. There were 48 participants. On each of 40 trials, subjects viewed a pair of numerals and had to decide whether or not they belonged to the same category, odd or even, with a button press. The odd and even numerals were equally chosen from small (>1, <10), medium (>10, <20), and high (>40, <50) magnitudes. The following procedures were employed: the first number appeared in the center of vision for 500 msec, followed, after a 500-msec delay, by the second number, which was exposed in either the left or right side of the visual fixation point, for 100 msec. On 20 successive trials the task for the subject was to decide whether or not both members of the pair were even numbers, whereas on 20 different successive trials they had to decide whether or not both members were odd numbers. Both the hand pressing the button (right versus left) and the order of the parity trials were counterbalanced across subjects.

Accuracy was very high in both visual half-fields. The RT results are illustrated graphically in Figure 4. A repeated-measures ANOVA was applied to the RT data with a within-subjects factor of Visual Field (LVF, RVF) and Parity (odd, even). There was a significant main effect for Parity, $F(1, 47) = 29.93$, $p < .00001$, reflecting faster responses to odd than to even membership. The factor of Visual Field was not statistically significant. The interaction of Visual Field $\times$ Parity was significant, $F(1, 47) = 5.55$, $p < .02$, reflecting greater sensitivity to parity in the LVF than in the RVF, as well as a significantly faster RT to even numbers in the RVF than in the LVF. Analysis for simple effects revealed a significant difference between odd and even in the LVF, $t(47) = 28.68$, $p < .00001$, as well as within the RVF, $t(47) = 9.66$, $p < .002$.

Logical categories are defined by strict necessary and sufficient inclusion criteria, and yet we continue to find hemispheric asymmetry in RT. Thus, the fuzzy boundaries of natural categories and the tight boundaries of

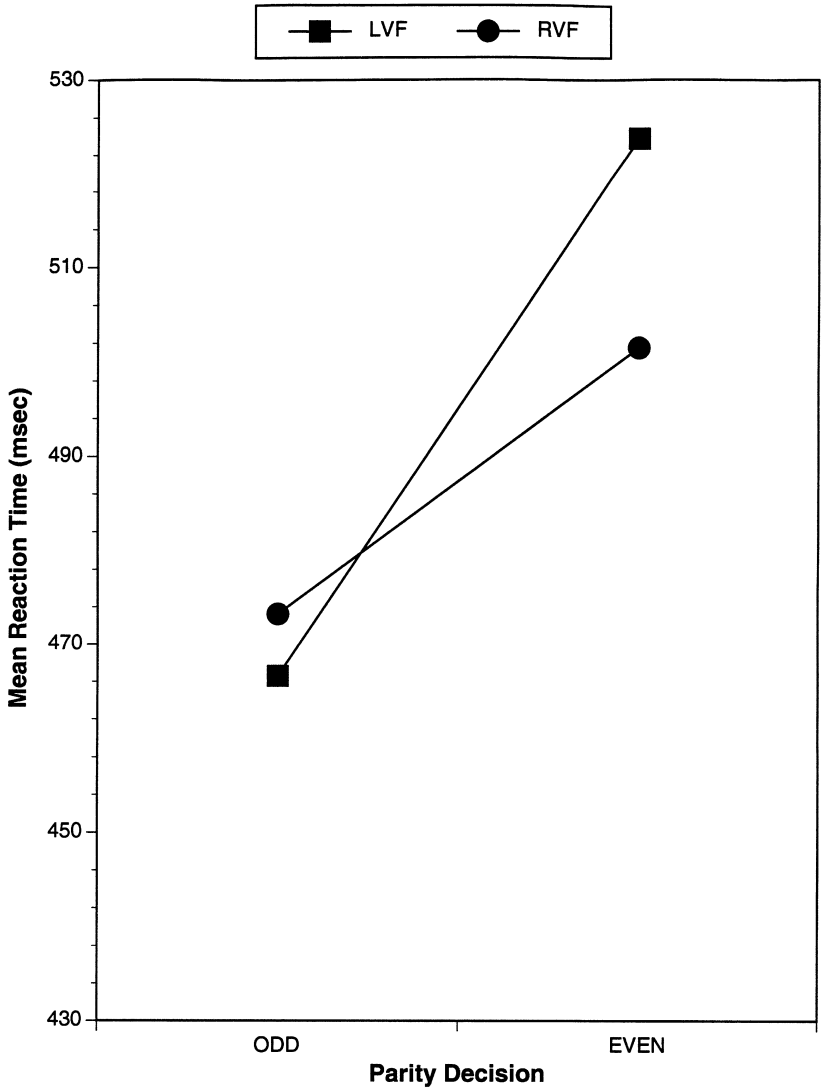


Fig. 4. Summary of mean reaction time (RT) in the left and right visual half-fields (LVF and RVF, respectively) to parity decisions (“odd” versus “even”) in a logical category.

logical categories do not distinguish between the hemispheres. Rather, a similar principle of knowledge categorization seems to apply in the hemispheric LTM. The hemispheric distinction lies in the nature of the organization of the knowledge system.

D. Schemas of Faces: A Case of a Highly Familiar Visual Object

The human face is a unique visual percept. It is a highly familiar visual pattern and is already extremely important in infancy (Fantz, 1964; Johnson, Dziurawiec, Ellis, & Morton, 1991). It is critical in human social interactions, and, as language, it appears to be modular (both functionally and structurally). Moreover, there is neuropsychological evidence that facial processing is specialized principally in the RH (Beardsworth & Zaidel, 1994; McCarthy & Warrington, 1990; Milner, 1975). RH damage, particularly in posterior regions, may lead to prosopagnosia, a neurological disorder in which the patient exhibits the inability to recognize people by their faces alone. Further, RH specialization appears right after birth, in the early postnatal period (de Schonen & Mathivet, 1989; de Schonen & Mathivet, 1990). Importantly, self-recognition can occur in the disconnected hemispheres in commissurotomy patients (Sperry et al., 1979). Following right, but not left anterior temporal lobectomy, patients have selectively poor memory for newly presented faces (Beardsworth & Zaidel, 1994), and the selective hemispheric role is seen even before the lobectomy, while the patients still suffer from repeated, drug-resistant epileptic seizures. However, work with complete commissurotomy patients clearly showed that both hemispheres play a role in face processing and that the relative contribution of each can be manipulated with task demands (Levy, Trevarthen, & Sperry, 1972).

The findings described in section II revealed an idée fixe in the RH regarding the face's appearance as contrasted with flexibility in the LH. Those results invited the question, What makes a face a face in each cerebral hemisphere? Consider aspects of the face which could selectively activate the face schema: (a) the outline contour, (b) the internal features, and (c) the spatial location of individual features. If the relationships between all of these components were violated systematically, what effect will this have on each hemisphere? In the next experiment, the outline contour was either present or absent, the configuration of the internal features was realistic (e.g., eye, nose, mouth, ear), and the spatial location of individual features was always seen according to normal "T shape". What was varied systematically, however, was the position of the internal features within the T. Figure 5 shows some examples of the stimuli in this experiment.

In this experiment the participants ($N = 24$) saw a single target in the center of vision for a 500-msec duration. This target represented a facial feature (e.g., nose, eye, ear, lips). At this duration and spatial location, both hemispheres could see the target. After a 500-msec delay, a normal face or face-like array (distorted face) was projected for 130 msec either in the LVF or RVF. Although the lateralized image was still on the screen,

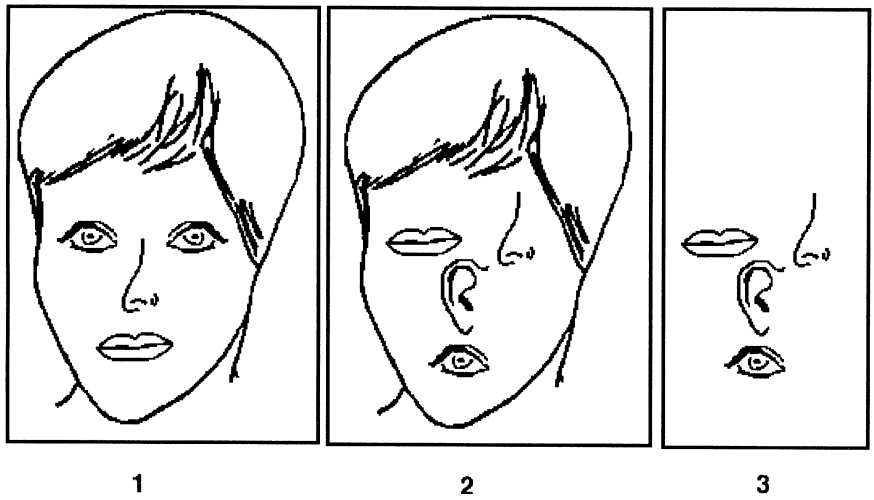


Fig. 5. Examples of the face stimuli and their distortions used in this experiment. 1 = a normal face (NF). 2 = a face distortion where the contour-frame is present but facial features are rearranged (PFR). 3 = a face distortion where the contour-frame is absent, only the internal features remain (AFR). Notice the rearrangement maintains the "T" shape of a normal face.

a small circular light cue appeared for 20 msec on top of one of the internal features in the lateralized array. This occurred 110 msec into the 130-msec exposure window. The task for the participant was to press the yes or no button to indicate whether or not the light cue had signaled a figure that matched the center-of-vision target.

The series of 48 trials consisted of normal faces (NF), face-like arrays where the contour frame was present (PFR), and face-like arrays (internal features only) where the contour frame was absent (AFR). In the face-like arrays, facial features were systematically rearranged. For every PFR, there was an AFR. These three types of stimuli were intermixed pseudorandomly within the series of trials.

The results are summarized in Figure 6. A repeated-measures ANOVA with a within-subjects factor of Visual Field (LVF, RVF) and Face Schema (NF, PFR, AFR) applied to the percent correct responses revealed a significant Visual Field $\times$ Face Schema interaction ($F(1, 23) = 4.12, p < .03$. The main interest was a distinction within each visual field between PFR and AFR. In the LVF, features in PFR were significantly less well localized than features in AFR ($t(46) = 8.83, p < .001$) or in the normal face ($t(46) = 7.23, p < .001$). In the RVF, there was no statistically significant difference between the three types of stimuli. Feature localization in the CFR was

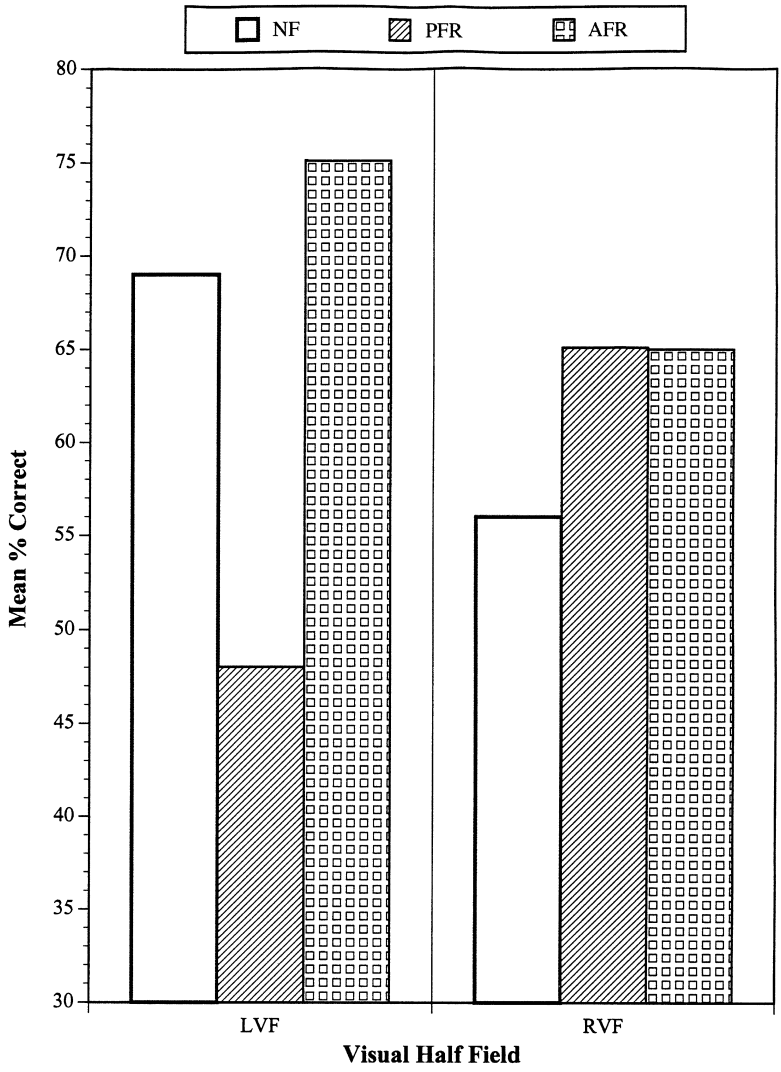


Fig. 6. Summary of mean accuracy in responding to face and face distortions in the left (LVF) and right (RVF) visual field of normal participants. Normal face (NF), contour-frame was present (PFR), contour-frame was absent (AFR).

significantly better in the RVF than in the LVF ($t(46) = 5.86$, $p < .002$), and for NF there was better localization in the LVF than in the RVF ($t(46) = 2.54$, $p < .01$).

These results are consistent with RH specialization for face processing in the normal brain. They show a RH reliance on a prototypical mental schema of a face. There was a selective reduction in feature localization for distorted faces where the contour was present and the features rearranged (CFR). The fact that absence of the contour frame (AFR) did not affect RH performance detrimentally suggests one of the mechanism in RH face processing, namely the context created by the contour. Deviations from this context may have contributed to congnitive interference in feature localization. On the other hand, the face schema was not activated in the RH with AFR stimuli. Other mechanisms not investigated in this experiment are undoubtedly involved as well, those provided by the T-shaped arrangement of the features, for example. In contrast, with LH performance we did not see a strong effect of stimulus type, and this suggests that face schema activation is fuzzier by comparison. One advantage in this type of activation is flexibility and room for cognitive maneuverability. On the other hand, one limitation is less precision in face identification.

III. Constraining Influences in the Interaction between Hemispheric Specialization and Long-Term Memory

Hemispheric specialization factors could exert constraints on interpretations of experience. In the above experiment with distorted faces, it remains to be explained why LH localization accuracy with normal faces (NF) was worse than in the RH. The answer may lie in hemispheric specialization: although the RH specializes in faces, the LH does not. It is reasonable to assume that hemispheric specialization in specific functions (e.g., language production, face recognition, topographical orientation) interacts with experiential knowledge stored in LTM. Language-related hemispheric work by Chiarello and associates supports this notion (Chiarello, 1998). Using the hemi-field technique in a lexical decision task (Faust & Chiarello, 1998a), participants decided which word best resolves an ambiguous sentence. In the RVF, incongruent words did not facilitate sentence ambiguity, only congruent words did so. In the LVF, facilitation of ambiguity occurred regardless of whether the word was congruent or incongruent. Thus, these results can be interpreted to show that when it comes to its nonspeciality, language, the RH can provide alternative word meanings flexibly, whereas the LH appears somewhat limited in that regard. In another sentence facilitation study (Faust & Chiarello, 1998b), the specialization of the LH in syntax emerged: Sentence completion was facilitated in the RVF regardless of the number of intervening words between the critical word in the sentence and a target word; in the LVF performance was limited by the

number of intervening words: the higher the number, the worse the performance. Thus, independent of organization and strategies in the hemispheric LTMs, there are constraining factors, and some of those involve principles of hemispheric specialization.

IV. Possible Answers to the Biological Logic of Hemispheric Long-Term Memory

A. Processing Experiences

Admittedly, we do not fully understand why speaking and visuospatial cognition are mutually exclusive and are specialized in opposite hemispheres, but perhaps we can be bold and generalize to various cognitive domains of our lives: Things that do not change rapidly are spatial layouts of the world such as geography, the position of the sun, moon, stars, constellations, oceans, rivers, mountains, valleys, canyons, vegetation, or roads. Human and animal faces have fixed configurations as well. There is clearly a biological advantage to having a concept and meaning system in the RH that is rigid with respect to spatial and facial layouts. The LH with its language-rich cognition may be neuronally wired to accept countless associations and distortions, allowing flexibility in interpretations of experience and even deriving new meanings. For example, humans frequently utter incorrect grammatical sentences and yet listeners understand the meaning, whether this occurs in the one-word stage of child development or in adulthood (D. Zaidel & Kasher, 1989). The essence of human language is its seemingly limitless combinatorial power. Perhaps the neuroanatomical underpinning of this power developed for the first time in human brain evolution, providing a cognitive system to support language communication where imperfections can still lead to meaning derivation. In my view of the cerebral hemispheres, surrealistic paintings by Magritte, Dali, or Archimboldo, in which physical, logical, and social violations of reality are depicted, typify this picture of LH cognition.

Although it is neither intuitive nor obvious why rigid and flexible cognitive systems need to be separated in the brain, this may have been the prerequisite in the evolution of human hemispheric specialization. The RH is vigilant with regards to what is well known in the geography, the terrain, faces, or etiquette, and the LH is receptive to new possibilities and interpretations of that which is known, familiar, or established in these very same domains. Hemispheric independence of such cognitive processes may in the long run be cognitively more efficient than if these processes occurred intra hemispherically.

B. Processing Facial Attractiveness in an Evolutionary Perspective

Mate selection strategies are critical features of biological survival. Consider the face of the owner and the brain of the observer. We have found that the organization of attractiveness in the face is sex-related; specifically, in women's faces the right half is significantly more attractive than the left half, whereas in men's faces there is no left–right difference in this regard (D. Zaidel, Chen, & German, 1995). On the other hand, we found that smiling is more salient in the left half of the face, and this is so in both women's and men's faces. What is the biological advantage of such arrangements? Surely it is not coincidental. We proposed that the arrangement is adaptive, having evolved in concert with brain evolution, with the progression in mammalian brains toward anatomical and functional asymmetry of the cerebral hemispheres. The context for the development of asymmetric functionality in the face is mate-selection strategies in humans. In the animal kingdom, attractiveness and health are highly linked and are essentially synonymous. We investigated this link in another study of human faces (Reis & Zaidel, 2000) and found that the pattern of judging left and right halves of women's and men's faces on the appearance of health was essentially identical to the pattern observed for judging the same faces on attractiveness. In humans, as in a large variety of nonhuman mammals, what is healthy is also attractive. However, because decision is made in the brain of the observer, there must be neuroanatomical features that process what is necessary. A woman's face has evolved to be what it is in response to the unique features of a male's brain, whatever they are. Similarly, a man's face has evolved to be what it is in response to a female's brain. Thus, we have proposed that the brain of the observer is biologically linked to the face of the observed, the two having evolved in concert.

V. Conclusions

Despite the fact that in humans the cerebral hemispheres are interconnected with the largest fiber tracts in the brain, the experiments described here show that they have separate strategies for handling knowledge stored in LTM. Considering millions of years of primate and human evolution in which the corpus callosum grew enormously in size and presumably so did interhemispheric communication, distinctiveness of the hemispheric LTMs suggests adaptive functional complementarity. It is not as if the hemispheres have wholly different stores in LTM, rather, the differences lies in their organization. Within each store, there must also be an interaction with

principles of hemispheric specialization *per se* (language in the LH and spatial cognition in the RH).

Convergent support for my characterization of the LH and RH knowledge-of-world systems has now been reported by others. Cronin-Golomb (Cronin-Golomb, 1995) studied the semantic networks of picture associations in the disconnected hemispheres of three complete commissurotomy patients (Bogen–Vogel, Caltech series). She concluded that the RH made typical and conventional associations and thus specializes in conventional meaning, whereas the LH made unusual associations to the target and thus specializes in processing deviations from standard meaning. Metcalfe and colleagues (Metcalfe, Funnell, & Gazzaniga, 1995) studied J. W., a collosotomy patient (who underwent a two-stage surgery), and based on six experiments concluded that the RH's specialization in veridical, literal information contributes to correct categorization of related events, whereas the LH's specialization in language contributes to inferences that deviate from veridical information.

Except for principles of hemispheric specialization where there can be sharp boundaries, there is by now little doubt that the cerebral hemispheres are not functionally as different from each other as ''night and day.'' LTM may contribute to interhemispheric communication, and LTM's organization and strategies may be the mental space where the colors of experience are combined with functional hemispheric specialization.

References

Barsalou, L. W. (1983). Ad hoc categories. *Memory and Cognition, 11,* 211–227.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, UK: Cambridge University Press.

Beardsworth, E., & Zaidel, D. W. (1994). Memory for faces in epileptic children before and after unilateral temporal lobectomy. *Journal of Clinical Experimental Neuropsychology, 16,* 738–748.

Bogen, J. E. (1992). The callosal syndromes. In K. M. Heilman & E. Valenstein (EDs.), *Clinical neuropsychology*. New York: Oxford University Press.

Bogen, J. E., Schultz, D. H., & Vogel, P. J. (1988). Completeness of callosotomy shown by magnetic resonance imaging in the long term. *Archives of Neurology, 45,* 1203–1205.

Bogen, J. E., & Vogel, P. J. (1962a). Cerebral commissurotomy in man. Preliminary case report. *Bulletin of the Los Angeles Neurological Society, 27,* 169–172.

Bogen, J. E., & Vogel, P. J. (1962b). Cerebral commissurotomy in man. Preliminary case report. *Bulletin of the Los Angeles Neurological Society, 27,* 169–172.

Bogen, J. E., & Vogel, P. J. (1975). Neurologic status in the long term following complete cerebral commissurotomy. In F. Michel & B. Schott (Eds.), *Les syndromes de disconnexion calleuse chez l'homme*. Lyon: Hospital Neurologie.

Chiarello, C. (1998). On codes of meaning and the meaning of codes: Semantic access and retrieval within and between hemispheres. In M. Beeman & C. Chiarello (Eds.), *Right*

hemisphere language comphrehension: Perspectives from cognitive neuroscience. Mahwah, NJ: Earlbaum.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8,* 240–248.

Cronin-Golomb, A. (1986). Comprehension of abstract concepts in right and left hemispheres of complete commissurotomy subjects. *Neuropsychologia, 24,* 881–887.

Cronin-Golomb, A. (1995). Semantic networks in the divided cerebral hemispheres. *Psychological Science, 6,* 212–218.

de Schonen, S., & Mathivet, E. (1989). First come, first served: A scenario about the development of hemispheric specialization. *European Bulletin of Cognitive Psychology, 9,* 33–44.

de Schonen, S., & Mathivet, E. (1990). Hemispheric asymmetry in a face discrimination task in infants. *Child Development, 61,* 1192–1205.

DeRenzi, E., Faglioni, P., & Spinnler, H. (1968). The performance of patients with unilateral brain damage on face recognition tasks. *Cortex, 5,* 274–284.

Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science, 46,* 669–670.

Faust, M., & Chiarello, C. (1998a). Sentence context and lexical ambiguity resolution by the two hemispheres. *Neuropsychologia, 36,* 827–835.

Faust, M., & Chiarello, C. (1998b). Constraints on sentence priming in the cerebral hemispheres: effects of intervening words in sentences and lists. *Brain and Language, 63,* 219–236.

Finger, S. (1994). History of neuropsychology. In D. W. Zaidel (Ed.), *Neuropsychology* (pp. 1–28). San Diego: Academic Press.

Galaburda, A. M. (1995). Anatomic basis of cerebral dominance. In R. J. Davidson & K. Hugdahl (Eds.), *Brain asymmetry* (pp. 51–73). Cambridge, MA: The MIT Press.

Hiscock, M., & Kinsbourne, M. (1995). Phylogeny and ontogeny of cerebral lateralization. In R. J. Davidson & K. Hugdahl (Eds.), *Brain asymmetry* (pp. 535–578). Cambridge, MA: The MIT Press.

Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborn's preferential tracking of face-like stimuli and its subsequent decline. *Cognition, 40,* 1–19.

Levy, J., Trevarthen, C., & Sperry, R. W. (1972). Perception of bilateral chimeric figures following hemisphere deconnexion. *Brain, 95,* 61–78.

McCarthy, R. A., & Warrington, E. K. (1990). *Cognitive neuropsychology*. London: Academic Press.

Medin, D. L., & Coley, J. D. (1998). Concepts and categorization. In Julian Hochberg (Ed.), *Perception and cognition at century's end*. San Diego: Academic Press.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32,* 49–96.

Metcalfe, J., Funnell, M., & Gazzaniga, M. S. (1995). Right-hemisphere memory superiority: Studies of a split-brain patient. *Psychological Science, 6,* 157–164.

Milner, B. (1975). Psychological aspects of focal epilepsy and its neurosurgical management. In J. K. Purpura, R. D. Penry, & R. D. Walters (Eds.), *Advances in neurology*. New York: Raven.

Reis, V. A., & Zaidel, D. W. (2000). Brain and Face: Communicating Signals of Health in the Left and Right. *Brain and Cognition, in press*.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12,* 1–20.

Rosch, E. (1975). Cognitive representation of semantic categories. *Journal of Experimental Psychology: General, 104,* 192–233.

Rosch, E., & Mervis, C. B. (1975b). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605.

Sperry, R. W. (1974). Lateral specialization in the surgically separated hemispheres. In F. O. Schmitt & F. G. Worden (Eds.), *The Neurosciences Third Study Program*. Cambridge, MA: MIT Press.

Sperry, R. W., Zaidel, E., & Zaidel, D. (1979). Self-recognition and social awareness in the deconnected minor hemisphere. *Neuropsychologia, 17,* 153–166.

Wittgenstein, L. (1953). *Philosophische Untersuchungen [Philosophical investigations]*. Oxford: Blackwell.

Zaidel, D. W. (1984). Cognitive functions in the right hemisphere. *La Recherche, 15,* 332–340.

Zaidel, D. W. (1985). Hemi-field tachistoscopic presentations and hemispheric specialization in normal subjects. In D. F. Benson & E. Zaidel (Eds.), *The dual brain: Hemispheric specialization in humans*. San Diego: Guilford Press.

Zaidel, D. W. (1986). Memory for scenes in stroke patients: Hemispheric processing of semantic organization in pictures. *Brain, 109,* 547–560.

Zaidel, D. W. (1987). Hemispheric asymmetry in long-term semantic relationships. *Cognitive Neuropsychology, 4,* 321–332.

Zaidel, D. W. (1988). Hemi-field asymmetries in memory for incongruous scenes. *Cortex, 24,* 231–244.

Zaidel, D. W. (1990a). Long-term semantic memory in the two cerebral hemispheres. In C. Trevarthen (Ed.), *Brain circuits and functions of the mind*. New York: Cambridge University Press.

Zaidel, D. W. (1990b). Memory and spatial cognition following commissurotomy. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (pp. 151–166). Amsterdam: Elsevier.

Zaidel, D. W. (1993). View of the world from a split-brain perspective. In E. M. R. Critchley (Eds.), *Neurological boundaries of reality*. London: Farrand Press.

Zaidel, D. W. (1994). Worlds apart: Pictorial semantics in the left and right cerebral hemispheres. *Current Directions in Psychological Science, 3,* 5–8.

Zaidel, D. W. (1995). Separated hemispheres, separated memories: Lessons about memory from split-brain patients. In R. Campbell & M. Conway (Eds.), *Broken memories*. Oxford: Blackwell.

Zaidel, D. W., Chen, A. C., & German, C. (1995). She is not a beauty even when she smiles: Possible evolutionary basis for a relationship between facial attractiveness and hemispheric specialization. *Neuropsychologia, 33,* 649–655.

Zaidel, D. W., & Frederick, K. (1988). Hemispheric asymmetries in logical categories: Convergent evidence from normal and split-brain subjects. *Society for Neuroscience Abstracts, 14,* 1138.

Zaidel, D. W., & Kasher, A. (1989). Hemispheric memory for surrealistic versus realistic paintings. *Cortex, 25,* 617–641.

Zaidel, D. W., & Sperry, R. W. (1977). Some long-term motor effects of cerebral commissurotomy in man. *Neuropsychologia, 15,* 193–204.

Zaidel, E. (1975). A technique for presenting lateralized visual input with prolonged exposure. *Vision Research, 15,* 283–289.

Zaidel, E., Zaidel, D. W., & Bogen, J. E. (1996). Disconnection syndromes. In J. G. Beaumont, P. M. Kenealy, & M. J. C. Rogers (Eds.), *The Blackwell dictionary of neuropsychology* (pp. 279–285). Oxford: Blackwell.

This Page Intentionally Left Blank

THE CAUSAL STATUS EFFECT IN CATEGORIZATION
An Overview

Woo-kyoung Ahn and Nancy S. Kim

When we categorize objects in the world or think about concepts, some aspects of entities matter more than others. In our concept of tires, for instance, roundness is more important than a black color. Even children believe that origins are more primary than certain specific aspects of appearances in categorizing animals (Keil, 1989). This chapter discusses why some features of concepts are more central than others. We will begin with a review of possible determinants of feature centrality. Then, the main body of this chapter will focus on one important determinant, the effect of people's causal background knowledge on feature centrality.

I. Different Approaches to Understanding Feature Centrality

In general, three approaches have been taken to understand feature centrality in concepts: the content-based approach, the statistical approach, and the theory-based approach. We will review each of these approaches, and discuss how they complement each other.

A. Content-Based Approach

In what we have termed the content-based approach, specific dimensions are documented as being central for certain categories or domains, but not

THE PSYCHOLOGY OF LEARNING AND MOTIVATION, VOL. 40

Copyright © 2001 by Academic Press.
All rights of reproduction in any form reserved.
0079-7421/00 $35.00

necessarily for others. For instance, it was found that for children, the shape of a rigid object is the most important dimension used to name the object. Landau, Smith, and Jones (1988) presented 2- and 3-year-old children with a small, blue, wooden inverted U-shaped object, and told the children that the object was a "dax." When asked to select other objects that were also a dax, children preferred objects with the same shape over those with the same size or material. The results, however, changed for different kinds of categories. If the named objects had eyes, children generalized names along the shape and texture dimension (Jones, Smith, & Landau, 1991). If the objects were made of nonrigid substances such as shaving cream, children generalized names along the color and texture dimensions (Soja, Carey, & Spelke, 1991).

Another example of the content-based approach comes from studies contrasting natural kinds with artifacts. These studies demonstrated that features internal to the object, such as molecular structure, were more important for categorization of an object as a natural kind than for categorization of an object as an artifact. In contrast, features external to the object, such as function, were generally shown to be more important for artifact category membership than for natural kind category membership. An intuitive example is illustrated in Smith Dick King's (1993) *Babe: The Gallant Pig* (later made into the movie *Babe*): A pig who wanted to be a sheepdog by performing all the correct functions of a sheepdog was scoffed at by other barnyard animals because performing the right function cannot change a natural kind's categorical identity.

Barton and Komatsu (1989) presented adult participants with either natural kinds (e.g., goat) or artifacts (e.g., tire) that had undergone various changes. The results showed that molecular changes (e.g., a goat having a changed chromosomal structure or a tire not made of rubber) mattered more for natural kinds, but functional changes (e.g., a female goat not giving milk or a tire that cannot roll) mattered more for artifact membership. Based on these findings, Barton and Komatsu (1989) drew a conclusion that takes the strong content-based approach, arguing that

> having a particular molecular or chromosomal structure is necessary and sufficient for (i.e., defining of) membership in a particular natural kind category and that having a particular function is necessary and, perhaps to a lesser extent, sufficient for membership in a particular artifact category. (p. 444)

Other researchers also found a similar pattern of results. Keil's studies (1989) show that for fourth graders and adults, changing the perceptual appearance of animals was less likely to change the animals' identity than changing the origin of the animals. However, the opposite pattern occurred for

artifacts. Similarly, Rips (1989) asked adult participants to read stories about various kinds of transformations. In one story, an animal called a sorp transformed from a bird-like creature into an insect-like creature as a result of either chemical hazards or maturation. Rips's (1989) participants rated chemical hazards as less likely to cause a change in the natural kind's identity than maturation. With artifacts, in contrast, the critical determinant in changing their category membership was the change in the function intended by its designer. It should be noted that unlike Barton and Komatsu (1989), Keil (1989) and Rips (1989) did not take the content-based approach, taking instead the theory-based approach, which will be discussed later.

A final example of the content-based approach is presented in Bloom (1996), who argued that ''the extension of artifact kind X to be those entities that have been successfully created with the intention that they belong to the same kind as current and previous X's'' (p.10). For instance, an object is a chair if it was successfully created with the intention that it would be a chair. Thus, for artifact kinds, the designer's intention to create that object is a defining, and the most central, feature. One critical problem with this argument lies in its circularity: The designer's intention that the object belongs to artifact kind X presupposes X, and therefore, it cannot explain why we have X in the first place.

More generally, a critical problem with the content-based approach is that it does not offer a more fundamental account for why a certain feature is more psychologically central in a concept. That is, simply asserting that a designer's intention or functional features are defining to artifacts does not explain why that became the case in the first place. In relation to the shape bias for rigid objects discussed earlier, Jones et al. (1991) proposed that this was a product of statistical regularities in language; rigid objects with the same name tend to share the same shape. Again, however, it does not explain why our language contains such statistical regularities in the first place. That is, what psychological constraints or mechanisms compelled us to label similarly shaped rigid objects with a common name? Thus, this approach has difficulty with making novel predictions as to what type of feature would be central in a novel domain.

The second general problem with the content-based approach is that the empirical results do not always agree with the claims. For instance, Malt and Johnson (1992) found that physical properties of artifacts (e.g., for boats: ''is wedge-shaped, with a sail, anchor, and wooden sides,'' as selected by undergraduate subjects in the study) were judged to be more important than, or just as important as, functional features (e.g., ''to carry one or more people over a body of water for purposes of work or recreation''). In a study that more directly examines the role of functional features, Ahn (1998) showed systematically conditions under which functional features

can be central for natural kinds and compositional can be central for artifact. We will later revisit this study.

B. Statistical Approach

The second approach to the issue of feature centrality is to examine statistical patterns of features in a category, such as category validity (the probability that an object has a certain feature given that it belongs to a certain category; e.g., the probability that an object has wings given that it is a bird) and cue validity, or diagnosticity (the probability that an object belongs to a certain category given that it has a certain feature; e.g., the probability that an object is a bird given that it has wings). For instance, Rosch and Mervis (1975) used category validity (''the number of items in a category that had been credited with that attribute,'' p. 578) as a measure of family resemblance, and showed that this measure was highly correlated with subjects' typicality ratings. That is, having features shared by many members of the same category makes an exemplar ''good.'' Tversky (1977) showed how the diagnosticity of features can affect similarity judgments. Kruschke's (1992) ALCOVE learns attention strengths (i.e., feature centrality) as a function of the diagnosticity of features. Corter and Gluck (1992) proposed a measure called category utility, which is the product of the base rate, cue validity, and category validity of a feature. They also showed that category utility is a good predictor for which level in a hierarchical structure is selected to be the basic level of categorization. Thus, the general consensus has been that the cue and category validities of a feature determine feature centrality to some extent.

However, the centrality of features does not seem to depend exclusively on the probabilities associated with features. For instance, a square basketball and a square cantaloupe both have category validities and cue validities of zero, but people are more willing to accept a square cantaloupe than a square basketball (Medin & Shoben, 1988).

Furthermore, the importance of cue and category validities in determining feature centrality depends on the type of categorization task. Cue validity would play a more important role in a task where one determines whether an object is A or B than in a task where one conceptualizes about an object. For instance, curvedness is a highly diagnostic feature in determining whether an object is a banana or a lemon. However, it is not as central when we think about bananas because a straight banana is acceptable (Medin & Shoben, 1988). Similarly, all purple seedless grapes are purple, but being purple is not so much conceptually central for purple seedless grapes in that it is merely a surface feature of purple seedless grapes rather than an essential feature of what really makes them purple

seedless grapes, such as their origin or DNA (Ahn & Sloman, 1997; Sloman & Ahn, 1999).

C. Theory-Based Approach

The third approach to feature centrality posits that the centrality of any given feature is determined by its importance in the principles or theories underlying the categories (Murphy & Medin, 1985). Categories are connected in many complex ways that resemble structured theories. For example, our concept of ''boomerang'' is connected with other concepts such as ''throwing,'' ''air,'' ''speed,'' and so on, all of which are intricately connected as in a scientific theory. Features that play an important part in commonsense or naive theories seem more essential in categorization than those that do not. For example, the feature *curvedness* plays a role in a theory of physics, and it thereby becomes a central feature for categorizing objects as boomerangs. In the case of bananas, however, curvedness does not play a critical role in a naive biological theory, and it is therefore not so central. Although the previous theory-based approaches have all discussed the importance of causal background knowledge, the exact mechanism has rarely been articulated (Gelman & Kalish, 1993; Murphy, 1993). That is, previous approaches have not specified how to determine the role a feature plays in one's domain theory. We extended the theory-based approach and proposed a causal status hypothesis (Ahn, 1998; Ahn, Kim, Lassaline, & Dennis, in press). This hypothesis provides one way to operationalize feature centrality in terms of one's causal background knowledge. This chapter will extensively discuss the causal status hypothesis.

Before we discuss the causal status hypothesis, it should be pointed out in all fairness that the theory-based approach alone cannot completely account for all issues in feature centrality even if its underlying mechanism is specified. Instead, there are a number of different kinds of determinants of feature centrality. At the very least, a feature's centrality is determined by the domain an object belongs to, a feature's statistical value, and a feature's role in one's background knowledge. For now, we will focus only on how one's causal background knowledge can determine a feature's centrality and will later return to the issue of different kinds of feature centrality.

II. Causal Status Hypothesis: General Introduction

In accordance with the theory-based view, we assume that concepts consist of richly structured features, rather than a set of independent features

(Murphy & Medin, 1985). We assume that people have naive theories about how these features are connected and that the majority of features of existing concepts are causally connected[1] (Carey, 1985; Gelman & Kalish, 1993; Wellman, 1990). Given such representational assumptions, the causal status hypothesis states that people regard cause features as more important and essential than effect features in their conceptual representations.

The causal status hypothesis is intuitive, and its real-life examples are abundant. We form an illness category based on the virus that causes the symptoms rather than by the symptoms per se. DNA structure causes many other properties of plants and animals, and hence is considered the most important feature in their classification (e.g., a plant that is found to lack tulip DNA will never be classified as a true tulip). In law, the severity of the crime often depends more on the suspects' intentions rather than their surface behaviors (e.g., killing someone by accident is a much less serious offense than intending to kill someone but inadvertently botching the plan). In judging whether somebody is nice or not, we look at their intentions rather than what they did.

At a moment's reflection, one can also generate several counterexamples to the causal status hypothesis. When doctors diagnose Alzheimer's disease, the diagnosis is not based on the cause but rather on symptoms alone. When we recognize plants and animals, we normally do not have direct access to their genetic structures, but we can confidently categorize them based only on their surface features. An orchestra selects its members based on their performance abilities rather than their training background. Then, the important question arises: to what extent does the causal status of features determine their centrality within our conceptual representations?

In the remainder of the chapter, we will focus on the causal status hypothesis. We will present our rationale for predicting the causal status effect and empirical results supporting the hypothesis under various contexts. We will describe previous categorization studies that can be accounted for by the causal status hypothesis. We will also discuss moderating factors for the causal status effect, including how other types of determinants for feature centrality reviewed in this section might alter how strongly the causal status effect is manifested. Followed by this discussion of the causal status hypothesis in the context of categorization literature, we will present one specific application of the causal status effect; namely, the effect of people's theories about mental disorders on their diagnosis decisions, and the possible ramifications of this effect for *DSM-IV* (APA, 1994) criteria. Finally, we will examine the potential consequences of focusing only on

[1] The last section of this chapter will provide more extensive discussion about these assumptions.

causal relations among features in studying the effect of lay theories on feature centrality.

III. Essentialism and the Causal Status Hypothesis

One traditional and pervasive view about concepts is that things have essences that make them what they are (e.g., Locke, 1894/1975). Essentialism is particularly related to the causal status hypothesis in that both are concerned with the causal potency of essential (or central) features. For instance, Locke stated, ''And thus the real internal, but generally in Substances, unknown Constitution of Things, whereon their discoverable Qualities depend, may be called their *Essence* [italics added]'' (p. 417). Similarly, Putnam (1977) wrote,

> If I describe something as a lemon, or as an acid, I indicate that it is likely to have certain characteristics (yellow peel, or sour taste in dilute water solution, as the case may be); but I also indicate that the presence of those characteristics, if they are present, is likely to be *accounted for by some 'essential nature'* [italics added] which the thing shares with other members of the natural kind. (p. 104)

It is controversial as to whether or not things actually have essences in the philosophical sense. Instead, Medin and Ortony (1989) proposed the idea of psychological essentialism: people act *as if* things have essences that are responsible for the surface features of objects. If psychological essentialism holds true, the deepest cause feature in a category would be our best guess as to what an essence would be, and it is reasonable to categorize objects based on the deepest cause feature[2] of a category (i.e., the essence in our conceptual representation). In this way, psychological essentialism provides a basis for the causal status hypothesis.

However, there are important differences between essentialism and the causal status hypothesis. (See Ahn et al., in press; Ahn, 1988, for more detail.) First, a strong version of essentialism might state that essences (whether known or unknown) serve as defining features of a category, and thus assume all-or-none categorization, where surface features do not affect categorization. (See Braisby, Franks, & Hampton, 1996; Diesendruck & Gelman, 1999; Kalish, 1995; Keil, 1989; Malt, 1994, for similar descriptions of essentialism.) In contrast, the causal status hypothesis does not make any statement about whether or not there is a certain class of features that serve as defining features. Indeed, the gradient structure of feature

[2] A causal chain is infinite and one might wonder how we decide when to stop explanations. We briefly discuss this issue later under VI. E. Domain Generality.

centrality is built into the hypothesis. That is, because a feature is more central than its effect but less central than its cause, it is logical to predict that the deeper a cause is in a causal chain, the more conceptually central that feature is.

Second, essences are often described as internal, hidden, or unobservable properties, such as atomic weight, genetic structure, and so on, whereas surface features are external or perceptual features such as wings and feathers in birds (e.g., Putnam, 1977). Locke (1894/1975) even stated that real essences are undiscoverable. The causal status hypothesis does not impose any restrictions on the type of cause or effect features, or the type of central or peripheral features.

Third, essentialism as claimed by philosophers (e.g., Kripke, 1972; Putnam, 1975) is a statement about things in the world, and it argues that categorization of natural kinds is independent of our knowledge of essential properties.[3] For instance, even if it turns out that water does not consist of H_2O, and instead consists of XYZ, what we have been referring to as water is still water. In that sense, natural kinds are like proper nouns in that even if we find out, for example, that Shakespeare did not write *Romeo and Juliet,* Shakespeare is nonetheless Shakespeare. However, Braisby et al. (1996) demonstrated that after discovering changes of essential properties (e.g., cats being robots controlled from Mars), people responded that category membership of the object should also be somewhat changed (e.g., the object formerly known as a cat is not a cat anymore). Likewise, the causal status hypothesis assumes that more responsibility is given to specific knowledge because feature centrality should change as our knowledge about the causal relations among features changes. For instance, if cats' behavior turns out to be caused by Martians rather than cat DNA, then cat DNA would not be as central anymore in categorizing cats.

A. Summary

Essentialism serves as a basis for the causal status hypothesis. If essences are the deepest cause in a category, both essentialism and the causal status hypothesis acknowledge the special status of essences in that they are the most central features in the category. However, unlike essentialism, the causal status hypothesis does not assume that there are defining features in concepts, categorization is all-or-none, deeper features are internal or hidden, or that categorization is independent of knowledge. In the next section, we summarize empirical support for these differences and for other predictions of the causal status hypothesis.

[3] It should be noted that the related construct of psychological essentialism (e.g., Medin & Ortony, 1989) does not make this claim, because psychological, as opposed to philosophical, essentialism is about mental representations of things in the world.

Recently, Stevens (2000) argued that previous studies by Gelman and her colleagues that are considered to be evidence for psychological essentialism can in fact be accounted for simply by assuming beliefs in causal laws without having to resort to claiming the existence of essences in our conceptual representation. For instance, Gelman and Markman (1987) showed children pictures of a gray squirrel, a kaibab (a kind of squirrel that looks like a rabbit), and a rabbit. When told that the gray squirrel eats bugs and that the rabbit eats grass, most 4-year-olds said that the kaibab eats bugs. Stevens argued that it is much more parsimonious to account for this kind of result by assuming that children believe that something about being a squirrel causes an animal to eat bugs, and that the fact that the something is an essence adds no explanatory power. On one hand, we hesitate to agree fully with his claim because no convincing evidence has been presented to rule out psychological essentialism. On the other hand, we agree that causal laws were most likely a critical driving factor for these results. After reviewing empirical support for the causal status hypothesis, we will also review how previous studies can be accounted for in terms of the causal status hypothesis.

IV. Main Empirical Results Supporting the Causal Status Hypothesis

The most direct test of the causal status hypothesis is reported in Ahn et al. (in press). Participants in their Experiment 1 read about three characteristic features of a target category (e.g., animals called ''roobans'' tend to eat fruits, have sticky feet, and build nests on trees). Participants in the control condition received no further information about the target category. In contrast, participants in the experimental condition were told that one feature tends to cause the second feature, which in turn tends to cause the third feature (e.g., eating fruits tends to cause roobans to have sticky feet because the fruit sugars are secreted through pores on the undersides of their feet, and that having sticky feet tends to allow roobans to build nests on trees because they can climb up trees easily with their sticky feet). Finally, all participants were presented with three exemplars, each of which had two features characteristic of the target category and one noncharacteristic feature (e.g., an animal that likes to eat worms, has feet that are sticky, and builds nests in trees). Participants were then asked to rate how likely it was that this animal is a member of the target category (e.g., a rooban).

For participants in the control condition, likelihood ratings remained constant regardless of which feature the exemplar animal was missing. However, in the experimental condition, the likelihood judgments varied

as a function of the missing feature's causal status. Specifically, when an exemplar was missing the target category's fundamental cause in the causal chain, the mean likelihood of being a target category member was lower than when an object was missing its intermediate cause in the causal chain, which in turn was lower than when an object was missing its terminal effect. That is, the deeper a feature was in a causal chain, the more central it was in a categorization judgment.

One alternative explanation for the above results is that participants might have assumed that category validities of features vary as a function of causal status. Whereas category validities of features in a category can be an important determinant of feature centrality as discussed earlier, high category validities do not necessarily entail high causal status. For instance, "being purple" in purple seedless grapes has a category validity of 1, but it has low causal status because it does not cause any other features in that category. Note, however, that many cause features in basic-level categories also tend to have high category validities. For example, birds' genetic codes are causally central, and they are also present in all birds. This observation suggests that, due to this correlation in real-life categories, participants in Experiment 1 of Ahn et al. (in press)—which did not explicitly specify category validities of features—might have assumed that the cause features must have high base rates. Thus, it could be argued that these inferred high category validities of the cause features determined their conceptual centrality rather than their causal status per se.

Experiment 2 of Ahn et al. (in press) controlled for this problem. In this experiment, participants observed 12 representative samples of the target category before they made judgments on transfer items (and before those participants in the causal condition received information about the causal relations). The actual category validities of three characteristic features of a target category were held constant in these samples. After observing 12 samples, participants were asked to judge the frequency of each feature. Following these frequency judgments, participants performed the identical tasks as in the first experiment, with half the participants in the causal condition and the other half in the no-causal control condition. The results showed that all participants judged frequencies fairly accurately; that is, there was no significant difference between features. Yet, the causal status effect was replicated.

Ahn et al. (in press, Experiment 3) also used a different categorization task and found the causal status effect. Imagine Jane who is depressed because she has low self-esteem. Now determine who should be categorized with Jane: Susan who is depressed because she has been drinking, or Barbara who is defensive because she has low self-esteem. Participants in this study who received sets of items similar to this example preferred to

categorize objects and persons based on a matching cause (e.g., categorizing Jane and Barbara together) rather than a matching effect (categorizing Jane and Susan together).

Thus, the basic causal status effect was documented in tasks involving both category likelihood judgments and free sorting. Furthermore, the causal status effect occurs above and beyond any confounding effect of category validities. The existence of the effect seems clear, but how far-reaching are its ramifications? In the next section, we will demonstrate how feature centrality phenomena shown in previous experiments can be explained by the causal status hypothesis.

V. Related Phenomena Accounted for by the Causal Status Hypothesis: Natural Kinds versus Artifacts

As discussed earlier under the content-based approach to feature centrality, previous studies (e.g., Barton & Komatsu, 1989; Gelman, 1988; Keil, 1989; Rips, 1989) have shown that different features are central for natural kinds and artifacts: in natural kinds internal or molecular features are more conceptually central than functional features, but in artifacts functional features are more conceptually central than internal or molecular features. As discussed earlier, it is tempting to take the content-based approach based on these findings and conclude that there is something inherently special about molecular features for natural kinds and functional features for artifacts.

In contrast, Ahn (1998) argued that the mechanism underlying this phenomenon is the causal status effect. That is, in natural kinds, internal/molecular features tend to cause functional features (e.g., cow DNA determines whether or not cows give milk) but in artifacts, functional features determine its compositional structure (e.g., chairs are used for sitting, and for that reason, they are made of a hard substance).

Experiments 1 and 2 in Ahn (1998) examined the real-life categories used in previous studies (Barton & Komatsu, 1989; Malt & Johnson, 1992). Participants were asked about causal relations among features within the same category. At the same time, they judged the centrality of features as measured by the degree to which a feature impacts categorization when that feature is missing. It was found that across natural and artifactual kinds, the more features any particular feature caused, the more influential the feature was in categorization. In addition, Ahn (1998) directly manipulated the causal status of features using artificial stimuli, and showed that when a compositional feature caused a functional feature, a compositional feature was more influential in categorization of both natural and artifactual

kinds, whereas the opposite was true when the causal direction was reversed. Thus, the results suggest that the more fundamental determinant of feature centrality is not the specific content of features but rather the causal role that a feature plays. This way, the causal status hypothesis offers a more fundamental account than has previously been put forth for why people might treat natural kinds and artifacts differently.

A. Category Use Effect

We speculate that additional existing findings in the literature may be viewed as having been mediated by a causal status effect. For example, Ross (1996, 1997, 1999) found that using category knowledge to perform a nonclassification task resulted in changes in feature weighting on a subsequent classification task. In a series of experiments in Ross (1997), participants learned to categorize hypothetical patients as having one of two novel diseases, on the basis of four relevant symptoms. During this training phase, participants also learned to prescribe one of two treatments for each of the diseases, on the basis of two relevant symptoms for each disease. For instance, four symptoms of buragamo disease were a fever, a runny nose, dizziness, and abdominal pain, and buragamo patients with a fever were treated by lamohillin, and buragamo patients with runny noses were treated by pexlophene. Ross found that although all four symptoms perfectly predicted the disease, symptoms that predicted the treatment to be prescribed were treated as more predictive of the disease too.

We conjecture that this result might have been obtained because the treatment-relevant symptoms were perceived to be causally central to the disorder. For instance, when a patient with a fever is always treated by lamohillin, as in Ross's (1997) study, this implies to participants that lamohillin acts specifically upon a fever. A treatment that is able to cure a disease is generally most likely to act upon the causal symptoms of the disease, not the peripheral effects. For example, one would not expect an ear infection caused by bacteria to be treated effectively with medicines that simply suppress symptoms (e.g., pain killers). Instead, to cure the ear infection, antibiotics that act directly upon the bacteria should be given. Moreover, it may be that treatment-relevant symptoms were considered by Ross's participants to be causally central in a broader sense, in that they determined which treatment plan to adopt. Thus, we suggest that Ross's category use effect may be mediated by a causal status effect. That is, within our causal status framework, the fact that Ross's participants felt that treatment-relevant symptoms were most important in diagnosing the disease can be traced to the experiment's implication that these symptoms are the most causal. Thus, it would be extremely difficult to pit the category use effect

against the causal status effect because whenever a feature is relevant for a certain use, it becomes a causal feature in that the feature determines the use of the category.

The causal status hypothesis also appears to be compatible with findings involving the use of categories in problem solving. For example, Chi, Feltovich, and Glaser's (1981) classic experiment assessed the categorical representations of physics problems in advanced physics graduate students and professors (experts) and undergraduates who had just completed a semester-long course in mechanics (novices). They found that experts consistently based their categorization of physics problems on deeper physics principles, such as conservation of momentum or the work–energy theorem, whereas novices categorized based on the surface features of the problems, such as whether there was an inclined plane or pulley in the problem. Their findings seem to be related to the causal status hypothesis in that both indicate a tendency to not categorize based on surface features when knowledge about a deeper structure is available. That is, experts in Chi et al.'s (1981) study were aware of both the surface features and the deeper structure, but preferred to categorize based on the latter. Similarly, novices in Chi et al.'s study did not categorize the problems based on even more surface features, such as the number of lines in the problem (see Keil, Smith, Simons, & Levin, 1998, for a similar argument). It should be noted, however, that in Chi et al.'s study, participants were specifically asked to sort the problems ''based on similarities of solution'' (p. 124). Thus, it remains to be seen whether they would sort problems by deeper properties (i.e., deeper properties within their understanding) even without these instructions. Given that Ahn et al.'s (in press) Experiment 3 described earlier showed that people spontaneously create categories based on a matching causal feature rather than a matching effect feature, we would predict that both experts and novices will spontaneously sort problems based on the deepest cause known to them.

B. Appreciation for Intentionality

Recently, several developmental studies demonstrated that even young children appreciate intentionality behind external representations (e.g., drawings) more than their appearance in naming the representations. Bloom and Markson (1998) asked 3- and 4-year-old children to draw, for instance, a balloon and a lollipop. As one might expect, these drawings looked almost identical. Later, when the children were asked to name these drawings, the children named the pictures on the basis of what they had intended for them to depict. Similarly, Gelman and Ebeling (1998) showed that when the same picture was produced either intentionally or accidentally

(e.g., drawings had been intentionally created to be a bear versus somebody spilled paint), intentional representation led to higher rates of naming responses (e.g., ''bear'') than did accidental representation. Clearly, the intention behind drawings is a causal factor of the drawings' appearance. Thus, appreciation for the drawer's intention when naming drawings can be construed as an example of the causal status effect.

C. Developmental Studies

It has been shown that even children discriminate features when they categorize objects or use concepts. For instance, Keil (1989) demonstrated that fourth graders base their categorization on origins of animals (e.g., being born from another raccoon) rather than on perceptual appearance (e.g., black with a white stripe on the back). Gelman (1988) found that second graders are more likely to generalize internal parts of an animal (e.g., a spleen inside a rabbit), rather than functional features, to other instances of the same kind. Once again, these results seem to be an example of the causal status effect: origins or internal parts of animals can be thought of as determining the surface features, and therefore, they are more essential in children's categorization.

Unfortunately, however, previous studies did not provide direct evidence that children believe that essential features cause other features. This leaves open the question of whether children's preference for deeper features is derived from accumulated past knowledge, or whether it instead manifests a belief brought to the task of acquiring knowledge, such as the causal status effect. For instance, Keil's (1989) transformation experiments further showed the content-specific effect in that fourth graders can differ from adults when different stimulus materials were used. Fully half the fourth graders in these experiments thought that changing a tiger into a lion by injection given early in life would make it a lion, whereas most adults did not. Keil (1992) argued that this difference is due to the fact that understanding is knowledge dependent. That is, the argument is that the difference occurred because the fourth graders did not have the right sort of knowledge rather than because they lack a bias toward weighing causal features.

One way of testing this claim is to use artificial categories and directly manipulate causal relations among features to see whether causal status can determine feature centrality in children's categorization. Using this framework, Ahn, Gelman, Amsterlaw, Hohenstein, and Kalish (2000) found the causal status effect in 7- to 9-year-old children. In this study, adults and children learned descriptions of novel animals, in which one feature caused two other features. When asked to determine which

transfer item was more likely to be an example of the animal they had learned, both adults and children preferred an animal with a cause feature and an effect feature than an animal with two effect features. Thus, children at this age group do indeed show the causal status bias. It remains to be seen whether even younger children have the causal status bias.

D. Summary

The causal status effect seems to be prevalent in various aspects of categorization and our use of concepts. It explains why different features are central for natural kinds and artifacts, in that internal/molecular features tend to be causally central in natural kinds, whereas functional features tend to be causally central in artifacts. It may be the underlying mechanism for phenomena involving the use of categories in reasoning, including Ross's (1997) category use effect and Chi et al.'s (1981) classic experiment on experts' and novices' sorting of physics problems. It may mediate our appreciation for intentionality as demonstrated by Bloom and Markson (1998). Finally, it is able to explain one way in which children weight features in categorization. Given that the causal status effect appears to be present in a variety of phenomena, a way to implement this general effect computationally would provide the means to facilitate further research. In the next section, we review a model of the causal status effect that fulfills this purpose.

VI. Computational Modeling

The causal status hypothesis provides one well-defined way of constraining feature weights, within the framework of the theory-based approach to categorization. An example of a computational implementation of the causal status hypothesis is presented in Sloman, Love, and Ahn (1998). This model is based on general "dependency" relationships rather than being restricted only to causal relationships. (See the last section of this chapter for further discussion about the difference between these two.) The idea of the model is that the more other features depend on (e.g., are caused by, are determined by, are followed by, etc.) a feature *i,* the more central this feature *i* becomes to the concept. More specifically, the model states that conceptual centrality of a feature ($C_{i,t}$) is a function of dependency as follows:

$$C_{i,t} = \Sigma A_{ij} C_{j,t-1}, \tag{1}$$

where A_{ij} is the strength of a dependency link from feature j to feature i. This formula states that the centrality of feature i is determined at each time step by summing across the immutability of every other feature multiplied by that feature's degree of dependence upon feature i. For instance, suppose feature X causes feature Y, which causes feature Z, and the causal strengths of both relations are 3, and the initial value for feature centrality was an arbitrary value of, say, 1. After two iterations, the centralities of features X, Y, and Z become 16, 7, and 1, respectively. These qualitative differences in feature centrality predicted by the model are consistent with the results found in Experiments 1 and 2 of Ahn et al. (in press). This model is just one of many possible ways of computationally modeling the causal status effect. In fact, there are results that this particular model cannot account for, as we shall see later in the chapter.

VII. Moderating Factors

Although the causal status effect is robust, it is important to specify its boundary conditions. This section discusses various factors that can moderate the causal status effect. Another goal of this section is to account for apparent counterexamples to the causal status effect in terms of these moderating factors.

A. Influence of Other Determinants of Feature Centrality

At the beginning of this chapter, we reviewed different approaches to feature centrality. Similarly, Sloman et al. (1998) examined conceptual centrality (mutability, or the degree to which a feature in a concept can be transformed while maintaining the concept's coherence), category centrality (perceived relative frequency of a feature within a category), diagnosticity (evidence provided by a feature for one category relative to a set of categories), and prominence (how salient the feature seems to people when thinking about the category). In order to show that these are empirically dissociable, Sloman et al. (1998) asked participants questions designed to measure each of these centralities. For instance, conceptual centrality was measured using a number of questions, including a measure of surprise (e.g., How surprised would you be to encounter an apple that did not grow on trees?), ease of imagining (e.g., How easily can you imagine a real apple that does not grow on trees?), goodness of example (e.g., How good an example of an apple would you consider an apple that does not ever grow on trees?) and similarity to an ideal (e.g., How similar is an apple that doesn't grow on trees to an ideal apple?). In a factor analysis, they found that questions selected to measure the same centrality loaded on the same factor, and

questions measuring different centralities loaded on different factors. (One exception to this result was that the questions measuring the conceptual centrality and those measuring categorical centrality loaded on the same factor. Sloman and Ahn [1999] later addressed this issue, as is discussed in a later section of this chapter.) Thus, in general, these results show that different kinds of feature centrality are empirically dissociable.

One intuitive example for understanding this phenomenon is illustrated in a Calvin and Hobbes cartoon in Figure 1. Calvin has just lost his beloved stuffed tiger Hobbes, so he is working on a "lost" ad. The ad should be designed to facilitate a perceptual categorization task, but instead of describing perceptually central features, Calvin listed Hobbes's features that are conceptually central to him, making this cartoon amusing.

Among these various types of feature centrality, the causal status effect is related to conceptual centrality; the measures of surprise and goodness of example are similar to the ones used in Ahn et al. (in press). Furthermore, Sloman et al. found that conceptual centrality was the only type of centrality that correlated reliably with a feature's status in a concept's dependency structure (i.e., centrality predicted by Equation 1) across various levels of categories (see also Ahn & Sloman, 1997; Sloman & Ahn, 1999). As explained earlier, Equation 1 is one way of computationally modeling the causal status effect. Thus, these results suggest that the causal status effect is likely to be most pertinent to tasks involving conceptual centrality, and may be less influential in tasks such as perceptual categorization (which would be determined by perceptual prominence) or discrimination tasks (which would be determined by diagnosticity of features).

B. Levels of Abstraction and Name versus Conceptual Centrality

The fact that there are different types of determinants for feature centrality that act on different types of categorization tasks can account for a set

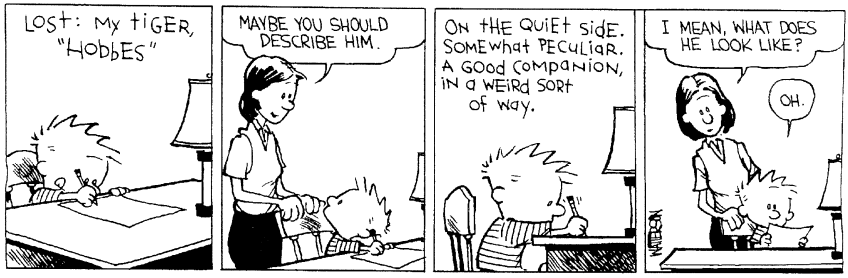


Fig. 1. Calvin and Hobbes cartoon. (© 1988 Watterson. Reprinted with permission of Universal Press Syndicate. All rights reserved.)

of examples that apparently contradict the causal status hypothesis. For instance, although ''being purple'' is not causally central in our concept of purple seedless grapes, it is a necessary feature for the category of purple seedless grapes. Similarly, in our concept of pine trees, ''having needles'' is not as causally central as ''having a trunk,'' but we would not call an object a pine tree if it did not have needles. These apparent counterexamples can be explained if we understand the distinction between name centrality and conceptual centrality.

Sloman and Ahn (1999) argued that the name centrality of a feature (i.e., the feature's power to determine the appropriateness of a category label for an object) is a function of feature frequency in a category, whereas conceptual centrality is a function of features' relational structure (e.g., causal status). In most cases, the two types of centrality are highly correlated (e.g., ''having wings'' for birds). But oftentimes, we discriminate between categories that do share conceptually central properties, and assign different names to them. Such cases tend to occur at a high level of specificity. For instance, grapes have conceptually central properties (e.g., is juicy, grows on vines), and we subdivide them into purple versus green seedless grapes. By virtue of specifying a distinct category, the name must refer to one or more features that make the distinction but are not necessarily conceptually central (e.g., ''being purple'').

The counterexamples we provided at the beginning of this section can be explained by this framework. For instance, ''having needles,'' which has low causal centrality, has high name centrality (i.e., we would not *name* an object a pine tree unless it had needles), but it still has low conceptual centrality (i.e., it is not difficult to imagine an object that is in all ways like a pine tree except that it did not have needles). Sloman and Ahn (1999) showed that participants also shared this intuition. The question format used to measure name centrality was, ''Suppose an object is in all ways like an X except that it does not have feature Y. How appropriate would it be to call this object an X?'' The question format used to measure conceptual centrality was, ''How easy is it to imagine an object that has ALL the features of an X except that it does not have feature Y?'' Sloman and Ahn (1999) found that features that distinguish categories at a specific level have high name centrality but low conceptual centrality. In addition, they found that few features of the same concept depended on these specific-level features, consistent with the causal status hypothesis. Hence, the above ''counterexamples'' of the causal status effect become moot when we distinguish name centrality from conceptual centrality.

C. The Case of a Common-Effect Structure

In the previous two sections, we focused on the fact that different determinants of feature centrality are dissociable. However, real-life categorization

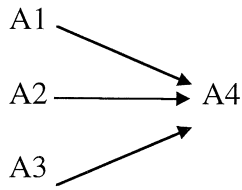


Fig. 2. Common-effect causal structure.

often involves multiple categorization processes. For instance, names of concepts might influence the way we think about concepts (e.g., Davidoff, & Davies, Roberson, 1999). Or when we think about object concepts, we might also visualize the object. Under such circumstances, different determinants of feature centrality can simultaneously act on categorization. For instance, a feature with high causal status but low category validity might end up having mediocre feature centrality. When other determinants of feature centrality heavily favor the effect feature over its cause, it is reasonable to expect that these multiple influences can override the causal status effect. Alternatively, the causal status effect can be enhanced when other determinants of feature centrality favor the cause feature over its effect.

This discussion about the influence of other determinants of feature centrality on the causal status effect is useful in analyzing a case that has been frequently posed as a counterexample to the causal status hypothesis. The case we will examine in this section is a common-effect structure, in which multiple causes have a single effect in common (see Figure 2 for an abstract structure). We will first discuss real-life examples involving a common-effect structure, followed by Rehder and Hastie's (1997) experimental results.

1. Real-Life Example

Some real-life concepts have a common-effect structure.[4] One example is a syndrome: a syndrome is a cluster of symptoms with some predictive value in terms of treatment and prognosis without a single known etiology, and thus it tends to have multiple causes for a common set of symptoms. Another example is pneumonia: it is a serious infection or inflammation of the lungs which can have over 30 different causes, including bacteria, viruses, chemicals, or even inhaled objects like peanuts or small toys.

[4] The fact that we have concepts containing common-effect structures might be considered as a counterexample to the causal status hypothesis. That is, why would we have any category that was not created based on a single cause? In section VI. E Domain Generality, we will discuss the issue of why people sometimes create concepts with a common-effect structure.

In these examples, a cause seems less crucial than its effect in determining category membership. For instance, in determining whether or not a person has pneumonia, the cause of the lung inflammation is less important than the lung inflammation itself.[5] As discussed above, however, feature centrality may be determined by other factors, such as category validity. In the case of pneumonia the effect feature has a category validity of 1 (i.e., all patients with pneumonia have lung inflammation), whereas each of the cause features has a much lower category validity because there are many possible causes for pneumonia. For the feature centrality of *each* of these causes, any amount of the causal status effect that occurs is canceled out due to its low category validity. Furthermore, in the case of the common-effect structure, the effect feature has a greater advantage over the cause features in terms of the number of relations in which it participates. Gentner and her colleagues (e.g., Gentner, 1989) have suggested that relational features are more important than isolated features in analogical reasoning. If the number of relations in which a feature participates also determines feature centrality, then an effect feature in a common-effect structure can outweigh one of the cause features because the effect feature participates in more relations than any one of these cause features.

Because these multiple counteracting forces act upon the cause features in a common-effect structure (low category validity, fewer relations to participate in), the overall centrality of one of these causes in a common-effect structure can end up becoming lower than the overall centrality of its effect feature. Exactly how all these potential determinants for feature centrality might be combined awaits more future research. With these examples we attempted to argue in this section that a failure to observe the causal status effect in common-effect structures as in the above situations is not evidence for a counterexample to actual occurrence of the causal status effect because in these situations, there are a disproportionate number of counteracting forces.

Consider again the case of pneumonia: the causal status hypothesis is supported when these counteracting forces are removed. Indeed, the American Lung Association (1998) states explicitly that ''pneumonia is not a single disease,'' precisely because there are many causes. Instead, they break the disease down into various subcategories including bacterial pneumonia, viral pneumonia, mycoplasma pneumonia, and so on, based on the type of cause, just as the causal status hypothesis would predict. Furthermore, note

[5] In diagnosing pneumonia, causes for lung inflammation are also crucial information because they determine different treatment plans. However, in simply determing whether or not someone has pneumonia, causes would not be as critical as lung inflammation per se because there are many potential causes, whereas lung inflammation is a necessary feature for pneumonia.

that each subtype no longer has a common-effect structure, and within each subtype, the category validity of the cause feature and the category validity of the effect feature become equal. For instance, the probability of *having lung infection* given that a person has bacterial pneumonia is the same as the probability of *being infected with pneumonia bacteria,* given that a person has bacterial pneumonia (i.e., 1). In addition, both of these features participate in the same number of relations. Now that all the other counteracting forces are removed, the cause feature (e.g., being infected with pneumonia bacteria) may be measured to be more central than the effect feature (e.g., having a lung infection). For instance, a patient infected with pneumonia bacteria, but who has not yet developed lung inflammation (missing effect) is at least more likely to be considered as specifically having bacterial pneumonia than a patient who has lung inflammation from inhaling a peanut (missing cause).

2. *Rehder and Hastie's Study (1997)*

Rehder and Hastie (1997) posed what seemed to challenge our causal status hypothesis, reporting that a cause feature in a common-effect structure was less central than its effect. Moreover, our argument in the previous section might not appear to hold in this case because they equated the category validities of features in their training exemplars. Thus, this study merits a detailed discussion in the current chapter. Readers who are not interested in the details of this issue can safely skip to the next section.

Participants in Rehder and Hastie (1997) learned novel categories' features on four dimensions (A1, A2, A3, and A4 in Table I), each with two

TABLE I

ITEMS USED IN REHDER AND HASTIE (1997) AND IN OUR REBUTTAL EXPERIMENT[a]

	Missing-effect items					Missing-cause items				
Task	A1	A2	A3	A4	Mean	A1	A2	A3	A4	Mean
Categorization	1	1	1	0	2.87	0	1	1	1	6.65
	1	1	0	0	3.41	0	0	1	1	5.20
	1	0	0	0	3.71	0	0	0	1	2.09
Inference	1	1	1	?	2.85	?	1	1	1	4.35
	1	1	0	?	3.17	?	0	1	1	3.95
	1	0	0	?	4.56	?	0	0	1	2.72

[a] Items for the Categorization task constitute a subset of items used in Rehder and Hastie's study (1997) and a complete set of items used in our experiment. The Inference task was used only in our experiment. The mean ratings are results from our experiment. In the common-effect condition, A4 was a common effect for A1, A2, and A3.

values (0 and 1 in Table I). Prototype values were 1s. In the Common-Effect condition, 1s on A1, A2, and A3 were described to cause 1 on A4, as in Figure 2. Participants learned a novel category by observing training exemplars, where 1s occurred equally frequently across the four dimensions. In the test phase, participants rated the category membership likelihood of 16 exemplars that were all possible combinations of binary values on the four dimensions. (Table I lists 6 of the 16 categorization test exemplars in pairs of missing-cause and missing-effect items with 3, 2, or 1 prototype values.) Their regression analysis, performed on average participants' ratings on category membership, showed that in the Common-Effect condition, the weight associated with A4 (i.e., effect) was higher than that associated with each of A1, A2, and A3 (i.e., cause), presumably contradicting the causal status hypothesis.

It should be noted that in Rehder and Hastie (1997) there is no mention of whether the participants were informed that the exemplars they saw during training were representative samples of each category. These would have been critical instructions to add, because the observed category validities of features during training conflicted with the category validities that one would expect from the common-effect structure. More specifically, we argue that due to the common-effect structure, participants expected that the effect feature should have a higher category validity than each of the causes, even if they happened to observe the same category validities across all features. That is, if they believe that A1 can cause A4, A2 can cause A4, and A3 can cause A4, it is reasonable to expect that A4 would be more frequent than A1, A2, and A3, and that they might have accidentally seen a subset of category members with a different distribution. Without specific instructions that the training exemplars were representative of the entire category, participants could have discounted evidence in favor of their expectations derived from causal background knowledge. In the following experiment we tested whether participants indeed have different expectations about category validities of cause-and-effect features in a common-effect structure.

For the sake of simplicity, we presented only the three pairs of missing-cause and missing-effect exemplars in Table I, because they represent all 16 test exemplars used by Rehder and Hastie, (1997) assuming equal salience of dimensions without the causal background knowledge (which was ensured through various measures they took). For instance, although we did not use 1011 and 1101, they are fundamentally redundant with 0111 because all three are missing-cause items with three prototype values.

As in Rehder and Hastie (1997), we first taught participants the common-effect causal background knowledge and then measured the likelihood of category membership for each of the six test exemplars. In addition, we

included an inference task which measured the likelihood of a nonprototype value (Value 0) for the missing dimension, given values on other dimensions as specified in Table I for the inference task. In this way, we measured the participants' expected category validity of value 0 on each of the dimensions, given the causal background knowledge of the common-effect structure.

The prediction was that measures from the categorization task would correlate with the measures of the inference task. For instance, Rehder and Hastie found that 1110 (missing-effect) yielded lower ratings from the categorization task than 0111 (missing-cause), and we argue that this occurred because given the common-effect structure, A4 = 0 in 111? is expected to be less likely than A1 = 0 in ?111. People would judge A4 = 0 in 111? to be very unlikely because all the causes have a value of 1. In contrast, A1 = 0 in ?111 is somewhat likely because there is no reason to expect 1 on A1. Similar predictions are made for 1100 and 0011.

However, 1000 and 0001 would be expected to lead to the opposite result. That is, A4 = 0 in 100? would be somewhat unlikely given that A1 = 1 and is expected to cause A4 = 1, but given that A2 and A3 = 0, it is not unreasonable to predict A4 = 0. In contrast, A1 = 0 in ?001 would be very unlikely because neither A2 nor A3 = 1, so A1 must have a value of 1 to account for the fact that A4 = 1. That is, the category validity expected for a value of 0 on the missing dimension is lower in the missing-cause item (?001) than in the missing-effect item (100?). Indeed, Rehder and Hastie (B. Rehder, personal communication, September 22, 1997) found that the categorization likelihood judgments for 0001 (missing-cause; M = 20.7) was lower than that for 1000 (missing-effect; M = 29.5), which runs counter to their results from the overall regression analyses.

The results of our experiment supported our predictions, as shown in Table I under ''Mean'' ratings. First, the categorization task results replicated the results from Rehder and Hastie's study. That is, missing-effect items led to lower categorization judgments than missing-cause items only from the 1110 versus 0111 and 1100 versus 0011 pairs. For the 1000 versus 0001 pair, the missing-cause item led to lower categorization judgments than the missing-effect item. More importantly, the directions of these results are also mirrored by the corresponding inference problem results. The correlation between the average likelihood judgments for the six categorization task problems and the six inference task problems was significant (r_s = .83; $p < .04$).

This experiment thereby provides support for the hypothesis that in a common-effect structure, missing effect features presumably mattered more than missing cause features in Rehder and Hastie (1997) because having a noncharacteristic value for the effect dimension in their stimulus materials is less likely than having a noncharacteristic value for cause dimensions.

Although the training exemplars equated for the category validities of cause and effect features in their materials, the causal relations among features were structured in such a way that effect features are more likely to be expected to occur in most exemplars (as shown in the inference task of this study). Based on this imbalance between the effect and the cause features, we submit that their results do not necessarily contradict the causal status hypothesis.

D. Plausibility of Causal Beliefs

Thus far, we have not considered how the causal status effect might be moderated by how plausible people find the causal relations in the first place. We believe that plausibility does moderate the causal status effect, such that low plausibility reduces the effect. In general, the causal beliefs that people have about concepts come in various strengths. People might strongly believe, for instance, that bottles have a bottom so that they can contain liquid, or that robins are robins because they were born from another robin. However, that Echinacea contains a substance that bolsters the human immune system, or that yellow furniture increases children's emotional intelligence might be implausible to some (or most) people. It seems reasonable to expect that the amount of the causal status effect would be moderated by the degree to which the causal relations among features are plausible. That is, the more plausible the causal relations between two features are, the stronger the causal status effect should be from these two features.

From among many possible ways of manipulating causal plausibility, Ahn et al. (in press, Experiment 6) chose to examine the effect of compatibility between old and new causal background knowledge. Most people would believe that viral infections or genetic defects cause symptoms in disorders, rather than the other way around. Because of this old causal background knowledge, the statement that "Virus XB12 causes a low insulin level" is more plausible than the statement that "a low insulin level causes infection of Virus XB12."[6] The stimulus materials that conformed to laypeople's existing knowledge were developed for the Canonical condition (similar to the first example in the above), and the ones that conflicted with it were developed for the Reverse condition (similar to the second example in the above). Indeed, the results showed that the causal status effect was much more pronounced in the Canonical condition than in the Reverse condition.

This effect of the plausibility of causal relations on the causal status effect suggests one interesting possible difference between experts' and novices'

[6] There are exceptions to this (e.g., AIDS), but in general cases, this difference in plausibility between the two types of items seems to hold up, as was confirmed in Ahn et al.'s (in press) pilot study.

categorization in their domain of expertise. Novices who have just completed initial training in a certain domain might not be confident in their background knowledge. They might not yet have internalized newly learned theories, and they have not had the chance to develop their own theories. Experts, on the other hand, might have a better understanding of the underlying mechanisms of the concept, which has been shown to be an important factor in increasing the plausibility of causal relations (e.g., Ahn, Kalish, Medin, & Gelman, 1995). In addition, experts, after years of practicing in their domain, might have developed their own idiosyncratic theories about, for instance, how symptoms are causally related or how different components of a machine interact. If so, the causal status effect might be more pronounced in experts than in novices. We are currently testing this issue in the domain of psychological disorders. (See section IX. Application for more detail.)

E. Domain Generality

Another potential moderating factor for the causal status effect is domain. Schwartz (1979) argues that nominal kinds, such as ''white things,'' are different from natural kinds, such as ''gold,'' in that nominal kinds do not have essences. According to Schwartz, nominal kinds are conventionally established and therefore, if the criterial properties change, it no longer belongs to the category. For instance, if an object is stained with mud, it no longer belongs to the category of ''white things.'' In support of the idea of domain differences in essentialism, Diesendruck and Gelman (1999) found more essentialistic, all-or-none categorization of natural kinds than of artifacts. Thus, if psychological essentialism is responsible for the causal status effect, it might be predicted that a stronger causal status effect would occur for natural kinds than for nominal kinds.

The results obtained to date, however, are somewhat at variance with that prediction because no significant differences for the causal status effect have been found across various domains, including diseases and symptoms, artifacts, natural kinds, and social situations (Ahn, 1998; Ahn et al., in press). However, these studies did not systematically test the issue of domain generality by using randomly selected categories and features.

In particular, domain differences might occur when people create new categories for the following reason. Causal chains can be infinite, and an explanation must stop at some point. For natural kinds, this point might be their essences, or the deepest cause known to people. For nominal kinds, it is up to humans to select a stopping point based on their goal in creating these categories.

Thus, for categories that people treat as natural kinds (i.e., kinds that naturally occur in the world), we would expect people to create new catego-

ries as deeper underlying causes are revealed. For instance, to laypeople, pneumonia might be a single disease, because laypeople do not know that there are multiple possible causes for it (thus, from a layperson's perspective, pneumonia does not have a common-effect structure.) Experts, on the other hand, do not treat pneumonia as a single disease, and partition pneumonia into more specific categories based on its different causes.

In contrast, in creating nominal kinds, one can choose any criteria that meet one's functional goal for the category, ignoring underlying causes. This is because nominal kinds are by definition conventionally fixed. For instance, the selection criteria for new orchestra members usually focus on the quality of playing (an effect) rather than the details of how the person was trained (a cause). The orchestra category is a nominal kind, and the creator of the category can therefore intentionally set up such criteria, in order to meet their goal of getting the best performers. Similarly, one might create an ad-hoc category based on a goal, such as, ''things to take out of a house in case of fire,'' without further querying about why one might have such a goal.

F. Fast versus Slow Judgments

All of the experiments showing support for the causal status hypothesis involved slow judgments, where participants responded at their own pace. It is still an open question as to whether the causal status effect would be manifested in a time-pressured task, and as to whether it can account for some of the time course differences found in the categorization literature, such as reaction time differences in responding to typical versus atypical instances.

There are reports of influence of background knowledge even under a very short response deadline (e.g., Lin & Murphy, 1997; Palmeri & Blalock, 2000). Lin and Murphy's (1997) studies are of particular relevance to the causal status hypothesis. Participants first learned the following background knowledge about novel categories. For participants in Group A, ''tuks'' were said to be animal-catching devices (Category A, henceforth), whereas for participants in Group B, they were said to be pesticide-spraying devices (Category B, henceforth). Although both groups saw identical pictures of the learning exemplars, features that were crucial to the function of Category A were different from features that were crucial to the function of Category B. After learning exemplars, participants were presented with, among various items, an item retaining the crucial part of Category A while omitting that of category B (consistent A item), and an item retaining the crucial part of Category B while omitting that of Category A (consistent B item). They found that, in general, Group A responded positively to

Consistent A items and negatively to Consistent B items, whereas Group B showed the opposite pattern. Most interestingly, they found the same pattern of results even when the stimuli were presented only for 50 ms followed by a mask. Note that, consistent with the causal status hypothesis, the features that had a greater impact in this study were the ones that had causal relevance to the function of each category. One reason to be cautious about generalizing Lin and Murphy's study to the causal status hypothesis is that their task measured perceptual recognition of stimuli, rather than directly measuring conceptual centrality per se. Future research can further examine this issue.

G. Summary

In this section, we have reviewed some moderating factors for the causal status effect. In general, if a sufficient number of other centrality-determining factors are pitted against the causal status effect, it may be difficult to observe in experimental measures. These other determinants of centrality include category validity and cue validity (diagnosticity), in addition to conceptual centrality, which we have argued embodies the causal status hypothesis most closely. We proposed that in many cases, apparent counterexamples to the causal status effect are actually instantiations of the case in which other factors are overwhelmingly pitted against it. In a study of the common-effect causal structure, we demonstrated some direct evidence for this proposal. In addition, low plausibility of the causal relations appears to diminish the causal status effect. When creating categories, domain differences appear, such that natural kind categories may be based on essences, whereas artifact categories are based on conventionally fixed, functional goals. Finally, we speculated that the causal status effect may not be affected much by time pressure, given related findings such as Lin and Murphy's (1997). Thus, this section has addressed concerns about counterexamples to the causal status effect and laid out some of its boundaries. In the next section, we take a step in the opposite direction, explaining what predictions the causal status hypothesis may make for features in a category that exist outside the main causal structure.

VIII. Causal versus Isolated Features

The causal status hypothesis concerns the difference between features within the same causal structure. What about the difference between the centrality of features that participate in a causal structure and the centrality of features that are not involved in any causal relationships with other

features? For instance, consider four symptoms of a mental disorder, A, B, C, and D, where the first three symptoms form a causal chain ($A \rightarrow B \rightarrow C$) and the last symptom, D, is isolated. Would D affect diagnosis as much as the most fundamental cause (A), the intermediate cause (B), or the most terminal effect (C)? Alternatively, could it be that D affects diagnosis even less than C? Previous studies testing the effect of causal status of features on categorization have not examined this issue, although it is important to do so because real-life concepts sometimes have isolated features. For instance, most people would find it difficult to explain why tires are black and how this feature is causally related to other features of a tire.

Some clues to answering this question come from the induction literature. According to Gentner's (1989) structure-mapping theory, relational features (statements taking two or more arguments; for instance, *x* is smaller than *y*) are more important than attributes (statements taking only one argument; for instance, *x* is yellow) in analogical inference. For instance, in the analogy "an atom is like the solar system," attributes such as "yellow," "hot," and "massive" for the sun are not useful in making the analogy and are therefore discarded. However, relational features such as "more massive than," and "revolves around" can be used to draw the analogy that electrons revolve around the nucleus in an atom as planets revolve around the sun in a solar system. Lassaline (1996) provides evidence for Gentner's theory in category-based induction. In a similar vein, Billman and her colleagues have shown that it is easier to learn a rule that links two features when the link is part of a system of correlations than when it occurs in isolation (e.g., Billman, 1989; Billman & Knutson, 1996). Given these findings, we hypothesized that features related to other features would be given greater weight in categorization than isolated features.

In our experiment (Kim & Ahn, 1999), we developed six artificial mental disorders, each comprised of six symptoms taken from the *Diagnostic and Statistical Manual of Mental Disorders, 4th edition* (*DSM-IV*) (APA, 1994).[7] For example, the six symptoms of "Methinismus" were (a) intense fear of gaining weight, (b) recurrent unjustified suspicions of the spouse's infidelity, (c) making fake suicide attempts, (d) perfectionism that interferes with task completion, (e) unreasonably scorning authority, and (f) exaggerated startle response. No two symptoms within an artificial disorder were taken from the *DSM-IV* (APA, 1994) disorder.

Each participant received two disorders in the First-Cause condition (the first three symptoms were causally connected), two in last-cause condition

[7] We were particularly interested in the domain of psychological disorders because these studies were carried out as a part of a larger project to understand expert–novice differences in the domain of psychological disorders. See section IX. Application for preliminary data on this.

(the last three symptoms were causally connected), and two in the No-Cause condition (none of the symptoms was causally connected). When the symptoms were causally related, they also received a plausible explanation (e.g., ''An intense fear of gaining weight causes these patients to have recurrent unjustified suspicions of their spouses' infidelity, because they fear having become fat and unattractive to their spouses.''). When the symptoms were not causally related, the symptoms were given a common grouping factor in order to make the length and the saliency of symptoms equivalent to the causal conditions[8] (e.g., ''At a recent conference on mental illnesses, a speaker talking about intense fears of gaining weight was praised because she remained collected when the microphone broke. At a recent conference on mental illnesses, a speaker talking about recurrent unjustified suspicions of the spouses' infidelity was praised because she had charisma. At a recent conference on mental illnesses, a speaker talking about making fake suicide attempts was praised because she handled the audience's questions well.'').

After receiving descriptions of six symptoms of each disorder, participants answered questions in the format of ''if a patient is in all ways like a typical person with [disorder X] EXCEPT that he or she does NOT have [symptom Y], does the patient have [disorder X?]'' on a scale of 0 (definitely no) to 100 (definitely yes).[9] For clarity of presentation, we inverted these scores in Figure 3 so that the higher the number is, the more conceptually central the symptom is to the disorder.

Figure 3 summarizes the results. In the No-Cause condition, the six symptoms did not differ. However, when the same symptoms were causally related, the causal depth determined the membership likelihood (i.e., In the First-cause condition, $A > B > C$; In the Last-cause condition, $D > E > F$), replicating Experiment 1 in Ahn et al. (in press). More importantly, isolated symptoms were significantly less central than even the most terminal effect in a causal chain (i.e., in the First-cause condition, $F < C$; in the Last-Cause condition, $C < F$).

These results are comparable to findings in Wisniewski's study (1995). He taught participants pairs of artifact categories and asked them to classify novel exemplars, manipulating whether they were told the artifacts' functions or not during training. He found that when people knew an artifact's function, they were more likely to classify new exemplars based on the exemplar features' relevance to that function than when they did not know the artifact's function. That is, causally isolated features (i.e., features that did not have any functional relevance) were less central to the concept

[8] The same pattern of results were obtained when no common factor was given in the control condition (Kim, 1999).

[9] The background knowledge was available to participants during this task.

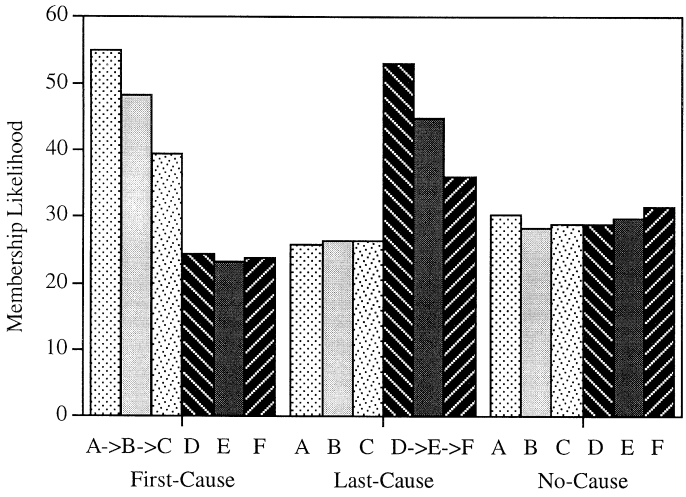


Fig. 3. Results from Kim (1999, Experiment 2).

than causal features. From this perspective, Wisniewski's results might have been due to the causal status of features rather than something inherent in the relevance of function in categorizing artifacts.

More recently, Gopnik and Sobel (1999) showed that 2-, 3-, and 4-year-old children categorize objects on the basis of causal power rather than similarity in appearance. Children saw that an object labeled as a "blicket" would set off a machine. When asked to select another blicket, they were more likely to choose an object that also set off the machine than an object similar in appearance. Because there was no clear causal connection between appearance and being able to set off the machine, these results demonstrate that children gave more weight to causal features than to isolated features (i.e., perceptual appearance) in categorization, consistent with our findings.

The finding that features participating in a relation are judged to be more central than isolated features have some implications for the model of Sloman et al. (1998). This model currently predicts that isolated features and terminal effect features will have the same weight, but this was not the case in the current study.

IX. Application

We now describe a study that is applied to real-life concepts. In this study, we were interested in the domain of psychological disorders, especially because of its potential implications for the *DSM-IV* (APA, 1994).

One of the earliest known classification systems of mental disorders, espoused by Kraepelin (1913), required necessary and sufficient features for the diagnosis of a disorder. This is an approach analogous to the bygone classical view in categorization research. Over the past century, this classical-view taxonomy of mental disorders has given way to the *DSM* system. The widely accepted *DSM-IV* (APA, 1994) was intended to reflect the later prototypical view of categorization, which allows for some flexibility (Barlow & Durand, 1999). For example, the prototypical patient with schizophrenia has five symptoms, but a presenting patient need only have two of those five symptoms for a diagnosis of the disorder. In addition, as in a standard prototype model of categorization, the *DSM-IV* represents disorders as lists of features that are not causally connected to each other. For example, clinicians using *DSM-IV* (APA, 1994) criteria would diagnose someone with obsessive-compulsive personality disorder if the patient has any combination of 4 symptoms out of a list of 8. Indeed, Medin (1989) has suggested that ''the *DSM-III-R* guidebook (APA, 1987) provides only a skeletal outline that is brought to life by theories and causal scenarios underlying and intertwined with the symptoms that comprise the diagnostic criteria'' (p. 1479).

A more serious challenge to the DSM system was put forth by Follette and Houts (1996). Their position is that the *DSM-IV* (APA, 1994), which claims to be atheoretical, is inadequate because it fails to provide ''an organizing theory that describes the fundamental principles underlying the taxonomy'' (Follette & Houts, 1996; p. 1120). The *DSM*'s purpose in not specifying underlying theories is to avoid battles between different theoretical schools as to which theories should be included or focused on in the manual. However, these authors argue that the advantage gained by such a solution is outweighed by disadvantages concerning the negative effect a taxonomy that is silent with respect to theory has on clinical research (Follette & Houts, 1996). Specifically, Follette and Houts contend that one major practical failure of the *DSM* system is that its lack of unifying theories makes research difficult and slow. They propose that the first step towards remedying this failing of the *DSM* is to strengthen those research programs based on multiple theories (Follette & Houts, 1996).

As the first step toward studying the nature of the theory-based categorization in the domain of psychological disorders, we tested laypeople's understanding of psychological disorders (Kim & Ahn, 1999). Unlike the assumption underlying the *DSM* system, laypersons seem to have theories about how the symptoms of a disorder are causally connected. For instance, consider three of the *DSM-IV* (APA, 1994) diagnostic criteria for anorexia nervosa: ''refuses to maintain minimal body weight,'' ''intense fear of gaining weight or becoming fat,'' and ''disturbance in the way in which one's

body weight or shape is experienced.'' It is easy to imagine a layperson thinking that people who have these anorexia nervosa symptoms ''refuse to maintain minimal body weight'' *because* they have an ''intense fear of gaining weight or becoming fat,'' and also that they have an ''intense fear of gaining weight or becoming fat'' *because* they have a ''disturbance in the way in which one's body weight or shape is experienced.'' Furthermore, we hypothesized that laypeople's causal theories would influence the way they perceive the importance of symptoms in their conceptualization of psychological disorders, as predicted by the causal status hypothesis.

In this experiment, we selected four mental disorders taken directly from the *DSM-IV* (APA, 1994) as stimuli. They included two Axis I clinical disorders (anorexia nervosa and major depressive disorder), and two Axis II personality disorders (narcissistic personality disorder and obsessive-compulsive personality disorder). The task was divided into two parts, a causal centrality task in which participants drew causal relations among symptoms within each disorder and assigned causal strengths to each relation, and a conceptual centrality task in which participants rated the importance of symptoms in each category. We hypothesized that each symptom is conceptually central to the extent that other symptoms are dependent on it (that is, to the extent that it is causally central). We will now describe the methods in more detail.

A list of ''criterial'' symptoms and ''characteristic'' symptoms was compiled for each of the four selected disorders. The criterial symptoms were taken from the list of diagnostic criteria for each disorder described in *DSM-IV*. We also selected additional characteristic symptoms that were not considered to be diagnostic criteria but were described as characteristics in the *DSM-IV* (APA, 1994).

For each criterial symptom, a question measuring conceptual centrality was developed, in the format of ''if a patient is in all ways like a typical person with X EXCEPT that he or she does NOT have the symptom of Y, does the patient have X?'' where X is one of the four mental disorders and Y is a symptom. The participant's answer was collected on a rating scale of 0 (definitely no) to 100 (definitely yes). A total of 30 such questions, presented in random order and blocked by disorder, served as the conceptual centrality task. For clarity of presentation, we inverted these ratings in the analysis reported below so that the higher the number is, the more conceptually central the symptom is to the disorder.

For the causal centrality task, participants were presented with both the criterial and characteristic symptoms,[10] and were asked to draw arrows

[10] We included the characteristic symptoms so that participants' causal structures could be measured as comprehensively and as accurately as possible (e.g., if a characteristic symptom X were not included, a criterial symptom Z could erroneously be measured as causally peripheral when, in fact, participants believed it to cause characteristic symptom X).

indicating causal relations among them. They then assigned numbers to each arrow indicating the strength of that causal relationship, on a scale of 1–5 (where 1 means ''X very weakly causes Y'' and 5 means ''X very strongly causes Y''). The order of the conceptual centrality task and the causal centrality task was counterbalanced across the participants.

The results showed that in general, the more causally central a symptom was, the more conceptually central it was. Figure 4 shows an example for the disorder anorexia nervosa. This figure presents averaged causal strengths among symptoms. (For simplicity of presentation, causal strengths lower than 1.0 are omitted from the figure.) Arrows point toward effect symptoms, and the symptoms circled with thicker lines are criterial symptoms. This figure also reports the mean conceptual centrality ratings, shown by the numbers within the circles. One interesting result to notice is that even among criterial symptoms, causal centrality seems highly variable. For instance, on average, participants believed that in anorexia nervosa, ''fear of being fat even when underweight'' causes many symptoms, including fear of eating in public, bingeing and purging, excessive dieting, and refusal to gain weight. However, ''absence of the period (in women) for 3+ menstrual

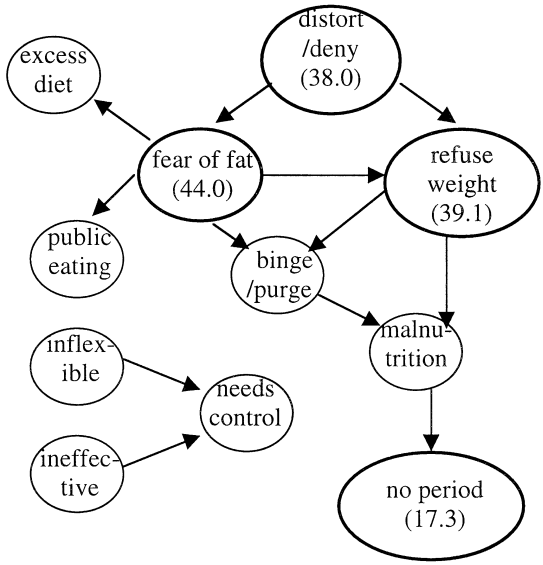


Fig. 4. Laypeople's mean causal dependency structure for the disorder anorexia nervosa. ''Distort/deny'' means ''Disturbed experience of body shape or denial of the problem'', ''fear of fat'' means ''Fear of being fat even when underweight'', ''refuse weight'' means ''Refusal to maintain body weight at or above minimal levels'', and ''no period'' means ''Absence of the period (in women) for 3 + menstrual cycles.''

cycles,'' another criterial symptom for anorexia nervosa, was not judged to cause any other symptoms in that disorder. More importantly, note that a conceptually central symptom (e.g., ''fear of being fat even when underweight'' in anorexia nervosa) is also causally central, and a conceptually peripheral symptom (e.g., ''absence of the period (in women) for 3+ menstrual cycles'') is also causally peripheral.

We quantified this result by implementing Sloman et al.'s (1998) model, described earlier. To implement the model, we derived the predictions of Equation 1 based on the ratings obtained in the causal centrality task and compared them with the participants' conceptual centrality ratings obtained from the experiment. A pair-wise dependency matrix for each participant and disorder was first determined from their responses in the causal centrality task. That is, the strengths participants assigned to the causal arrows constituted the cells of the matrix. For each disorder, the matrices were averaged over all participants to yield a single matrix. Model-predicted conceptual centrality ratings were set to the initial arbitrary value of 0.5, following the procedure of Sloman et al. (1998). The matrix multiplication was performed repetitively until the Spearman rank correlation of the model-predicted conceptual centrality ratings and the conceptual centrality ratings given directly by the participants converged to its terminal stable value. Rank correlations between the predicted and actual conceptual centrality ratings for each feature showed that the two factors were indeed positively correlated $r_s = .73$.

Thus, this experiment presents a number of important findings concerning laypersons' conceptual representations of mental disorders. First, laypersons seem to have richly structured causal theories about mental disorders, as shown in Figure 4. Participants were specifically told that they could leave the sheet blank if there were no causal relations among symptoms, but they spontaneously draw rich causal connections among symptoms. Second, this experiment presents the first demonstration of how naive theories on mental disorders correlate with laypersons' diagnoses. In accord with our hypotheses, symptoms that undergraduates rated as being causally central were also conceptually central.

A. Possible Implications for *DSM-IV*

We have shown that in laypeople at least, diagnosis is influenced heavily by causal theories. Therefore, another question that necessarily arises from an investigation of causal theories and diagnosis is whether the same issue applies also to clinicians. Elstein (1988), in his review of research on the cognitive processes involved in clinical inference and diagnosis, points out that clinicians do not, in general, make diagnoses by matching all the

symptoms of the client to the lists of criteria in the *DSM*. Instead, clinicians are much more likely to devote most of their attention to several symptoms thought to be ''prototypical'' of a specific disorder (Elstein, 1988). Cantor, Smith, French, and Mezzich (1980) have also argued that clinicians' reasoning adheres most closely to this prototype view of categorization. Our speculation is that the extra weight given to these ''prototypical'' symptoms is due in part to their place in the causal structure of clinicians' theories. Such a link between causal theories and diagnosis in clinicians seems likely. Indeed, Einhorn (1988) suggested that ''cues-to-causality,'' or cues that indicate probable causal relations, may be used by clinicians as constraints to hypothesis construction (p. 53). That is, the clinician may use these cues to narrow down possible causal scenarios for the yet-undiagnosed disorder, which at the same time narrows down the set of possible diagnoses.

We currently have a project under way to investigate the relation between causal theories and diagnosis for clinicians. If it turns out that clinicians also show a strong causal status effect, it may shed some light on a well-known problem in *DSM-IV* (APA, 1994) diagnosis—in particular, the reliability problem in diagnosing personality disorders. The *DSM-IV* (APA, 1994) specifies the more important symptoms for depression and other clinical disorders. For instance, it specifies that more weight should be given to ''depressed mood'' and ''lack of pleasure'' than to the other seven symptoms of a major depressive episode. For these disorders, there is high reliability of diagnosis between clinicians. However, the *DSM-IV* (APA, 1994) does not make such specifications for the personality disorders, and it is for these disorders that reliability of diagnosis is notoriously low. We suspect, given the findings in laypersons reported here, that clinicians may simply be administering their own weights to symptoms in diagnosis according to their personal theories of the personality disorders. We do not necessarily expect that this will be the whole answer to the problem of reliability in diagnosing personality disorders, but we do believe that it is a potentially important factor deserving further study. Future studies of the effects of clinicians' causal theories on their diagnoses may have serious ramifications for this aspect of the *DSM* system.

X. Other Types of Dependency Relations

As a final issue, this section discusses the research strategy we took, namely, focusing only on the causal relations among features in concepts. Certainly, concepts consist of relations other than causal relations. We have focused on the causal relations because they are considered to be the most critical part of theory representation (Carey, 1985; Gelman & Kalish, 1993; Well-

man, 1990). How much explanatory power do we lose because of this? We will discuss this issue by analyzing both bidirectional and directional relations among features.

A. Bidirectional Relations

Previous research has found that correlations are often noticed and used in categorization tasks (Medin, Altom, Edelson, & Freko, 1982; but see Murphy & Wisniewski, 1989). However, Malt and Smith (1984) presented the finding that for many real-world categories, properties that are correlated in participants' feature listings are not always noticed as such. Furthermore, they reported that these unnoticed correlations do not affect categorization. In part 1 of their Experiment 1, they compiled from participants a list of properties that are generally true of various exemplars of bird, furniture, tree, fruit, and flower. These were then used as the input to the correlation analyses. Across all the categories, an average of 33.5% of the property pairs were statistically reliably correlated. The property correlations obtained in this experiment were statistical cooccurrences, and were not necessarily reflected in people's actual representations of those concepts. For instance, ''eats insects'' and ''sings'' were found to be correlated in the bird concept. Indeed, Malt and Smith found that breaking such correlations did not affect typicality judgments. Thus, not all correlations mattered in categorization.

In Malt and Smith's (1984) Experiment 2, participants determined whether each of the correlated property pairs from Experiment 1 was perceived to have a relationship. Of the original pairs, 33% were explicitly rated as correlated. When they manipulated these explicit correlations, Malt and Smith found that the correlations now determined participants' typicality ratings. Specifically, the exemplars with the correlations intact were judged to be more typical than the exemplars with the correlations broken.

The question that Kevin Lee and the first author of this chapter raised was why only a subset of the correlations originally found in Malt and Smith's (1984) Experiment 1 were perceived to be correlated and affected typicality judgments. We hypothesized that the property pairs that participants explicitly rated as belonging together were noticed because of a perceived dependency relation between them. For example, people may have rated ''eats fish'' and ''is near ocean'' to go together for birds because eating fish depends on being near the ocean. On the other hand, we posited that ''eats insects'' and ''sings'' were not perceived as being correlated (even though they were, in fact, positively correlated) because people generally do not have an explanation for how those two features are directly related.

In our experiment, we presented each of the property pairs found to be correlated and in Malt and Smith's (1984) first experiment to 18 undergraduate students at Yale University. Pairs that were rated as correlated in part 1 of their Experiment 2 were labeled as explicit correlations, while the remainder were labeled implicit. For each property pair, two statements were generated, one to reflect each direction of dependency (A → B and B → A henceforth). For example, using the properties "eats fish" and "lives near the ocean," one test item was, "For a bird, whether or not it eats fish depends on whether or not it lives near the ocean," and the other test item was, "For a bird, whether or not it lives near the ocean depends on whether or not it eats fish." Along with each item, a scale marked from 1 for "strongly disagree" to 9 for "strongly agree" was displayed.

The idea is that the more asymmetric the dependency relation is between the members of a correlated pair, the greater the absolute value of the difference between ratings on A → B and B → A would be. Indeed, the absolute value of the mean difference score between A → B and B → A was greater in the explicit pairs ($M = 2.0$) than in the implicit pairs ($M = 1.4$), $p < .01$. We also examined the maximum score from each property pair. The mean maximum score for the explicit pairs (5.20) was significantly greater than that for the implicit pairs (2.96), $p < .05$. These results indicate that explicit pairs had more asymmetric dependency relations associated with them than implicit pairs. These results supported our hypothesis that perceived theoretical relationships, such as dependency, might allow correlations between some property pairs to be noticed, subsequently affecting categorization.

Let us return to the issue of the explanatory power of the causal status hypothesis. In this section, we attempted to show that correlations that are actually noticed and influence categorization are, in fact, likely to be directional relations. This proposal is consistent with Murphy and Medin's (1985) suggestion that

> people are not only sensitive to feature correlations, but they can deduce reasons for those correlations, based on their knowledge of the way the world works. Perhaps, then, the connection between those features is not a simple link, but a whole causal explanation for how the two are related. (p. 300; see also Murphy & Wisniewski, 1989)

Thus, by focusing only on directional relations, the causal status hypothesis does not seem to be too limited in accounting for feature centrality. We will now examine cases with directional relations.

B. Directional Relations

We are not aware of any empirical studies that have systematically examined all possible types of directional relations among features in people's concep-

tual representations. Thus, we will discuss two types of directional relations (other than a causal relation) that were examined in Sloman et al. (1998). In this study, participants were explicitly told that symptom A, for example, does not cause symptom B but that symptom B follows symptom A (temporal dependency), or that the presence of symptom B depends on the presence of symptom A (contingency; e.g., the presence of a moustache is contingent upon the presence of a mouth). The results showed that temporally preceding features or features on which other features are dependent were judged to be more conceptually central. What are the implications of these results for the causal status effect? There are at least three possibilities.

First, these results might have occurred because participants imposed complex causal interpretations on the temporal dependency and contingency relations (e.g., symptom A might not directly cause symptom B, as was specified, but it might indirectly cause symptom B). In this case, the effect of dependency structure can be viewed as a special case of the causal status effect.

Second, it might be that these ''noncausal'' relations are in fact, a type of causal relations in a loose sense. For instance, in a restaurant script (e.g., Schank & Abelson, 1977), there is a sequence of scripted events, such as ''customer goes into restaurant,'' ''customer goes to the table and sits down,'' ''customer picks up menu,'' and so on. These events are related in a temporal sequence, and it is difficult to say that one event caused another event in a strict sense. For instance, a customer's going into the restaurant did not make that customer sit down at the table. However, in a loose sense, these temporally related actions are causally linked in that changes in one may lead to changes in another. That is, a customer's going into the restaurant *allowed* the customer to sit down at the table. Sloman et al. also examined a ''contingency'' relation, such that the presence of a moustache is contingent upon the presence of a mouth. Again, it is difficult to say that a mouth caused a moustache to be present. But in a loose sense, there is a causal relation in that a mouth allowed a moustache to be present. If so, the causal status effect equals the effect of dependency relations, and Sloman et al.'s findings thereby suggest that there are no critical differences between at least some different types of causal relations.

Third, it could be that the causal status effect is a special case of a more general phenomenon occurring in any kind of asymmetrical dependency structure. This last possibility does not threaten the present claim that cause features are more central than their effect features, because the key ideas converge. Furthermore, even if the causal status effect is a special case of a more general phenomenon, it nonetheless appears to be a major portion of that general phenomenon, as indicated by studies showing that causal

relations alone can account for a large amount of variance in feature centrality for natural categories (e.g., Ahn, 1998; Kim & Ahn, 1999). Indeed, causal relations are prevalent and serve as essential components of relations that features have in our conceptual representations (e.g., Carey, 1985; Wellman, 1990).

XI. Conclusion

In this chapter, we have provided one mechanism by which feature centrality is determined. Our causal status hypothesis, derived from the idea of psychological essentialism, proposes that people regard cause features as more important and essential than effect features in that cause features affect category membership decisions more than effect features do. We reported several empirical studies that robustly support these main predictions of the causal status hypothesis, and reviewed a number of related phenomena that it can account for. We presented one computational model that implements the main gist of the causal status hypothesis, and reviewed direct empirical evidence for the model's validity. Boundary conditions of the causal status effect were stated and evidence for them reviewed. Empirical evidence modifying the predictions of the causal status hypothesis for feature weighting of causal versus isolated features was also discussed. A study was presented that successfully applied the causal status hypothesis to a class of real-world concepts, that of mental disorders. Finally, we provided a rationale for our decision to study only causal relations with respect to feature centrality, among the many types of possible dependency relations.

Acknowledgments

We would like to thank Charles Kalish, Douglas Medin, and Thomas Palmeri for their extremely helpful comments on the earlier draft of this article. This project was supported by a National Institute of Mental Health grant (R01 MH57737) awarded to Woo-kyoung Ahn, and a National Science Foundation Graduate Research Fellowship awarded to Nancy S. Kim.

References

Ahn, W. (1998). Why are different features central for natural kinds and artifacts?: The role of causal status in determining feature centrality. *Cognition, 69,* 135–178.

Ahn, W., Gelman, S. A., Hohenstein, J. & Kalish, C. W. (2000). Causal status effect in children's categorization. *Cognition, 75,* 1–9.

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54,* 299–352.

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (in press). Causal status as a determinant of feature centrality. *Cognitive Psychology*.

Ahn, W., & Sloman, S. (1997). Distinguishing name centrality from conceptual centrality. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 1–7). Mahwah, NJ: Lawrence Erlbaum Associates.

American Lung Association (1998). *Pneumonia* [On-line]. Available: http://www.lungusa.org/diseases/lungpneumoni.html

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed. revised). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Barlow, D. H., & Durand, V. M. (1999). *Abnormal psychology* (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Barton, M. E., & Komatsu, L. K. (1989). Defining features of natural kinds and artifacts. *Journal of Psycholinguistic Research, 18,* 433–447.

Billman, D. (1989). Systems of correlations in rule and category learning: Use of structured input in learning syntactic categories. *Language & Cognitive Processes, 4,* 127–155.

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22,* 458–475.

Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition, 60,* 1–29.

Bloom, P., & Markson, L. (1998). Intention and analogy in children's naming of pictorial representations. *Psychological Science, 9,* 200–204.

Braisby, N., Franks, B., & Hampton, J. (1996). Essentialism, word use, and concepts. *Cognition, 59,* 247–274.

Cantor, N., Smith, E. E., French, R., & Mezzich, J. (1980). Psychiatric diagnosis as prototype categorization. *Journal of Abnormal Psychology, 89,* 181–193.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: Plenum.

Chi, M. R., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152.

Corter, J. E., & Gluck, M. A. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin. 111,* 291–303.

Davidoff, J., Davies, I., Roberson, D. (1999). Calour categories in a stone-age tribe. *Nature., 398,* 203–204.

Diesendruck, G., & Gelman, S. A. (1999). Domain differences in absolute judgments of category membership: Evidence for an essentialist account of categorization. *Psychonomic Bulletin & Review, 6,* 338–346.

Einhorn, H. (1988). Diagnosis and causality in clinical and statistical prediction. In D. C. Turk & P. Salovey (Eds.), *Reasoning, inference, and judgment in clinical psychology* (pp. 51–70). New York: Free Press.

Elstein, A. S. (1988). Cognitive processes in clinical inference and decision making. In D. C. Turk & P. Salovey (Eds.), *Reasoning, inference, and judgment in clinical psychology* (pp. 17–50). New York: Free Press.

Follette, W. C., & Houts, A. C. (1996). Models of scientific progress and the role of theory in taxonomy development: A case study of the *DSM*. *Journal of Consulting and Clinical Psychology, 64,* 1120–1132.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology, 20,* 65–95.

Gelman, S. A., & Ebeling, K. S. (1998). Shape and representational status in children's early naming. *Cognition, 66,* B35–B47.

Gelman, S. A., & Kalish, C. W. (1993). Categories and causality. In R. Pasnak & M. L. Howe (Eds.). *Emerging themes in cognitive development* (Vol. 2, pp. 3–32). New York: Springer Verlag.

Gelman, S. A., & Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development, 58,* 1532–1541.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the nonobvious. *Cognition, 38,* 213–244.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge, UK: Cambridge University Press.

Gopnik, A., & Sobel, D. (1999, April). *Detecting blickets: How young children use novel causal information in categorization and induction*. Paper presented at the Society for Research in Child Development, Albuquerque, New Mexico.

Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development, 62,* 499–516.

Kalish, C. W. (1995). Essentialism and graded membership in animal and artifact categories. *Memory & Cognition, 23,* 335–353.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: The MIT Press.

Keil, F. C. (1992). The origins of an autonomous biology. In M. R. Gunnar & M. Maratsos (Eds.), *The Minnesota symposia on child psychology: Vol. 25. Modularity and constraints in language and cognition:* (pp. 103–137). Hillsdale, NJ: Erlbaum.

Keil, F. C., Smith, W. C., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition, 65,* 103–135.

Kim, N. S. (1999). *The role of naïve causal theories on lay diagnoses of mental illness*. Unpublished master's thesis, Yale University, New Haven, CT.

Kim, N. S., & Ahn, W. (1999, November). *The influence of naïve causal theories on lay diagnoses of mental illnesses*. Poster session presented at the 40th annual meeting of the Psychonomic Society, Los Angeles, CA.

King, S. D. (1993). *Babe: The gallant pig*. New York: Crown Publishers Inc.

Kraepelin, E. (1913). *Psychiatry: A textbook*. Leipzig: Barth.

Kripke, S. (1972). Naming and necessity. In D. Davidson & Harman (Eds.), *Semantics of natural language*. Dordrecht: D. Reidel.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3,* 299–321.

Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 754–770.

Lin, E. L., & Murphy, G. L. (1997). Effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 1153–1169.

Locke, J. (1975). *An essay concerning human understanding*. London: Oxford University Press. [Original work published 1894]

Malt, B. C. (1994). Water is not H_2O. *Cognitive Psychology, 27,* 41–70.

Malt, B. C., & Johnson, E. C. (1992). Do artifact concepts have cores? *Journal of Memory and Language, 31,* 195–217.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning & Verbal Behavior, 23,* 250–269.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 12,* 1469–1481.
Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 8,* 37–50.
Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge, UK: Cambridge University Press.
Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology, 20,* 158–190.
Murphy, G. L. (1993). A rational theory of concepts. In G. V. Nakamura, R. Taraban, & D. L. Medin (Eds.), *The psychology of learning and motivation* (Vol. 30, pp. 327–359), San Diego: Academic Press.
Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.
Murphy, G. L., & Wisniewski, E. J. (1989). Feature correlations in conceptual representations. In G. Tiberghien (Ed.), *Advances in cognitive science: Vol. 2. Theory and applications* (pp. 23–45). Chichester, UK: Ellis Horwood.
Palmeri, T. J., & Blalock, C. (2000). *The role of background knowledge in speeded perceptual categorization*. Manuscript submitted for publication.
Putnam, H. (1977). Is semantics possible? In S. P. Schwartz (Ed.), *Naming, necessity, and natural kinds,* pp. 102–118. Ithaca, NY: Cornell University Press.
Rehder, B., & Hastie, R. (1997). The roles of causes and effects in categorization. *In Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 650–655). Mahwah, NJ: Lawrence Erlbaum Associates.
Rips, L. J. (1989). Similarity, typicality, and categorization. In A. O. Stella Vosniadou (Ed.), *Similarity and analogical reasoning* (pp. 21–59). New York: Cambridge University Press.
Roberson, D., Davidoff, J., & Davies, I. (1999). *Are color categories universal? New evidence from a traditional culture*. Paper presented at the 40th annual meeting of the Psychonomic Society, Los Angeles, CA.
Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–47), Hillsdale, NJ: Erlbaum.
Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605.
Ross, B. H. (1996). Category representations and the effects of interacting with instances. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22,* 1249–1265.
Ross, B. H. (1997). The use of categories affects classification. *Journal of Memory & Language. 37,* 240–267.
Ross, B. H. (1999). Postclassification category use: The effects of learning to use categories after learning to classify. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 25,* 743–757.
Schank, R. C., & Abelson, R. (1977). *Scripts, plans, goals, and understanding,* Hillsdale, NJ: Erlbaum.
Schwartz, S. P. (1979). Natural kind term. *Cognition, 7,* 301–315.
Sloman, S. A., & Ahn, W. (1999). Feature centrality: Naming versus Imaging. *Memory & Cognition, 27,* 526–537.
Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science, 22,* 189–228.
Soja, N. N., Carey, S., & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meaning: Object terms and substance terms. *Cognition, 38,* 179–211.
Stervens. M. (2000). The essentialist aspect of naïve theories. *Cognition, 74,* 149–175.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,*(4), 327–352.
Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.
Wisniewski, E. J. (1995). Prior knowledge and functionally relevant features in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 449–468.

This Page Intentionally Left Blank

REMEMBERING AS A SOCIAL PROCESS

Mary Susan Weldon

I. Introduction

In 1932, Sir Frederic C. Bartlett published his famous book *Remembering: A Study in Experimental and Social Psychology*. The title makes three interesting points about the study of memory. First, the use of the term *remembering* rather than *memory* emphasizes his view that remembering is a process rather than an entity or mental faculty. Second, the title suggests that the study of remembering requires special attention to the experimental methods one uses in psychology. And third, the title implies that remembering is a topic in social psychology. The first two points have been discussed and debated extensively in psychology, both before and after Bartlett's time. But it is the third point that is of interest here. Bartlett's argument that remembering is social has received little serious treatment in psychology, and has had no perceptible influence on how memory has been conceptualized or investigated in mainstream experimental work. Behaviorists sought to understand learning and memory in terms of the stimulus–response associations acquired by the individual organism, and modern cognitive psychologists in terms of the mental processes and cognitive structures that comprise the individual mind.

My goal in this chapter is to take a serious look at the proposition that remembering is a social process, and to do so from the perspective of a cognitive psychologist. The main questions I want to explore are, first, what does it mean to characterize remembering as a social process, and second, what does this imply for how one might study memory?

THE PSYCHOLOGY OF LEARNING AND MOTIVATION, VOL. 40

Copyright © 2001 by Academic Press.
All rights of reproduction in any form reserved.

0079-7421/00 $35.00

In the first section, I briefly examine some assumptions of the information-processing metaphor to provide a background for considering and contrasting alternative perspectives. I then present a sample of ideas about the social aspects of memory and cognition, discussing relevant arguments by Durkheim, Halbwachs, Bartlett, Vygotsky, and Hutchins. This is not intended to provide a comprehensive overview, but rather to highlight some of the important and distinct perspectives that have been taken on the topic, and hopefully persuade the reader that there is something to be gained by thinking about memory as a social process. This is only half the struggle, however, because even if one is persuaded by this general idea, one next has to figure out how to incorporate this perspective into one's theories and methods in cognitive psychology. It turns out that there is no general consensus about an appropriate perspective or set of methods, and the second section provides an overview of different theoretical, conceptual, and methodological approaches that investigators have taken in efforts to investigate the social nature of learning and memory, some of which appear fundamentally incompatible with traditional cognitive psychology. It is important to note that the theoretical concepts discussed in the first two sections do not always focus on memory per se, but sometimes on the social nature of cognition in general; however, the arguments easily apply to memory. Finally, in the last section, I present findings from two lines of research that I have undertaken in my endeavor to study memory as a social process, one looking at the effects of collaboration in remembering, and the other an effort to adopt graph theoretic notions from social network analysis to characterize collective memory.

II. Arguments for the Social Nature of Memory

For many decades much of the field of cognitive psychology has employed an information-processing view of cognition, using the computer as a metaphor for mental architecture and processes. In the area of memory, for example, this metaphor underlies a variety of theoretical constructs such as memory buffers, short-term memory, long-term memory, memory search, input, encoding, retention, retrieval, parallel versus serial processes, rehearsal, codes, representations, and so forth. The metaphor has provided many useful concepts, but the limitations are widely acknowledged either to exclude or fail to account for important phenomena such as reconstructive and creative processes, emotional influences, and social and cultural processes in cognition. As Gardner (1985) noted, emotional, social, and cultural processes have been set aside while investigators have sought to understand individual cognition, the obvious assumption being that the

processes of individual cognition could be separated from emotional and sociocultural considerations without compromising the integrity of the system. This approach has had an important impact on how emotional and sociocultural processes later have been conceptualized. Specifically, rather than being treated as integral to the processes of cognition, they have been treated as contextual factors whose influences could be controlled as independent variables or accounted for as subject variables, that is, as inputs to the system rather than as components of the system, or as part of the contents of mind rather than part of the processes of mind.

Although modern cognitive psychology is dominated by this modular and individualistic approach to cognition, the idea that human thought is a social process, or at least that it cannot be understood independently of the social context, has received serious treatment in other disciplines. Here I would like to provide a flavor of this idea by presenting five interesting lines of argument for the social nature of memory, those of Émile Durkheim, Maurice Halbwachs, Frederic C. Bartlett, L. S. Vygotsky, and Edwin Hutchins. There are many other thought-provoking arguments and data that the interested reader also should consult (e.g., Baltes & Staudinger, 1996; Cole, Engeström, & Vasquez, 1997; Lave & Wenger, 1991; Resnick, Levine, & Teasley, 1991; Tobach, Falmagne, Parlee, Martin, & Kapelman, 1997).

A. Émile Durkheim

Émile Durkheim (1858–1917; Durkheim, 1903/1985, 1915/1985) is widely considered the father of modern French sociology, and is credited for introducing certain empirical methods into the study of society. Interestingly, he spent some time in Wilhelm Wundt's laboratory where the fledgling science of experimental psychology was taking shape. Most of Durkheim's work addressed issues of morality and social justice, and as such he was interested in the origins of religion and its role in society. In one of his most important works, *The Elementary Forms of the Religious Life* (Durkheim, 1915/1985), he argued that religion was one of the first systems of collective representation of beliefs and knowledge, in other words, that religion serves as a collective memory. He suggested that these early belief systems arose from the human tendency to classify nature, and furthermore, that the classification schemes that humans used were modeled on their observations of how their own societies were organized. In other words, Durkheim argued that the first logical categories were social categories. Because humans are grouped, and think of themselves as organized into groups, they group other things in nature as well, and so logical hierarchy is just another aspect of social hierarchy. Thus, it is the natural social order that gives rise to more abstract mental concepts about the categorical

structure of nature in general. Furthermore, Durkheim argued that it is not only the idea of categories but also the nature of the logical relations among categories that is social in its origin. For example, original social groups were organized into families, families into clans, clans into tribes, and so on, and this gave rise to the commonplace nested structure that characterizes categorical organization of other things in the world.

Durkheim also argued that there is a fundamental relation among the universal categories employed in thought, religion, and collective activity. In his view, ideas such as time, space, class, number, cause, substance, and personality are the most fundamental and universal of concepts, and are also universal properties of religious cosmologies. But this is no coincidence, because these concepts arise from collective experience and represent collective realities. His theory of totemism explained how religious rituals, symbols, and artifacts served as a reflection of the group consciousness that arose through collective activity, and further served to sustain and re-create shared mental states. Religions formed the basis of nearly all great social institutions, including science. ''That which makes a man is the totality of the intellectual property which constitutes civilization, and civilization is the work of society'' (p. 127; Durkheim, 1915/1985). The mental life of each individual depends on collective life and collective action. Durkheim did not deny the importance of the individual, but rather emphasized that individuals, society, and nature could be abstracted from one another only artificially.

B. Maurice Halbwachs

Maurice Halbwachs (1877–1945), a student of Durkheim's, explored the social nature of memory in his treatise *La mémoire collective* (*The Collective Memory;* Halbwachs, 1950/1980). In this work, he discussed the interrelations among individual, collective, and historical memory, as well as the social structuring of time and space in relation to memory. Halbwachs argued that ''a person remembers only by situating himself within the viewpoint of one or several groups and one or several currents of collective thought'' (p. 33). At any point in time a person's interpretation and remembering of experience originates in particular social milieus, and so all memory is socially based, reflecting viewpoints constituted through one's membership in various social groups. Figure 1 illustrates my understanding of Halbwachs's perspective. The circles represent the various social groups to which a person belongs and that are in play at any point in time. The influence of the group on one's viewpoint during experiencing and remembering depends both on one's physical presence and one's mental engagement with that group at the particular point in time, and these two influences

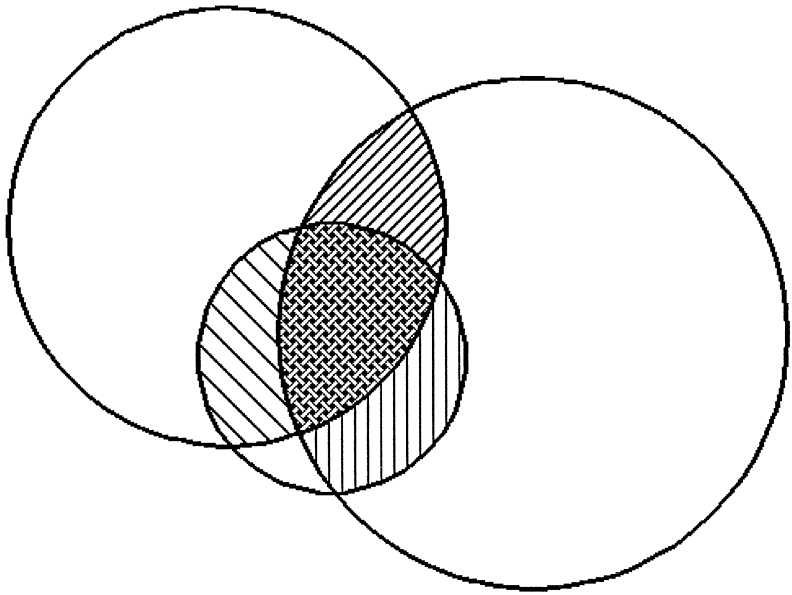


Fig. 1. A depiction of Maurice Halbwach's (1950/1980) explanation of the illusion that some experiences and memories are independent of social groups. The influences of multiple groups can be in play at any point in time, each contrasting and reinforcing the other, giving rise to a unique viewpoint that is misattributed to one's independent self.

can vary relatively independently. The impression that one has individual memories (not in the sense of being private, but in the sense of originating solely within one's self) can be thought of as a memory illusion. When thoughts arising from more than one group come together simultaneously, they contrast and reinforce each other, each changing the appearance of the other by virtue of their juxtaposition. This gives rise to a unique or unfamiliar experience that the individual alone enjoys, and thus creates the illusion that the thought is independent. The individual loses track of the fact that he or she is considering the experience from the viewpoint of multiple groups simultaneously, each from the viewpoint of the other. The individual experience or remembrance is the intersection of these collective influences, even though the person is inclined to experience it as a unique event arising only from the self.

One's involvement in different social groups waxes and wanes over time, and thus so does the strength of their influences. To the extent that one loses ties to social groups, remembrances also are lost, and as such, knowledge cannot be separated from the contact with the social group. For

example, events that are experienced from the viewpoint of two groups are to some extent dependent on sustained contact with those groups, and remembering will be best in a situation permitting both of these influences to combine their action. As one loses contact with one of the groups, the remembrance will seem less familiar and the collective factors determining it will become hidden. Halbwachs (1950/1980) saw this process almost as a type of source amnesia or unconscious plagiarsm, to use modern terms: ''How many people are critical enough to discern what they owe to others in their thinking and so acknowledge to themselves how small their own contribution usually is?'' (p. 45).

C. Frederic C. Bartlett

Sir Frederic C. Bartlett (1886–1969) was a contemporary of Halbwachs and was familiar with his work, although it does not appear that they ever met. Whereas Halbwachs arrived at his interest in collective memory through philosophy and sociology, Bartlett came to his ideas about the social nature of memory through a very different path, and with very different emphases. Bartlett started his research in psychology in 1913 as a student in the laboratory of C. S. Meyers at the University of Cambridge, where Ebbinghaus's methods had been introduced. Although Bartlett appreciated the experimental control Ebbinghaus had achieved, he became disillusioned with the use of nonsense syllables and the method of rote repetition that was employed to learn them. He argued that nonsense syllables did not really simplify the understanding of the relation between stimulus and response because they did not really rid the stimuli of meaning, and that ''uniformity and simplicity of structure of stimuli are no guarantee whatever of uniformity and simplicity of structure in organic response, particularly at the human level'' (Bartlett, 1932, p. 3). In other words, even if two people are presented with the same nominal stimulus, they will notice different features, impose different meanings, and respond with different behaviors, and Ebbinghaus's methods could not address this. Bartlett also argued that Ebbinghaus's methods resulted in the study of isolated habits, not remembering as a complex process integrated with normal functioning. Because remembering is not an independent mental function, but rather is integrated with perceiving, imagining, and thinking, any effort to isolate memory processes results in an artificial task in which the investigator is no longer studying the very thing that initially aroused his or her interest. Thus, Bartlett rejected the reductionist approaches that characterized the research conducted by Ebbinghaus and psychophysicists, and instead tried to study remembering as a constructive process arising through the dynamic interplay of past experience, present perception, thought, affect, and the

temperament of the individual. A natural upshot of this approach was also to recognize that social groups play an important role in the process of remembering.

Bartlett (1932) was not concerned with establishing the existence of a ''group mind'' or ''collective memory'' per se, a quest he considered unlikely to be successful. Rather, his emphasis was on the fact that individual behavior cannot be separated from its group constituency; social influences govern behavior every bit as much as physical stimuli such as light or bodily sensations, or information such as the news in a telegram. Therefore, to understand behavior one has to consider the social contexts in which individual behavior is situated. He defined a group as an organized collection of people held together by an active influence, such as a belief system, a cause, or other psychological common ground. This special identity of the group, or its ''bias,'' gives rise to its characteristic practices, customs, attitudes, ideals, perspectives, institutions, and even its group temperament, all of which give persistence to the group culture across time and space. Bartlett argued that the group identity was more than simply a background or context for interpreting individual action, a perspective that would be relatively trivial. He wanted to make the stronger claim that the group culture serves as a schema that biases the members' attitudes toward and perceptions of events, and also serves as a framework for reconstructing memories of events.

Bartlett asserted that ''both the manner and the matter of recall are often predominantly determined by social influences'' (Bartlett, 1932, p. 244). For example, social customs and interests prescribe the practices people employ to remember and recount their experiences. To take an example reported recently, Mayan children may appear to have poor recall when retelling an event to an adult, but the interpretation of this behavior changes when one realizes that it is culturally inappropriate for Mayan children to speak freely to an adult, so they must employ linguistic conventions for showing deference and respect (Rogoff & Mistry, 1985). Likewise, the contents of memories reflect that which is valued by the social group and important to one's successful functioning in it. For example, Bartlett (1932) reported that Swazi herdsmen had excellent memory for cattle transactions conducted over a year in the past, including the seller, type, age, and cost of each head bought. One might imagine that a real estate broker has similar skill remembering clients and property prices.

Bartlett's ideas about the social nature of remembering are not particularly radical as compared to the perspectives of Durkheim and Halbwachs. Bartlett is more comfortable distinguishing between the individual and the social influences that modify and direct the individual's behavior. Nevertheless, Bartlett's ideas about the social underpinnings of memory have had

no recognizable influence on the study of learning and memory that has taken place in the mainstream traditions of behaviorism, verbal learning theory, and cognitive psychology, which followed the publication of his work. In contrast, his ideas about schema and reconstructive processes in individual memory have been enormously influential, albeit lacking any reference to his hypotheses about the social roots of these processes.

D. L. S. Vygotsky

Many contemporary sociocultural approaches to cognition have roots in the ideas of L. S. Vygotsky. Vygotsky was a Soviet psychologist writing in the early part of the twentieth century, and who was answering the call to develop a Marxist psychology. He was concerned with the process of development, including the development of mental functioning, and his ideas have been highly influential in modern concepts of cognition as a socially situated activity, particularly within developmental psychology. Vygotsky claimed that to understand the individual one must understand the social networks in which he or she develops. He stated that internal mental processes and structures have their origins in the external world, specifically in social interactions. Peoples' interactions are also mediated, meaning that they depend on ''tools,'' which are both technical (e.g., the table of elements) and psychological (e.g., language), and which are internalized over time. Thus, higher mental functions are social in several senses. First, the concept of mental function applies to both individual and social activity. Second, individual psychological functions (or *intrapsychological* processes) are transformations of psychological interactions among people engaged in activity in the external social world (or *interpsychological* processes). Because both social interaction and its mediational tools are social in origin, their internalized forms are also social. Third, the individual's own mental functions are transformed as the interpsychological processes and mediational tools are internalized; for example, memory processes are transformed by the internalization of language. Fourth, interpersonal processes and mediational tools are products of sociocultural evolution and so are situated in the larger sociocultural context, in both their external forms and internal forms. This summary represents a great oversimplification, and more extensive treatments are available elsewhere (Minick, 1997; Rogoff, 1998; Wertsch, 1991; Vygotsky, 1962, 1978).

Vygotsky's social learning theory has been adopted, developed, and modified in many sociocultual approaches to cognition that have taken inspiration from him, including *activity theory* and theories of *situated learning*. In general, the important assumption of these cultural-historical perspectives is that there is interdependency among the individual, activity, society, culture,

and their respective histories, and they are mutually constituted in ongoing processes. Therefore, the individual cannot be studied in isolation because to do so would separate these component processes from their normal integrated functioning. This would be akin to trying to understand how the respiratory system works by removing the lungs from the body, or trying to understand the nature of the ocean by examining a glass of seawater.

E. Edwin Hutchins

To conclude this overview, I would like to outline some ideas of Edwin Hutchins (Hutchins, 1994), a contemporary cognitive scientist whose line of argument about this topic is very different in nature from those discussed so far, in part because his point of departure is modern theory in cognitive science. Hutchins criticizes the computational metaphor in which the architecture of the individual mind is modeled on a standard computer and in which perception, attention, memory, language, decision making, and so forth are identified as separate areas of study and treated as independent entities. Hutchins rejects this modular perspective and argues that not only do mental processes need to be understood in a more integrated fashion, but also that cognition needs to be understood as a cultural process.

When cognitive science puts formal symbol systems into the head, the person is eliminated from the model of the cognitive system because, by doing so, sensation, action, emotion, context, and culture are separated from human cognition, and so the model no longer characterizes how the individual functions. In his view, the problem with the computer model of mind is not the hypothesis that minds process symbols per se, but rather that human cognition is necessarily symbolic (also see Kolers & Roediger, 1984, for other arguments and data challenging propositional representations). When the whole of cognition is characterized simply as a formal symbol-manipulating system inside the head, cognition is detached from the eyes, ears, and hands that support interaction with the material world, which gave rise to the ability to create and use formal symbol systems in the first place. Treating the individual mind as the appropriate unit of cognitive analysis leads to pretheoretical assumptions that are fundamentally flawed. For example, if one has the impression that individual minds operate in isolation, then one may mistake the properties of complex sociocultural and technological systems as reflecting the properties of advanced minds, and try to account for differences in technological advances between different groups as reflecting differences in their cognitive capacities.

According to Hutchins, culture is an adaptive process that takes place both within and outside the minds of people, and cognitive activity is part of this process, so the unit of cognitive analysis must extend beyond the

physical boundary of the individual mind. If one relegates all of cognition to internal mental processes, then one is required to pack all the explanatory machinery of cognition into the individual mind as well, leading to misidentification of the boundaries of the cognitive system, and the overattribution to the individual mind alone of the processes that give rise to intelligent behavior. For example, formal symbol systems such as mathematics and logic are products of sociocultural processes in which humans play a part and which have evolved over hundreds of years. But the activity of the individual's mind is not the same as the activity of the sociocultural system as a whole, nor does the mind necessarily function in the same manner as any particular product of the cultural system, such as a system of formal logic. The computational metaphor describes what the culture does (i.e., manipulates abstract symbols to solve problems or communicate), but not what the individual mind does.

In reality, according to Hutchins, the development of both mind and cultural practices takes place simultaneously through the conduct of ongoing daily activities, including everything from ancient routines like preparing meals, to cutting-edge technological activities like surfing the web, to activities that require highly complex technology and social coordination such as performing surgery or navigating a ship. Mental processes, cultural processes, and human activity are dynamic, ongoing processes, and in this sense cognition is fundamentally a cultural process. For Hutchins, the cognitive system is more meaningfully viewed as comprising the person acting in coordination with a social group to perform activities, using tools and practices that have evolved over time within a culture. Hutchins's view is distinctive in that his goal is not to account for individual mental life in terms of sociocultural processes per se. Rather, his goal is to address the role of culture by characterizing the information-processing system as larger than the individual, as comprising the sociocultural processes themselves.

To summarize, I have tried to give a flavor of five different perspectives on the social nature of cognition, and in particular, the social nature of memory. Although these perspectives are distinctive in their views of the relation between individual and social processes, they all share a fundamental assumption. It is not sufficient simply to acknowledge that remembering takes place in a sociocultural context, and that this context can provide some interesting insights into how people think and what they think about. Rather, it is impossible to understand memory without simultaneously accounting for the social processes that give rise to the processes of learning and remembering. Accordingly, any account of remembering based only on an account of the individual is not simply incomplete, but is flawed.

III. How Can Remembering Be Studied as a Social Process?

If one is persuaded by arguments that cognitive processes cannot be understood in isolation from social processes, or at least that there is something to be learned by considering the social nature of cognition, then how is one to proceed? The implications of integrating mind, society, and culture are profound in many respects, bearing on the pretheoretical assumptions one brings to investigations of cognition, the units of analysis one observes, the models and metaphors one employs to understand phenomena of interest, the data one collects, the form in which the data are presented, and the generalizations one can make about human cognitive behavior.

Investigators interested in the social nature of memory and cognition have proposed and implemented a variety of theoretical stances and methodological approaches, and there is no general consensus about the best conceptual frameworks or set of methods that should be adopted. The aim of this section is to provide an overview of these approaches and point out some of the important conceptual and methodological distinctions among them. I am going to characterize these approaches as ranging from conservative to radical, based on the degree to which they depart from traditional assumptions and methods in cognitive psychology, and thus the degree to which they require a major conceptual stretch for the traditionally trained cognitive psychologist (of which I count myself as one). However, the classification scheme should not be seen as representing mutually exclusive categories because there is considerable overlap between some frameworks, and some frameworks could fit into more than one category.

A. The Individual in the Social Context

1. Social Cognition

A relatively conservative approach to this endeavor is reflected in some paradigms that have been used in social psychology to study social cognition. In this approach, the basic information-processing structure of traditional cognitive models is adopted, and the stimulus inputs and outputs are simply changed to be ''social'' stimuli, such as photographs of faces or lists of personality traits, rather than traditional ''cognitive'' stimuli such as letters, words, pictures, or discourse. In addition, participants are typically tested individually and in isolation, so as not to be contaminated by social interaction.

As an example, early studies of person memory reported that when people read a list of personality traits about a target person (e.g., ''religious,'' ''respectful''), and then later read about various activities of the target person, people remembered more of the behaviors that were inconsistent

with their expectations of the target (e.g., got drunk and missed church services) than were consistent with expectations (e.g., said grace before dinner; Hastie & Kumar, 1979; Srull, 1981). This result can be explained with a standard information-processing model distinguishing between working memory and long-term memory. Specifically, because the research participant is trying to make sense of the inconsistent behaviors, he or she processes them more elaboratively and for a longer period of time in working memory, and as a result these behaviors are more likely to be recorded in long-term memory. In this paradigm the fundamental nature of the cognitive machinery intervening between input and output remains the same whether one is attending to word lists (''cognitive'' stimuli) or a person's traits (''social'' stimuli). (Incidentally, this result does not hold up with more socially realistic methods, see Fiske & Goodwin, 1994.)

This approach has been criticized by many social psychologists interested in social cognition because it fails to capture the social nature of the processes of interest. The perceiver is treated as a hermit, isolated from other people (Fiske & Taylor, 1991). Ickes and Gonzalez (1996) emphasize two critical shortcomings of this paradigm: (a) it assumes independence instead of interdependence of participants' responses; and (b) it treats social cognitive processes as subjective rather than intersubjective, that is, as arising from reciprocal interaction. This leads to the odd situation in which it is assumed that to study social cognition it must first be removed from the social interaction that engenders it. Any effort to understand social cognition in isolation from its intersubjective context, when social processes are by definition intersubjective, will fail to capture the essential nature of social cognition.

This paradigm also seems to be an inappropriate approach for studying remembering as a social process. Remembering is not social simply in the sense that one can remember social information, such as parties, faces, personalities, and so forth. Rather, the processes of remembering are themselves social processes, and the paradigm does not capture this. Even when one is sitting alone, remembering to oneself, mental activity is part of a sociocultural process that cannot be excised from the individual's mental life. For example, the process of talking to one's self is simply the internalization of a social activity. In a sense, one needs to invert the traditional approach to the relation between cognitive psychology and social cognition. Rather than starting with a model of individual cognition and trying to fit social processing to it, I am suggesting that we should start with the social processes and see how cognition arises as part of these. In other words, the interest here is not in the cognition *of* social information, but in cognition *as* social.

2. *Sociocultural Influences*

A second approach that is relatively conservative, but is more in line with the idea of studying cognition as embedded within a sociocultural context, is to think of the sociocultural context essentially as an independent variable, and incorporate it into the factorial structure of the traditional research design. An example of this approach is illustrated in recent work examining the hypothesis that cultural belief systems influence the cognitive styles of the individual members of a society (Nisbett, Peng, Choi, & Norenzayan, in press). Nisbett et al. hypothesized that members of Eastern Asian cultures that descend from ancient Chinese civilization should differ from members of Western cultures that descend from ancient Greek civilization in terms of how they perceive and interpret the world. Ancient Greeks had a strong sense of personal agency, the idea that individuals have a free will and the ability to take independent action. They also had an interest in discovering laws to explain and predict nature, which led to great advances in science and mathematics. In contrast to the Greeks, the Chinese possessed a stronger sense of social obligation than of individual rights, and although they had a technologically more advanced civilization than the Greeks, their innovations arose from a more practical, intuitive approach to life rather than formal models of nature. Based on these characterizations of the root civilizations, Nisbett et al. suggest that these different sociocultural practices and values have led to important differences in the basic cognitive processes of the individual members of the descendent civilizations, influencing reasoning, inference, categorization, and ways of attending to the environment.

An example of a test of this idea is illustrated in Figure 2. Nisbett et al. (in press) hypothesized that East Asians' tendency to believe that causality is located in the environment will lead them to see the environment as an interrelated whole, whereas Westerner's tendency to look for causality in individual objects will lead them to focus on individual objects in the environment. Among a variety of interesting tests of this idea, they showed Japanese and American students animated underwater scenes that included prominent focal fish in a context of other underwater objects such as smaller fish, rocks, seaweed, and so forth. When later asked to report what they had seen, Americans and Japanese were equally likely to mention the focal fish. However, Japanese were significantly more likely to mention "background" features such as seaweed, the pond, and so forth. Furthermore, Japanese were more likely to describe relations among the components of the pictures, such as how the fish swam by the seaweed. On the other hand, when given a recognition test for the focal fish placed in a new background, the Japanese students' recognition was harmed by the change of context, whereas the Americans' was not.

	Approach to Environment	
	Japanese students Relatively holistic, relational	**U.S. students** Relatively analytic
Memory for "focal" objects	Good	Good
Memory for "background" objects and relational features	Good	Poor
Recognition of focal objects in changed background	Poor	Good

Fig. 2. Factorial design to test the role of culture in modifying cognitive processes. (From Nisbett et al. in press, *Psychological Review.* Adapted with permission.)

The authors interpreted these findings as evidence that individual cognitive processes may differ as a function of different worldviews that are propagated through the different cultural histories of a society. Of course, one should be careful not to interpret these data in terms of essentialist models of cultural traits, but rather as reflecting average group tendencies influenced by different cultural milieus. But the main point I want to make here concerns the conceptual approach represented by this style of research. This approach to sociocultural influences in cognition is fairly traditional in that the individual serves as the unit of analysis, the information-processing metaphor is intact (or at least, not challenged), and culture serves as an independent variable that simply changes something about the individual's cognitive processes, in this case changing how the scene is parsed during perception and reconstructed during recall.

By labeling this as a traditional approach I do not mean to imply that the findings or their implications are trivial. Certainly they lead to important questions about the degree to which one can characterize at least some components of the cognitive architecture as comprising a fixed or hard-wired system. (For another example, see Lancy & Strathern, 1981, who demonstrate that hierarchical taxonomy is not a universal organizational strategy.) Rather, I mean to suggest that the approach (a) employs standard experimental methods and design; (b) uses the individual as the unit of analysis; (c) implies that cultural influences can be conceptualized and described within a traditional factorial structure, in which culture is independent from other factors, such as type of pictorial element or type of memory

test; and (d) models cognition as an information processing system within the individual. In this approach culture is simply added on as an additional factor in a factorial design and, as the logic of this approach goes, if culture is important in cognitive functioning then it should produce at least a main effect, but hopefully an interaction. (See Figure 2; simple main effects such as mere differences in the level of performance across different cultures are usually difficult to interpret in this kind of work). This type of design has been used to uncover a variety of interesting cross-cultural differences in cognition, and is the same one typically used to study sex differences, age differences, and other individual differences. The assumption is often that the so-called sociocultural factor can be added or subtracted from the study of cognition according to one's interest. However, the standard methods for studying cognition are not considered to be compromised when sociocultural factors are ignored, and the information-processing view of cognition typically is not seriously challenged by this approach. Rogoff (1998) refers to this as the ''social influences'' approach to the study of social and cultural factors in cognition.

B. The Group and Society as Cognitive Systems

A somewhat more adventurous approach to the idea that cognition and remembering are social processes is to conceive of the group at large as the cognitive system, and therefore as a necessary unit of analysis. Approaches of this type have taken different emphases, varying in the degree to which they focus on the group or the group-and-individual, as well as the degree to which they attempt to characterize the cognitive properties of the group versus those of the individual within the group. Note that in this approach the group members do not necessarily need to be physically copresent to serve as a cognitive system, but the group can extend across time and space.

1. The Group as an Information-Processing System

A recent review by Hinsz, Tindale, and Vollrath (1997) provides an overview of research that treats small groups as information-processing systems. Although the authors do not posit that the group-level architecture is necessarily identical to that of the individual mind, they suggest that a generic individual-level information-processing model, such as that presented in Figure 3, can be used as a starting point to explore the properties of the group-level system. Just like individuals, groups have goals, make efforts to direct their attention appropriately, engage in learning, retention, and retrieval of events and knowledge, process and integrate information, make decisions and judgments, and produce responses. However, groups

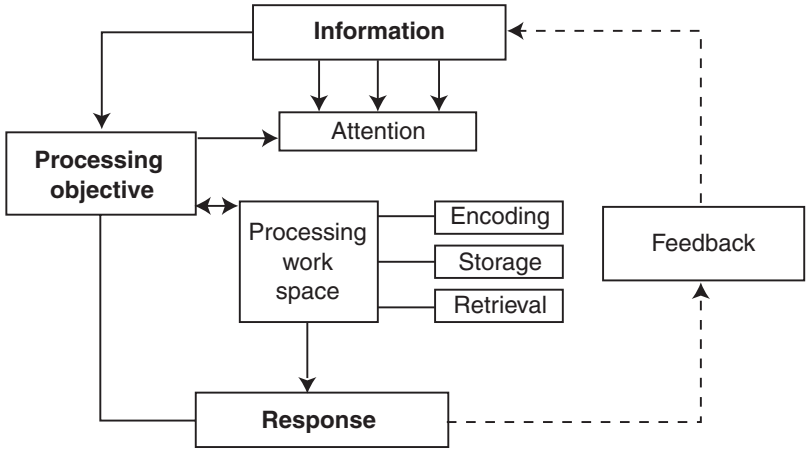


Fig. 3. Generic information-processing model representing the individual's cognitive system. (From "The emerging conceptualization of groups as information processors," by V. B. Hinsz, R. S. Tindale, and D. A. Vollrath, 1997, *Psychological Bulletin, 121,* p. 44. Copyright 1997 by the American Psychological Association. Reprinted with permission.)

have additional layers of complexity that are not normally considered in individual cognition. For example, in addition to managing the cognitive demands of the task, group members need to coordinate their individual goals, attention, knowledge, decision making, and responses, which requires complex social management.

Essentially, Hinsz et al. (1997) use the information-processing model as a framework for organizing the extensive research findings about small group performance. For example, individual group members may share some task-relevant knowledge while also uniquely knowing other information that the group might need, and so the processes whereby group members access and discuss shared and unshared knowledge can be modeled as the retrieval processes of the group-level information-processing system (e.g., Stasser & Titus, 1987). One advantage of this approach is that it provides a ready working model for describing the architecture of group-level cognitive activity, and suggests testable hypotheses. Furthermore, the terminology and concepts are familiar to the cognitive psychologist and thus provide a straightforward extension from individual to group cognition. At the same time, the approach is more radical than the social-influences approach described above in the sense that it treats the group as a unit of analysis. An important issue to be addressed in the future is whether the information-processing model provides the most suitable vocabulary and conceptual framework for modeling group-level cognition, and whether it

has the ability to capture the interdependent, reciprocal, and recursive aspects of group cognitive activity.

2. *The Individual, Group, and Society as an Information-Processing System*

Although Hinsz et al. (1997) do not deny the importance of broader sociocultural aspects of cognition, their theoretical focus is on the cognitive dynamics of the small group, and they do not describe how they conceptualize the sociocultural milieu in this framework, for example, whether they see it as a contextual influence or a constituent process. However, other investigators have made efforts to give more explicit attention to the sociocultural milieu as an integral component of the information-processing system.

As discussed earlier, Hutchins (1994) also employs the information-processing metaphor to describe cognitive activity at the group level, even though he roundly rejects it at the level of the individual. Underlying his information-processing metaphor are several important pretheoretical assumptions about the system.

Social customs and practices have evolved in such a way that cognitive labor is distributed across individuals, activities, and work settings. Even an individual working alone is doing the work within a complex social system. Because societies organize their cognitive tasks in a distributed fashion, he assumes that groups will have some cognitive properties that cannot be predicted or understood by understanding only the individuals. To illustrate such emergent properties, Hutchins stowed away on the bridge of a navy ship for an extended period of time to observe the complex team activity of ship navigation. He explored the computational systems that exist within functional cognitive units larger than the individual mind, such as the person in interaction with a navigational tool, the team as a whole, and the system of navigational practices that have developed over time within the culture.

In his analysis of the team-level cognitive system, for example, a variety of terms familiar to cognitive psychologists appear, such as *bottom-up* and *top-down processes, daemons, parallel processes, buffers,* and *distributed memory*. However, there is an important difference from their traditional usage. Here, it is the *social organization* rather than the individual mind that is hypothesized to give rise to this cognitive architecture. For example, within a navigational team knowledge is distributed in such a way that not all knowledge is held entirely by any one person, some knowledge is shared among many or all team members, and some knowledge is unique to specific individuals. The distribution of knowledge is partly determined by the way

societal practices have evolved, leading to particular organizations and distributions of tasks within work groups. In addition, social conventions prescribe certain rules of conduct and interaction, usually within hierarchical social structures. Individual team members understand how knowledge and tasks are distributed in the group as well as the social rules that govern activity in the hierarchy, and these conventions influence the behavior of each team member as he or she participates in the activities that lead toward the desired outcome. It is this structure that serves as the computational architecture of the cognitive system. The computations of the system are carried out through interactions among the participants with each other and with the various tools and representational systems that they bring into coordination to accomplish their task (also see Kirsh & Maglio, 1994). Each person has certain knowledge (e.g., how to take a bearing or how to plot the recorded bearings to get a position), certain technologies (e.g., a clock or a navigation chart), and a certain status, and the group has procedures for sequencing the work. Successful navigation requires that all these elements be coordinated, but the process of bringing the right actions, tools, and knowledge into play at the right times is not simply a social negotiation, but rather it actually *comprises* the computations of the cognitive system. Thus, it makes little sense to try to characterize the cognitive activity of the individual mind independent of the social system because the individual cognition only exists as part of this larger functional unit. Furthermore, the group is able to accomplish something that the individual alone cannot, and so the group constitutes a cognitive system in its own right.

Note that although both Hinsz et al. (1997) and Hutchins each use the information-processing model to describe group-level cognitive activity, Hutchins places much more emphasis on sociocultural processes in the sense that they explicitly constitute the cognitive activity that is being described. The methods Hutchins uses to describe the system also are different. Whereas Hinsz et al. employ experimental and laboratory studies, Hutchins's methods are more ethnographic. His data are descriptions of the activities, dialogues, interactions, procedures, manipulations of tools and symbols, and historical events that comprise the computations being performed by the sociocultural system. These descriptions are then organized in terms of information-processing concepts, such as production systems, distributed memory, and parallel processing. Hutchins provides a fascinating comparison of the navigational practices and representational systems of sailors in Micronesia and Western Europe, two cognitive systems that are the same in terms of computational outcomes but differ at the representational/algorithmic and implementational levels of analysis (see Marr, 1982).

One additional point to be made is that the unit of analysis in Hutchins's approach is not fixed. Efforts to bound the cognitive system in either space or time, or at the level of the individual, the work group, or a particular sea voyage will necessarily lead to artificial ideas about where the computational system begins and ends. Social organizations, tools and technologies, and computational practices have been evolving for hundreds of years, and they simultaneously create and are created by coordinated human activity. Thus, any given cognitive analysis involves a pragmatic decision to look at only a particular subcomponent at a particular point in time, but does not represent the actual bounds of the system.

In summary, a viable approach to understanding the sociocultural nature of cognition is to study functional units larger than the individual mind. Such approaches vary along at least two important dimensions: (a) the degree to which the individual mind is seen as a computational or information-processing system within this larger functional unit, and (b) the degree to which information-processing concepts are used to conceptualize the activity of the larger functional system itself. The two frameworks outlined here both employ fairly standard computer models, but other computational metaphors might also be useful for representing group-level cognitive systems, such as connectionist networks or intelligent agents (e.g., Brent & Thompson, 1999).

C. Social Development Theory, Activity Theory, and Situated Learning

The influence of Vygotsky's social development theory is seen largely in a collection of frameworks in which the activity is taken as the unit of analyses, and various components of the activity are analyzed to reveal the relations among activity, society, history, and person. For example, Leont'ev, a student of Vygotsky, proposed that activity be analyzed in terms of the goal, motive, operations, actions, objects, and shared consciousness about the meanings of these things in the context of the activity and its history (Axel, 1997).

Theories of situated learning are related to this idea in that they interpret cognitive development and learning in terms of the activities in which the developing human continuously participates. Knowledge is not an abstract representation in the mind, but rather a change in behavior evolving through interactions among persons, things, and situations in an effort to adapt effectively. Therefore, the locus of learning is not the individual mind, but the process of participation in particular situations. Even general or abstract knowledge takes on meaning only in interaction within a specific situation (Greeno, Moore, & Smith, 1993; Lave & Wenger, 1991).

Although laboratory studies are sometimes employed in these traditions, empirical work more often takes the form of descriptions of learning situations. In general the data are ethnographic, comprising transcripts of activities and conversations that take place in classrooms or apprenticeships, or descriptions of natural behavior that illustrate how traditional theories of knowledge acquisition fall short in describing how people actually learn and express their knowledge. This approach represents a radical departure from the types of theories, methods, and data usually employed in cognitive psychology, and reflects a fundamental difference in the goals and assumptions of the two frameworks. Cognitive psychology assumes that learning and memory are in the mind, and its goal is to build and test models of individual-level mental processes that can be generalized across people. To understand these processes in their pure form, they are isolated from ordinary activity, the individual is isolated from others, and multiple observations are averaged together to provide a ''representative'' description. In contrast, activity-based sociocultural theories assume that learning is a process of mutually constitutive individual, social, and cultural transformation, which takes place through peoples' participation in activities. Activity is the unit of analysis, and the goal is to describe and interpret these processes and transformations by observing them, describing them, and comparing them across multiple natural settings. Ethnographic data often provide interesting accounts of cognition in real-world tasks in ordinary social settings (for example, see Lave's observations of arithmetic strategies among tailors, supermarket shoppers, and Weight-Watchers; Lave, 1988, 1997) and provocative challenges to traditional laboratory-based concepts of cognition, but they often fail to provide the sorts of generalizable principles and testable models favored in cognitive psychology. However, for many investigators a sociocultural approach to cognition requires a major departure from many assumptions and methods that underlie traditional cognitive psychology, particularly the heavy reliance on lab-based studies, which run the risk of becoming so detached from the original behavior of interest that they can lead to incomplete or erroneous descriptions of cognitive behavior (e.g., DeLoache & Brown, 1997; Siegal, 1991). Rogoff (1998) provides a more detailed discussion of some of the key differences in the types of research questions and methods that guide traditional versus sociocultural approaches to the study of cognition.

D. The Collective Memory

The idea of a group mind is an intriguing yet elusive notion. Although it has a certain romantic appeal, the important question is whether it has any psychological or scientific tenability, and whether it represents a potentially

interesting way to conceptualize cognition as a social process. This seems to depend on how one defines it. The idea that there is a high-level, coordinated, organic mind or consciousness shared by all people, such as suggested in Jung's ideas about the collective unconscious and archetypes, has not received scientific support. Bartlett (1932) himself did not believe that the social nature of memory should be characterized in this way, nor that it was likely to be a useful concept to explore. On the other hand, there is no question that people behave differently in groups, and groups often exhibit emergent properties that cannot be predicted simply by knowing about the individuals who comprise the group. The study of collective behavior is an important field of social psychology, examining how cognitive processes such as decision making, attitude formation, and persuasion are altered by group dynamics and lead to important consequences, such as deciding whether to ignore or help a person in need, whether to riot or call for calm, whether to go to war or negotiate peace. With respect to cognitive behavior, the group setting can have powerful influences on perception, recall, and decision-making processes. For example, Loftus has demonstrated that misleading information or false suggestions from other people can be incorporated as facts into one's own memory (Loftus, 1991, 2000).

One possible way to define the group mind or collective memory is in terms of the behavior of the group, that is to treat the group outcome as the unit of analysis. A well-known example of group cognition is the phenomenon of *groupthink,* in which certain conditions lead decision-making groups to make riskier or poorer decisions than the group members would make on their own (Janis & Mann, 1977; McCauley, 1989; Pratkanis & Turner, 1999). Studies of collaborative brain-storming and remembering show that when groups work together they produce fewer ideas and recall less total information than if they work alone and then pool their separate answers (Taylor, Berry, & Block, 1958; Weldon & Bellinger, 1997). Thus, one can think of the group as engaging in cognitive behaviors that produce group-level outcomes, that is, certain outcomes can be viewed as belonging to the group as a whole rather than the individuals per se.

A second possibility is to define collective memory as knowledge that people share either through direct experience or from second-hand sources (e.g., Moscovici & Duveer, 1998). For example, members of different generations and cultures share different experiences, which results not only in different memories but also different identities and worldviews (e.g., Conway, 1997). Collective memories also can be shared across generations through historical accounts, photographs, documentaries, and stories. Different sociopolitical systems also give rise to different collective memories, which are captured and transmitted through media such as news reports,

films, books, encyclopedias, and so forth. For example, Table I shows descriptions of the Cuban missile crises taken from two different sources, the *Great Soviet Encyclopedia* (Prokhorov, 1983) and the *American Academic Encyclopedia* (1996), which would create entirely different shared accounts among their respective audiences. Collective memory also is reflected in many of the references people use to communicate their own perceptions of events, as well as to influence other peoples' interpretations of events. For example, a victim of recent tornadoes in Oklahoma declared, ''Hiroshima . . . It just looked like the pictures of Hiroshima'' (Sack, 1999). Even though she herself had not witnessed the bombing of Hiroshima, nor have many people in the United States, it represents such a significant event in our shared knowledge that the single word can be used to convey a vivid picture of the destruction she was witnessing as well as her emotional state. Ideally, the study of this type of collective memory should examine not only the content of the shared information, but also the processes that lead to the establishment, maintenance, and transformation of collective memories and collective amnesias, as well as the implications for peoples' decisions and actions in relevant situations. Some preliminary efforts to measure and characterize collective memory by using the mathematics of social network analysis are presented in the last part of this chapter.

A third concept of group-level memory is the idea of *transactive memory* proposed by Wegner (1987). Wegner suggests that people in ongoing relationships develop a shared memory system in which a set of individual memory systems are combined through communication among the individu-

TABLE I

ACCOUNTS OF THE CUBAN MISSILE CRISIS

From the *Great Soviet Encyclopedia*[a]	From the *American Academic Encyclopedia*[b]
Through a number of political and diplomatic measures the Soviet government worked to alleviate the grave international crisis and eliminate the threat of an invasion of Cuba.	After nearly two weeks of unprecedented tension, the Soviet government of Nikita Khrushchev yielded.
The steps taken by the Soviet Union and the firm resolve of the Cuban people . . . forced the government of the USA to evaluate the emerging situation more soberly.	Despite this concession, all sides regarded the outcome as a substantial victory for the United States, and Kennedy won a reputation as a formidable international statesman.

[a] From the Prokhorov (1983), copyright 1983 by Macmillan. Reprinted with permission.
[b] From the *Academic American Encyclopedia* (1996), copyright 1996 by Grolier.

als. Members of the system are aware not only of what information they themselves know, but also what the other members know, a type of social metacognition (Jost, Kruglanski, & Nelson, 1998). Over time, particularly in task-oriented groups and close social relationships, a division of labor is established such that responsibility for remembering certain things is allocated to different members, typically based on interest or expertise. For example, I have a terrible memory for what movies I have seen, but this is not a handicap at the video store because my husband has an excellent memory in this domain, so I simply pick up a video and ask him if I've already seen it. On the other hand, I have a good memory for places so I keep track of what services and attractions are available where. Thus, over time we have developed a transactive memory system through which we keep track of numerous things, more than either of us could do alone. However, note that each person in the system individually lacks critical pieces of information.

Wegner conceives of the transactive memory system as a property of the group, and as such a group mind. It is an information-processing system that emerges through communication between group members, is constructed over time, and is not traceable to any individual alone. The system impacts what the group remembers as a whole. Nevertheless, there is a reciprocal relation between the individuals and the group-level system because the propensities of the individuals influence how the system develops, and the emergent system influences what the individuals remember both within and outside the group.

To demonstrate the operation of a transactive memory system, Wegner, Erber, and Raymond (1991) recruited couples who had been dating at least three months. During the experiment, participants were tested either with their own partner or were re-paired with a person from another couple. Each pair was given a set of 64 different sentences to learn, with 9 or 10 sentences representing each of several different categories of information (e.g., statements about television, science, alcohol, etc.). In the assigned-expertise condition, each member of the pair was assigned special responsibility for learning a particular subset of the categories. In the no-assignment condition, the members were simply instructed to learn the items. When groups were left to their own devices (the no-assignment condition), intact couples remembered more total information than re-paired couples. However, when the couples were assigned categories, re-paired couples remembered significantly more than intact couples. Thus, the transactive memory system develops in such a way that the members optimize group performance, but if the organization of the system is disrupted then it actually underperforms. Recent findings about collaborative remembering in older couples lends additional support to this idea. Although elderly individuals

exhibit memory deficits relative to younger adults, when elderly couples who have been married for 40 years or more are allowed to work together, they remember just as much as young couples (Dixon & Gould, 1996).

In sum, collective memory can be conceptualized in such a way as to render it measurable, testable, and psychologically interesting. This is done by evaluating group-level behaviors and knowledge. At the moment, techniques are being developed to examine the content of these group memories, as discussed later in this chapter, and work also is underway to better characterize the group-level processes that comprise the transactive system (e.g., Wegner, 1995).

E. Summary

The purpose of this section was to discuss examples of some of the major conceptual frameworks and methods that have been developed in efforts to explore the social and cultural nature of cognition and memory. There are many different ways that one can conceive of remembering as a social process, as demonstrated by the important differences in the starting assumptions, units of analysis, and methods of these various frameworks. The approaches include (a) social cognition, in which social information is fed into the individual's information-processing system, just like any other stimulus; (b) the social-influence approach, in which culture is treated as an independent variable or contextual variable that can affect the structure and content of cognitive processes; (c) efforts to look at cognition as computations or information processing in a larger functional system, such as a person-and-tool, a small group, a society, or a culture; (d) efforts to describe the socioculturally situated activities of goal-directed individuals and groups; and finally (e) the concepts of group mind and collective memory. These are not mutually exclusive approaches, and many share some concepts while having distinct differences in other respects.

The discussion was not intended to provide a comprehensive review of the dozens of specific frameworks and arguments that have been developed, nor to review all the relevant empirical literature, nor to champion any particular perspective. Rather, the intent was to explain some of the issues that have motivated investigators' interest in understanding cognition as something more than computations in the individual mind. I also tried to point out some of the important differences among the various approaches, and highlight some of the issues that might be of particular concern to a cognitive psychologist. Although many of these frameworks talk about cognition in general rather than memory in particular, there is no reason that the arguments cannot be extended to memory, and I tried to illustrate the concepts with memory examples when possible.

In the last part of this chapter, I will discuss some of the empirical work that I have conducted in efforts to take a more social perspective on memory processes. I confess up front that this work does not provide clear-cut examples or paradigm-shifting demonstrations of the concepts discussed above, and it is better seen as my initial efforts to straddle both my traditional training as a cognitive psychologist and my growing interest in remembering as a social process.

IV. Collaborative Remembering and Collective Memory

I first became interested in the social aspects of remembering several years ago, before our psychology department had a good system for keeping minutes at faculty meetings. The beginning of each meeting would be spent trying to reconstruct what had happened at the last meeting. I wondered how collaborative remembering is similar to and different from individual remembering, and whether it exhibited similar principles of interference, reconstruction, encoding, retrieval, and so forth. This quickly gave rise to a variety of interesting questions, such as, are there important differences between group and individual memory? What impact does the individual have on group remembering, and the group on individual remembering? What aspects of the social setting affect what is remembered and how it is remembered? When multiple accounts of a situation are possible, who or what determines the emergence of the prevailing memory? How is historical memory different from memory for experienced events? How can social and cultural factors be accounted for in the description of group memory? What are the appropriate units of analysis, the individuals within the sociocultural context, the groups as a whole, or the activity itself defined by the tasks, goals, and sociocultural setting? Can collective memory be characterized and studied in a meaningful way? What kinds of models and conceptual frameworks can provide useful theoretical tools for studying remembering as a social process?

These questions represent broad issues in this domain, but my initial empirical work has been more modest in scope, trying to get a handle on some basic behavioral phenomena, some of the methodological tools that can be employed to study social aspects of memory, and some theoretical concepts that might be useful. This section is divided into two parts in which I illustrate two general lines of work in this effort, the first concerning the process of collaborative group remembering, and the second outlining a very preliminary effort to operationalize the concept of collective memory by adopting some methods from social network analysis.

A. Collaborative Remembering

1. Collaborative Inhibition

What happens when people work together to remember a past event? In an effort to study the group as a cognitive system, Weldon and Bellinger (1997) conducted two experiments to examine collaboration in recall. Most peoples' intuition is that, as a general rule, collaboration improves performance, a belief that underlies familiar problem-solving strategies such as brainstorming. In fact, it is usually the case that when a few people work together to generate ideas or remember information, the group does do better on the average than individuals working alone (e.g., see review in Clark & Stephenson, 1989). But it is also the case that when people work together, the group produces less than the sum of the individuals working alone, an outcome that has often been interpreted as *social loafing* (see review in Karau & Williams, 1993). It is interesting to note that this interpretation reflects our bias to view solitary behavior rather than social behavior as normative, even in activities that are usually social (D. L. Medin, personal communication, January 2000). This ''group deficit'' has been reported for many physical and cognitive tasks such as rope-pulling, clapping, shouting, vigilance tasks, and brainstorming. But what about remembering?

From the standpoint of a memory researcher, there are two equally plausible predictions about what might happen in remembering. First, collaboration could improve performance through the process of *cross-cuing,* in which the recollections of one person serve as cues to help others remember related information (Meudell, Hitch, & Boyle, 1995; Meudell, Hitch, & Kirby, 1992). On the other hand, the process of collaboration could produce interference in which the recall of some items interferes with peoples' ability to access the remaining material. This could be analogous to the part-list cuing effect, a phenomenon in which providing some of the to-be-remembered items to the rememberer actually impairs the person's ability to recall the remaining items (e.g., D. Basden & Basden, 1995; Roediger & Neely, 1982; Slamecka, 1969).

To answer this question, K. D. Bellinger and I conducted two experiments (Weldon & Bellinger, 1997). In Experiment 1, people were tested in groups of three and studied mixed lists of pictures and words. After a brief distractor task, the participants recalled the items either individually (people worked alone and wrote their answers on their own paper), or collaboratively (all three people recalled aloud and one person in the group wrote their answers on a single paper). The number of targets recalled was scored for both the individuals and the collaborative groups. However, the key feature was the scoring of *nominal groups,* which are groups in name only, and are formed by pooling the nonredundant responses of individuals who

recalled alone. For example, in the individual recall condition, Person 1 might recall *hat, bean, geyser, zebra,* Person 2 *hat, bean, martini,* and Person 3 *hat, fan, book, chalk, puzzle.* Average individual recall is (4 + 3 + 5)/3 = 4 items, but nominal group recall is 9 items (*hat, bean, geyser, zebra, martini, fan, book, chalk, puzzle*), which represents the total number of items that are actually recallable by the group, at least in that circumstance. Nominal group performance can be thought of as representing the potential recall of the group.

The results of the experiment are presented in Figure 4 (a subset of the original data). Overall, collaborative groups recalled more than individuals, which is not surprising and replicates other work (e.g., Clark & Stephenson, 1989). The more interesting finding is the comparison between collaborative and nominal groups, which shows that collaborative groups recalled significantly less than the pooled individuals, an outcome we have called *collaborative inhibition.* In other words, when people collaborate they recall less than their potential (also see Lorge & Solomon, 1961). Similar results have been obtained with dyads (Andersson & Rönnberg, 1995), although nominal dyads sometimes exhibit no difference from collaborating dyads (Meudell et al., 1992).

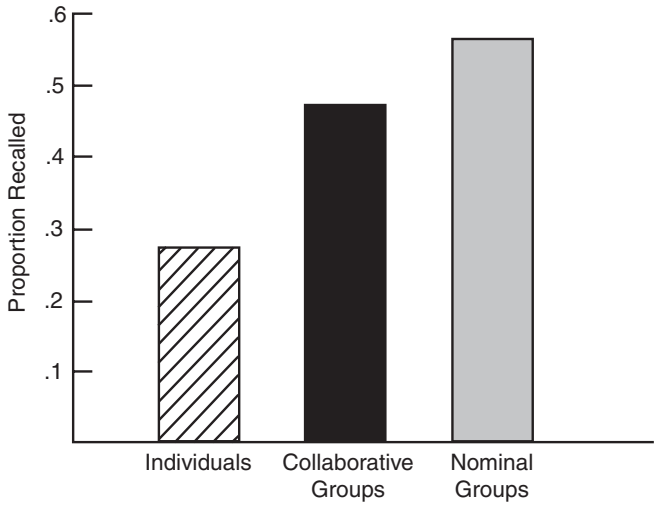


Fig. 4. Collaborative inhibition in recall of a list of picture and words, showing that collaborating groups recall less than their potential. (From ''Collective memory: Collaborative and individual processes in remembering,'' by M. S. Weldon and K. D. Bellinger, 1997, *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* p. 1164. Copyright 1997 by the American Psychological Association. Reprinted with permission.)

One concern about the generality of this finding is that list recall may be a relatively artificial task for groups, even more artificial than for individuals. In social settings, people probably rarely sit around trying to remember lists of disconnected items. More likely, they try to reconstruct meaningful connected information that they want to share such as stories, movies, news reports, and events they have experienced together. Therefore, people may have learned strategies for effectively collaborating to recall rich information like events and stories, but do not or cannot readily generalize these strategies to help them recall random word lists. In addition, stories have a hierarchical structure or grammer (e.g., Mandler & Johnson, 1977), and if people share these organizational schemes they may be able to use them to cross-cue each other and improve recall. In Experiment 2 we tested story recall to see if it would benefit from collaboration.

We recorded the story "War of the Ghosts" (from Bartlett, 1932) twice on an audiotape, and the two recording were played in succession for each group of three participants. After a brief distractor, participants were instructed to recall the material either individually or collaboratively. Recall was scored using the 42 idea units suggested by Mandler and Johnson (1977) and the results are presented in Figure 5. Once again, although

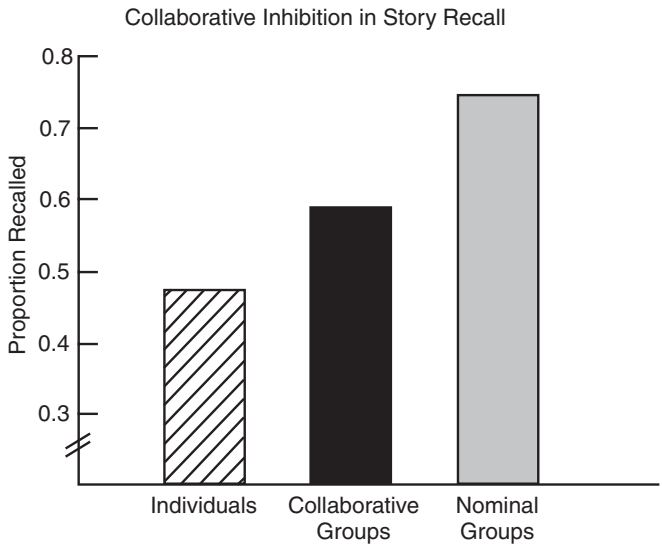


Fig. 5. Collaborative inhibition in recall of "The War of the Ghosts." (From "Collective memory: Collaborative and individual processes in remembering," by M. S. Weldon and K. D. Bellinger, 1997, *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* p. 1169. Copyright 1997 by the American Psychological Association. Reprinted with permission.)

collaborative groups recalled more than single individuals, they recalled significantly less than nominal groups, replicating the collaborative inhibition obtained with word lists. Therefore, the structured story did not eliminate poor performance in the collaborating group.

Why do people recall less than their potential when they work together in groups? We suggested a few possibilities. First, when people work in groups they may be less motivated to contribute to the group, for any number of reasons. For example, group members may engage in social loafing (Latané, Williams, & Harkins, 1979) and not put forth their best effort either because (a) they experience a diffusion of responsibility, feeling less personally responsible for the group outcome; or (b) they want to maintain equity of effort and not overwork themselves while others slack off; or (c) they do not feel personally accountable because their own contributions cannot be uniquely identified (Karau & Williams, 1993). Another possibility is that although people may not be loafing per se, they may feel more timid in a group and set a high standard for the value or correctness of their contribution before they will speak up, which will lead to fewer contributions overall. In a recent series of studies designed to test this motivation hypothesis, Weldon, Blair, and Huebsch (in press) manipulated motivation in various ways, such as (a) offering a large cash reward to the group that recalled the most information, (b) setting a minimum level of recall that the group had to achieve before they could leave the laboratory, and (c) increasing group cohesion. Overall, motivation did not appear to be correlated with the level of collaborative inhibition. Even with high levels of motivation, collaborating groups recalled significantly less than nominal groups, suggesting that impaired recall is not due to reduced motivation to put forth one's best effort in the group. This is not to say that there cannot be circumstances under which reticence or social loafing do occur during collaborative recall. For example, it is likely that people will slack off in very large groups or situations in which they are relatively anonymous, unlike the present experiments in which the groups were small and each person was relatively conspicuous. Also, there might be situations in which the risk of embarrassing oneself is truly salient, and staying quiet seems like a better option than saying something incorrect. Nevertheless, these experiments demonstrated that above and beyond motivational factors that may potentially operate in some situations, there are other processes that give rise to impaired performance in collaborative recall.

An alternative hypothesis is that collaborative recall produces retrieval interference. This idea was proposed and tested by Basden, Basden, Bryner, and Thomas (1997), who suggested that when a person hears the other group members recalling the material aloud, it disrupts his or her own idiosyncratic organization of the material and thereby disrupts strategies

for recalling it, leading to lower recall. They reasoned that if material is highly organized such that there is little room for individualized organization, collaborative inhibition should disappear. To test this they presented participants with lists containing either fifteen 6-item categories, or six 15-item categories. The logic is that the small categories do not allow for much variety in how individuals organize the items, and so there should be considerable overlap among individuals' organizations. Therefore, for small categories retrieval should not be very susceptible to disruption because each person's strategy will be pretty similar to the others'. In contrast, the items in the 15-item categories will probably be organized in a less consistent fashion across individuals. For example, the category labels might be relatively more overloaded as cues, and participants will need to resort to other strategies for organizing, interrelating, and cuing the information. In this case, recall will be more susceptible to disruption, and so collaborative inhibition should occur. These predictions were borne out, and the results are presented in Table II, which presents both category recall and instance recall. Category recall is the number of categories represented in recall, without regard to the number of items recalled from each. For this measure there are ceiling effects with the six large categories, but there is significant collaborative inhibition when fifteen categories were represented in the list. With so many categories, there is room for idiosyncratic organization and hence disruption in group recall. In contrast, when one looks at the number of items recalled per category (conditionalized on having recalled at least one item from the category), a different pattern emerges. When the categories are large, that is with the 15-item categories, collaborative inhibi-

TABLE II

Category and Instance Recall in Collaborative and Nominal Groups[a]

	Six 15-item categories		Fifteen 6-item categories	
Recall measure	Collaborative	Nominal (Pooled individuals)	Collaborative	Nominal (Pooled individuals)
Category recall	6.0	6.0	12.1	13.5
Instance recall	6.86	8.11	2.9	2.9

[a] Category recall refers to the number of categories represented by the recall of at least one item from that category; the number of items recalled in each category is not reflected in this measure. Instance recall refers to the average number of items recalled in each category. (Adapted from Basden, Basden, Bryner, Thomas, 1997).

tion is evident in instance recall. However, when the categories are small, as with the 6-item-categories, collaborative inhibition disappears. The results support the retrieval disruption hypothesis. When circumstances are such that people tend to have similar organization of the material, hearing other peoples' recalls is less likely to be disruptive, and collaborative inhibition should be eliminated or significantly reduced.

Additional evidence supports this hypothesis. For example, when people study word pairs (e.g., *coat–apple*) and then receive the stimulus term (e.g., *coat*) as a retrieval cue, collaborative dyads retrieve just as many words or slightly more than nominal dyads (Finlay, Hitch, & Meudell, in press). Presumably the retrieval cue coordinates retrieval processes in the group.

In a related idea, Andersson and Rönnberg (1997) hypothesized that collaborative inhibition arises from reduced *cue specificity* in the collaborating group, that is to say collaborative recall suffers because groups do not communicate sufficiently well to produce retrieval cues that enable them to access the desired information. In contrast, individuals can optimize their own performance by using cues that provide the most useful retrieval information for themselves (see Mäntylä & Nilsson, 1983). To test this, during an encoding phase dyads worked either individually or collaboratively to generate associates to 90 target words (e.g., see *telephone,* generate *call* as associate). In the collaborative encoding condition the two participants had to agree on the associate. During the retrieval phase, participants received their associate (e.g., *call*) as retrieval cues, with half the dyads in each of the original encoding conditions working individually and half collaboratively. The results are shown in Table III. Andersson and Rönnberg noted that overall, collaboration leads to worse performance than working alone and pooling the results (nominal dyads), showing that groups cannot optimize their use of retrieval cues, and supporting the *cue specificity*

TABLE III

PROPORTION OF WORDS RECALLED AS A FUNCTION OF ENCODING AND RETRIEVAL CONDITIONS FOR INDIVIDUALS AND DYADS[a]

	Retrieval condition	
Encoding condition	Collaborative dyad	Nominal dyad (pooled individuals)
Individual generation	.42	.80
Collaborative generation	.67	.61

[a] From "Cued memory collaboration: Effects of friendship and type of retrieval cue," by J. Andersson J. Rönnberg, 1997, *European Journal of Cognitive Psychology, 9,* pp. 273–287. Copyright 1997 by the *European Journal of Cognitive Psychology.* Reprinted with permission.

hypothesis. This interpretation is also consistent with Basden et al.'s (1997) idea that collaboration leads to disruption of individuals' idiosyncratic organization of material and retrieval strategies, and suggests that people can optimize strategies for themselves but have difficulty when they have to coordinate with someone else.

The authors' interpretation captures one aspect of their data, but another perspective also jumps out. The data exhibit an encoding-retrieval interaction, a pattern that corresponds with the predictions of the encoding specificity and transfer-appropriate processing frameworks (Tulving & Thomson, 1973; Morris, Bransford, & Franks, 1977, respectively). These frameworks hypothesize that retrieval will benefit to the extent that encoding and retrieval processes are similar, and so suggest another way to think about collaborative inhibition. Specifically, in Andersson and Rönnberg's data, one can see that in the Individual-generation encoding condition the usual collaborative inhibition effect is obtained, because collaborating dyads recalled less ($\bar{x}$ = .42) than nominal dyads ($\bar{x}$ = .80). In fact, even when the collaborative dyad was given both cues that the two individuals had generated separately during encoding, they still performed worse than the nominal dyads (for collaborative recall with two cues, $\bar{x}$ = .67, versus $\bar{x}$ = .80 for the nominal dyads)! This half of the design is the same as that used by Weldon and colleagues and Basden et al., that is, people study the material individually but then recall it either individually or collaboratively. The interesting aspect of Andersson and Rönnberg's data occurs in the collaborative encoding condition. When dyads collaborated at encoding, collaborative inhibition in retrieval not only disappeared, but collaborating dyads actually remembered more than the nominal dyads composed of two individuals who recalled alone. Thus, there is a crossover interaction showing that collaborative inhibition can at least be partly attributed to a mismatch between encoding and retrieval conditions, and this selectively disadvantages collaborative groups when they have encoded the information individually. However, when collaborative encoding takes place, then individual recall is at a disadvantage (collaborative dyad $\bar{x}$ = .67; nominal dyad $\bar{x}$ = .61; a statistical test was not reported for this particular comparison). Of course, there is still an overall disadvantage of collaboration with this task, which was the main interest of Andersson and Rönnberg, but the crossover interaction suggests that collaboration effects must also be interpreted in terms of the relation between the encoding and retrieval situaton.

These data also are compatible with the ideas of Basden et al. (1997) because they show that when people share an organizational scheme, group retrieval will be good. However, these data extend their findings by showing that the principle works the other way as well—if people initially share an

organization it can be disrupted by preventing them from working together, and thereby impair their retrieval.

2. *Does Collaborative Inhibition Occur on Other Tasks?*

In presenting findings on collaborative inhibition, I am often met with extreme resistance by the audience. The result seems to fly in the face of most peoples' intuitions or assumptions about what should happen, which is that collaboration should be beneficial. As evidence, people often mention the experience of sitting around with friends reminiscing about a shared experience, and how when someone cannot remember a particular detail, everyone joins in the effort to reconstruct the event, offering hints and tip-of-the-tongue information until the desired fact is remembered. This kind of experience certainly gives the impression that collaboration helps retrieval, although in reality it is only demonstrating that a group can remember more than an individual, which many data already show. Nevertheless, it does raise questions about how generalizable the phenomenon is for different kinds of retrieval tasks. For example, the free recall tasks used by Basden et al. (1997) and me (Weldon & Bellinger, 1997; Weldon et al., in press) are episodic memory tasks that require the retrieval of relatively disconnected and poorly learned material. Even the story ''War of the Ghosts'' used by Weldon and Bellinger can be considered impoverished because of its unfamiliar content and vague story line.

Perhaps collaboration will help retrieval of relatively richly interconnected material that people have learned over longer periods of time or with which they have many personal associations, such as movies, vacations, or important historical events. Such material might enable people to provide and use cross-cuing more effectively. For example, the process might unfold something like the following dialogue between two people trying to remember the name of a movie:

> ''Remember, it happened the night after we saw . . . What was that movie? Remember? It was the one where Arnold went to Mars or something.''
>
> ''Oh, yeah . . . Darn, what was that called? Remember, I couldn't watch the part where the person got a hole in his space suit and his eyes popped out.''
>
> ''Shoot, what . . . oh, yeah, I think that was also the movie where Sharon Stone made her big debut. I remember I didn't think much of her, but I guess everyone else did. (laughs) And he gets his memory replaced or implanted in this place where you can go and have a vacation just by having a chip inserted in—''
>
> ''*Total Recall!* That was it, *Total Recall.*''

In this case, there is a rich set of information and clues that the two people build upon to recall the information, which is difficult to do with word

lists. Therefore, the benefits of collaboration may not be able to manifest themselves with relatively impoverished materials.

To better test the generalizability of collaborative inhibition I conducted pilot work with a few new tasks, a map-reconstruction task (which has often been suggested to me as a task on which collaboration would be beneficial), a general knowledge test, and reconstruction of the Rey-Osterreith nonsense figure (Osterreith, 1944). These tasks were selected to provide retrieval demands different from the word-list and story recall used in the earlier experiments. First, whereas list and story recall tested memory for a specific episode, the map reconstruction and general knowledge tasks test memory for generic knowledge. The distinction between these two classes of retrieval tasks, episode-specific tests versus generic knowledge tests, has been considered theoretically important because they often dissociate, showing contrasting patterns of results. Tulving has hypothesized that they actually tap into two functionally independent memory systems, which he calls episodic and semantic memory (Tulving, 1985). Second, the Rey-Osterreith line-drawing task is potentially interesting because it assesses recall of a stimulus that is relatively more visuospatial, and is also an episode-specific task. In sum, the Rey-Osterreith, word-list, and story recall tests all assess episode-specific memory, whereas the map-reconstruction and general-knowledge tests assess generic knowledge. In addition, the map-reconstruction and Rey-Osterreith tests have a visuospatial component, whereas the general-knowledge, word, and story recall tasks are primarily verbal.

These new tests were included as distractor tasks in other experiments, and participants received only one of them. The data presented here should be considered pilot results, but provide some interesting preliminary findings. For all tasks, participants were tested in groups of three, and worked either collaboratively or independently. Nominal data were computed by pooling across the three individuals in the individual recall condition. The data are presented in Table IV, and the number of groups tested in each condition are presented in parentheses.

For the map-reconstruction task, participants were given an outline of the 48 contiguous United States, and asked to fill in as many states as possible. They were also told that if they could remember a state but not its location, they should simply write the name of the state below the map. In the collaborative condition participants were more lively and interactive than with list recall; they seemed to enjoy the task quite a bit. An example of one of the maps drawn by an individual is shown in Figure 6a, and an example of a collaboratively produced map in Figure 6b. The maps were scored two ways, (a) total number of states recalled, even if not drawn on the map; and (b) the number of state borders correctly adjacent on the

TABLE IV

PROPORTION CORRECT IN COLLABORATIVE AND NOMINAL GROUPS ON MAP RECONSTRUCTION, REY-OSTERRIETH FIGURE RECALL, AND THE GENERAL KNOWLEDGE TEST

	Recall condition[a]	
Type of test	Collaborative group	Nominal group (pooled individuals)
Map reconstruction		
State names	.91 (20)	.95 (20)
Borders	.50 (20)	.58 (20)
Rey-Osterrieth figure	.61 (10)	.69 (10)
General knowledge test		
Question Set 1	.58 (11)	.49 (11)
Question Set 2	.37 (41)	.30 (29)

[a] Numbers in parentheses indicate the number of groups tested in each condition.

drawing. A one-tail *t*-test reveals a significant disadvantage for the collaborative groups in state-name recall, $t(38) = 1.79$, $p < .04$. For border recall, the mean difference is fairly large but only marginally significant, $t(38) = 1.44$, $p < .08$, probably due to the large variance in this measure. Thus, although these data should be viewed cautiously, they provide no support for the hypothesis that the more interactive collaboration that accompanies map reconstruction leads to beneficial effects of collaboration.

The Rey-Osterreith figure (Osterreith, 1944; see Figure 7) is normally used as a neuropsychological assessment tool, but the task can be made sufficiently difficult to test individuals with normal memory. Participants looked at the figure for 3 min, then drew it after a 12-min filled delay. Figure recall was scored by counting correct reconstructions of 18 specific components. The means shown in Table IV exhibit a trend toward a collaborative inhibition effect in this task. Although the effect is not significant, there is low power in the comparison so more data would need to be collected to establish the reliability of this difference. In conjunction with the results of the map-reconstruction task, these results indicate that there is nothing unique about verbal materials as opposed to visuospatial materials in terms of their susceptibility to interference during collaborative retrieval.

For the general knowledge test participants were given a set of relatively difficult questions from Nelson and Narens's (1980) norms. These questions probe knowledge of history, geography, science, arts, and popular culture, for example, ''What is the name of the Chinese religion founded by Lao

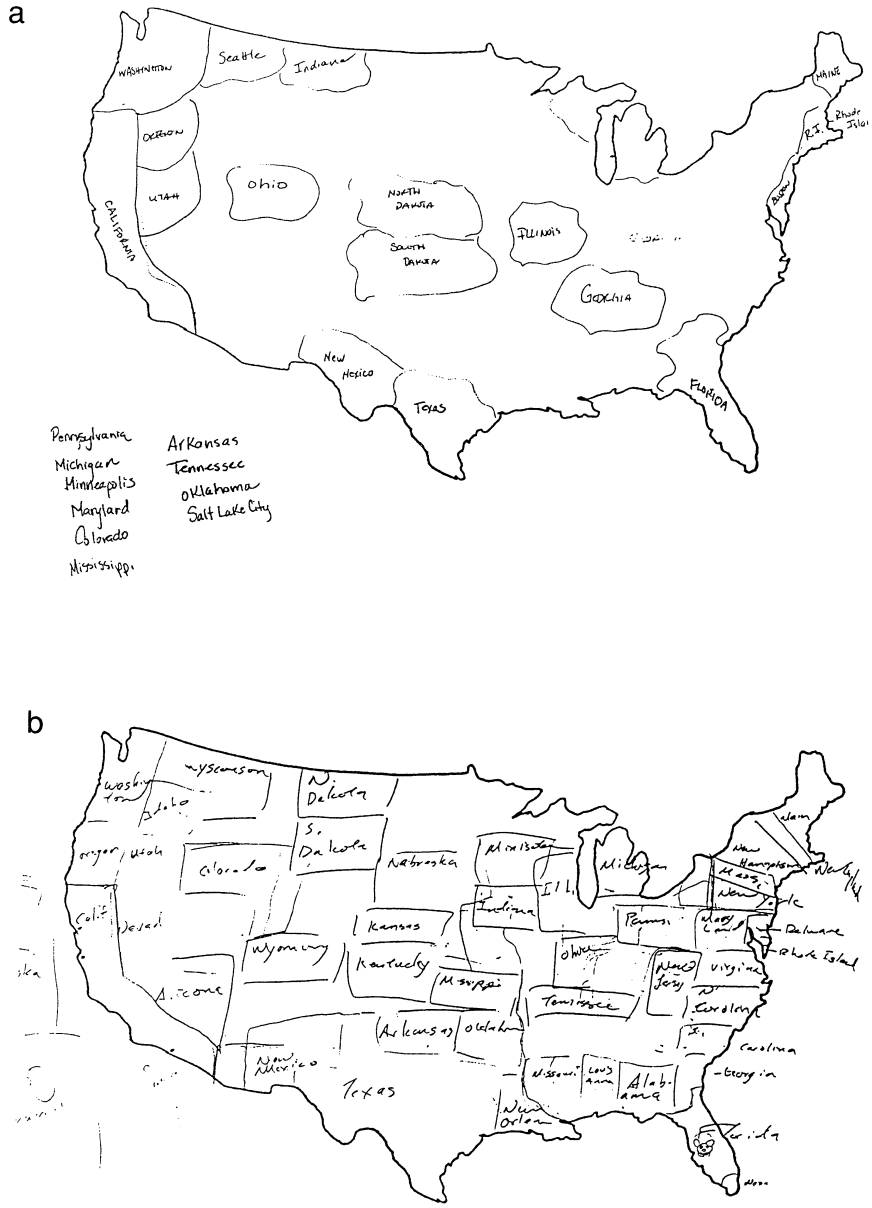


Fig. 6. Examples of map reconstructions. (a) Drawn by an individual working alone; (b) drawn by three people collaborating.

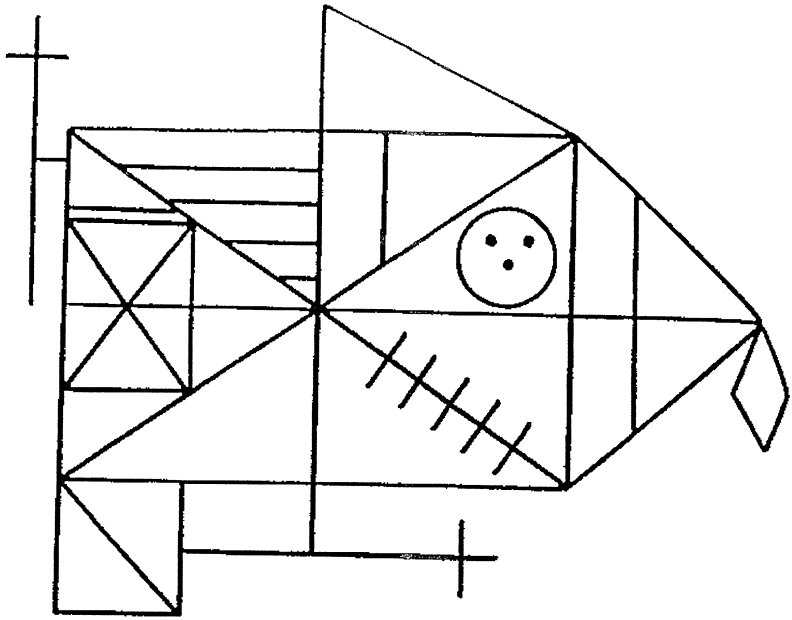


Fig. 7. The Rey–Osterreith figure (From Osterreith, 1944, reprinted with permission of Editions Médicine et Hygiène.)

Tse?'' and ''What American track athlete was the star of the 1936 Olympics?'' Two different question sets were used in two different experiments. Set 1 contained 35 questions and was tested on 11 groups in each condition, and Set 2 contained 46 questions and was tested on 29 groups in the nominal condition and 41 groups in the collaborative condition. The results are shown in Table IV, and show a reversal from the other tasks. Not only was collaborative inhibition eliminated, but in both experiments there was a noticeable trend toward better performance in the collaborative conditions (for set 1, $t(20) = 1.19$, $p = .25$; for set 2, $t(68) = 1.71$, $p = .09$; both two-tailed tests). These data provide credible evidence that collaborative inhibition is eliminated on the general knowledge test, and even suggest that collaboration benefits performance on this task (although more work is needed to determine if this latter trend is reliable). What is it about the general knowledge test, in contrast with free recall, map reconstruction, and figure recall, that protects it from collaborative inhibition? The key feature may be that it involves cued recall.

To explore this, the results of some of the experiments discussed so far are organized in Figure 8 (the episode-specific cued recall data refer to Finlay et al., in press, reported earlier). Each memory test is classified according to whether it is an episode-specific or a generic knowledge task, and whether it involves free recall or cued recall. My initial hypothesis was that generic memory test would be immune to collaborative inhibition, or even show benefits of collaboration because people would have richer associations and knowledge with which they could cross-cue each other. However, the distinction between episode-specific and generic knowledge does not seem to account for the pattern of collaborative inhibition effects. Instead, the distinction between cued and free recall appears to be the important factor. This is consistent with Basden et al.'s (1997) retrieval disruption hypothesis. In free recall, people have idiosyncratic organizations of the material, and retrieval strategies are disrupted by hearing other members in the group recall the items in different orders. However, cued recall can largely bypass dependence on individual organization because the cues provide more distinctive, direct, target-specific information to facilitate retrieval. Furthermore, cues may serve to coordinate the efforts of the collaborating group. In free recall different individuals are frequently searching for different items, or often just any item from the list with no particular target in mind. With retrieval cues, the retrieval processes of the group are more controlled because the cues point everyone in the same direction, such as remembering the target that was associated with *coat,* or the track star of the 1936 Olympics.

The hypothesis that cued recall eliminates collaborative inhibition is preliminary, based on results from different experimenters, labs, tasks, and materials. It needs to be verified through more systematic testing, which is currently underway in my lab. In addition, the processes of collaboration

	Episode-Specific	Generic Knowledge
Free recall	Word recall: Coll < Nom Story recall: Coll < Nom	Map recall: Coll < Nom
Cued recall	Word pair cued recall: Coll = Nom	General knowledge: Coll ≥ Nom

Fig. 8. Effects of collaboration on different retrieval tests, organized by two important task dimensions. The word-pair cued recall results are from Finlay et al. (in press), and the remaining data from Weldon (Weldon & Bellinger, 1997; Weldon, this chapter).

need to be analyzed more carefully to understand how cues affect group performance. For example, do they help by improving item-specific retrieval, by coordinating group retrieval strategies, by improving the efficacy of cross-cuing, or combinations of these factors? The answers to these kinds of questions will yield important insights into the relation between individual and group retrieval processes, and will be helpful in determining what kinds of models will best capture the dynamics and outcomes of collaborative remembering.

B. Collective Memory

As mentioned earlier, the concept of collective memory usually brings to mind a variety of ideas that are not considered to be scientifically tractable, such as Jung's theory of universally shared archetypes or Freud's ideas about universally shared dream symbols. Nevertheless, testable concepts have been proposed, such as Wegner's idea of transactive memory (e.g., Wegner, 1987). Recently, I have been exploring the possibility of using the methods of social network analysis to characterize collective memory, and in this last section I explain the technique and present some data illustrating its use.

1. Social Network Analysis: Methods for Studying Relational Data

Social scientists use network analysis to study relational structures such as the patterns of interaction among members of a social group, and the methods are commonly employed in sociology, anthropology, economics, and in some areas of psychology. For example, the concepts have been used to formalize and test Heider's balance theory, the idea that the valances of attitudes and relationships among the individuals within a group should have certain symmetries (e.g., Cartwright & Harary, 1956; Heider, 1946). Social network analysis uses a wide variety of graphical, mathematical, and statistical tools, of which only a tiny subset is presented here. A very accessible introduction to these basic ideas is provided by Scott (1991), and a more extensive treatment is available in Wasserman and Faust (1994). The terminology varies slightly in these sources, and I will use the terms used by Scott. The analyses presented here were performed using UCINET 5.0 (Borgatti, Everett, & Freeman, 1999).

Relational data can take a variety of forms. To illustrate, imagine that one wants to explore an intellectual network in terms of particular scientists and their affiliations through professional activities. Figure 9a illustrates this type of relationship in a matrix in which the rows represent individual cases such as professors, and the columns represent their affiliations, for

Affiliations
(Conferences or Idea units)

Cases (Professors)	A	B	C	D	E
1	1	1	1	1	0
2	1	1	1	0	1
3	0	1	1	1	0
4	0	0	0	0	1

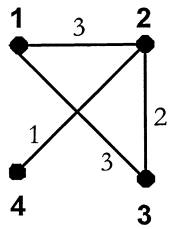
(a) Incidence matrix: Cases by affiliations

	1	2	3	4
1	4	3	3	0
2	3	4	2	1
3	3	2	3	0
4	0	1	0	1

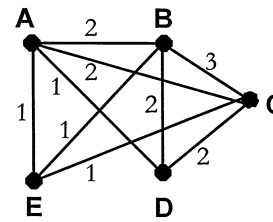
(b) Adjacency matrix: Cases (Professors)

	A	B	C	D	E
A	2	2	2	1	1
B	2	3	3	2	1
C	2	3	3	2	1
D	1	2	2	2	0
E	1	1	1	0	2

(c) Adjacency matrix: Affiliations (Conferences or Idea units)



(d) Graph: Cases (Professors)



(e) Graph: Affiliations (Conferences or Idea units)

Fig. 9. Relational matrices. See text for explanations of the traditional use in describing social networks and a proposed use for describing collective memory.

example, what conferences they have attended during the past year. A "1" in a cell indicates the professor attended the conference, and a zero that he or she did not. This case-by-affiliations *incidence matrix* thus represents which conferences were attended by each individual, and simultaneously which individuals attended each conference. (Later this network will be

interpreted in terms of collective memory by substituting "idea units" for conferences.)

The incidence matrix can be transformed into two *adjacency* matrices. One represents the connections among the individuals in terms of how many conferences each pair of individuals attended in common (Figure 9b), and in this example one might see this as an index of their shared interests or their opportunities to collaborate. For example, Professors 1 and 2 attended three conferences in common, whereas Professors 1 and 4 attended none in common. Notice that this type of relationship is *nondirectional,* meaning that if Professor 1 attends a conference with Professor 2, then Professor 2 is attending the conference with Professor 1. Thus, the matrix is symmetrical because the relations in this network are symmetrical by definition. Other types of data relationships may be directional, or not necessarily reciprocal, such as asking a colleague for an opinion.

These data are also *valued,* meaning that the relationship is quantifiable, in this case it represents the number of conferences attended in common. This contrasts with other types of data which are *nonvalued* or dichotomous, meaning that the relationship either exists or it does not (e.g., two people either have met or have not met; also see the incidence matrix in Figure 9a). The methods of network analyses are better developed for dichotomous than valued data, which limits the techniques that will be of use here because the data of interest are valued. Finally, note that the diagonal in this example simply represents the total number of conferences attended by each individual. The diagonal will not be of interest in the methods explored here.

The second adjacency matrix (Figure 9c) represents the connections among the conferences in terms of how many professors each pair of conferences had in common. This can be seen as an index of how the conferences are linked through patterns of attendance. For example, conferences B and C shared three professors, whereas conferences D and E shared none. This might suggest that conferences B and C are more important or capture more central themes with respect to the interests of this group of professors.

The relational matrices can be represented in graphic form. For example, the adjacency matrix representing the relations among the professors is illustrated in graphical form in Figure 9d, with the nodes or vertices representing the individual professors, and the lines or arcs representing links among them in terms of their attendance at conferences. Two nodes are *adjacent* if they are connected by a line, and a node and a line are said to be *incident* with each other. The lines for these data are valued, but not directional. Figure 9e represents the relations among the conferences in terms of the number of professors who attended each pair of conferences.

In some cases a graphical form can be helpful for visualizing the network. Note that the spatial positioning of the nodes does not convey any meaning and is governed by what is convenient and visually clear.

Once the relationships have been represented in adjacency matrices, a wide variety of mathematical operations can be performed to describe properties of the network, such as how densely connected the nodes are, whether there are any subgroups or cliques, and whether there are individual nodes that essentially hold the network together. I will come back to some examples shortly.

The examples discussed so far illustrate the typical use of matrix representations in the study of social networks, that is to represent how people are interconnected. In terms of social remembering, such networks could be used to track how ideas and knowledge flow among people. For example, social network analysis has been used to study the diffusion of ideas through scientific communities (Doreian & Fararo, 1985; Mullins, Hargens, Hecht, & Kick, 1977). However, I have been exploring a different possible use of the technique, which is to study how ideas are organized across a group of individuals, and thereby provide an operational definition of collective memory.

In this conceptualization, the incidence matrix is used to represent the relation between individuals and conceptual units. Returning to Figure 9a, the rows still represent individuals but the columns now represent elements or idea units that comprise the knowledge that the individuals might share, such as an event, a story, or a set of beliefs. To collect the data, each individual provides his or her own account of the event, and each cell entry is either a 1 if the individual mentioned the idea in his or her account, or a zero if not. The adjacency matrix in Figure 9b then represents how the individuals are connected through the concepts or ideas they share (or more precisely, those they access and express in the context). The corresponding graph in Figure 9d could be thought of as a representation of the network in terms of shared beliefs and knowledge, and might identify distinct cliques or subgroups who have knowledge or attitudes in common with each other but different from that of other subgroups. With respect to the concept of collective memory, however, the adjacency matrix in Figure 9c is of particular interest because it represents how the concepts are organized or distributed among the group members, and as such can serve as an operational definition of the collective memory. This is a quite abstract way to represent shared knowledge, but it has the potential to capture structural properties that are not reflected in traditional measures like the number of items correctly recalled. These properties can serve as the dependent variables in research examining how the different experiences of different groups are reflected in their shared knowledge or collective mem-

ory. To illustrate some of these potential properties, I will first introduce them using the hypothetical matrices and graphs in Figure 9, and then provide an example using some of the data for recall of ''War of the Ghosts'' collected by Weldon and Bellinger (1997).

There are dozens of properties that can be calculated for relational matrices, but not all of these can be used with valued data, and not all have obvious psychological interpretations in terms of collective memory. Also, some seem more suitable for certain types of research designs and materials than others. I will illustrate just a couple of the potentially interesting properties that seem suitable for Weldon and Bellinger's data.

One important node-level property that can be derived from a relational matrix is *degree,* which refers to the number of nodes linked to (or *adjacent* to) a particular node of interest. This indicates the importance of that particular node in terms of the number or strength of the ties to it from other nodes in the network. In the case of a simple nondirectional unvalued graph, this is simply the number of lines incident to a node. If Figure 9d did not contain valued lines, then the degree of Professor 4 would be 1 and that of Professor 2 would be 3. With valued lines such as those actually present in Figure 9d, the degree of each node is the sum of the values of its incident lines, so the degree of Professor 4 is 1 and that of Professor 2 is 6. This indicates that Professor 2 has stronger ties to the other members of the network than does Professor 4. In terms of collective memory, the degree of an idea unit can reflect how important it is in the organization of the concepts. For example, in Figure 9e concepts B and C have the highest degrees and E the lowest. An idea unit with a high degree has strong connections to many other idea units across the network, and therefore may play a central role in organizing, accessing, and binding knowledge across the group.

At the level of the network, various indices are available to assess the overall cohesion or integration of the network. One statistical index is the *variance of the degrees* of the nodes across the network, which indicates whether the centrality of the nodes (the number and frequency of connections to other nodes) is relatively homogeneous across nodes (all the nodes have similar numbers of connections with other nodes) or relatively heterogenous (some nodes have high connectivity and others are more peripheral). Lower variance is associated with more even distribution of degrees across the nodes in the network. One can also look at the *mean degree* of all the nodes in the network to get an idea of the average connectedness of the nodes on the whole. However, the variance is more informative in terms of describing network-wide integration because it indexes the dispersion of the strength of the nodes across the network.

A commonly used measure of network cohesion is *density*. In an unvalued graph it simply refers to the number of possible lines that is actually present in the network. That is, it is the ratio of the number of lines present to the total that is possible if every node were linked with every other node. Higher density indicates higher interconnectedness and cohesion in the network. Unfortunately, for a valued graph density is a little trickier to compute and there is some debate about the best way to do this. Wasserman and Faust (1994) recommend calculating the numerator as the sum of the values of all lines and the denominator as the total possible number of lines, thereby reflecting the average strength of the lines in the valued graph. Higher values reflect higher average strengths of connection overall, that is, in the network as a whole. This has some obvious interpretive difficulties, however. For example, if only two or three nodes are strongly interconnected and have very high values on their lines, the same average density can be obtained as when many nodes are weakly connected. Thus, for valued graphs an index of the variance may be a more meaningful indicator of network-level cohesion. The variance would indicate whether the links tend to have relatively similar strengths, or whether some internode links are stronger than others. In any case, note that valued graphs can be converted to unvalued graphs by establishing a cut-off (e.g., links with a value ≥ 2 will be retained or assigned a 1, and <2 removed or assigned a 0), enabling one to employ a wider range of analytical tools.

For the network in Figure 9a, the mean degree is 6.0 and the variance of the degrees is 3.6. The average density is 1.5, and the standard deviation of density is .81. These numbers are not particularly meaningful in isolation, but are interesting in comparisons across networks. In terms of collective memory, indices of network cohesion can reflect how information is interconnected among individuals. For example, are the idea units unevenly interconnected among the individuals such that some elements of the knowledge set are more strongly interconnected than others? This question would be answered by looking at the variances of the degrees and density. Is the knowledge set as a whole weakly or strongly connected across the group members? This is addressed by measuring average degree.

The final property I would like to mention is the idea of a *clique*. In evaluating networks it is often desirable to identify cohesive subgroups, that is subgroups that are highly interconnected within themselves and simultaneously relatively weakly connected to other members or subgroups in the network. A clique is a specific type of subgroup in which all members are connected to all other members, and must contain at least three members. With a valued network, one can identify subgroups in which the ties are strongest. To do this, the values of the lines are used as cut-offs for inclusion in the clique. For example, one may want to identify subgroups

of professors who have all attended the same conferences three or more times during the past year. In Figure 9d, no such clique exists. Notice that Professors 1 and 3 have attended three conferences in common, and Professors 1 and 2 have as well, but because Professors 2 and 3 have attended only two conferences in common, the subgroup containing Professors 1, 2, and 3 does not meet the requirements for a clique. If the criterion is relaxed to two conferences, then Professors 1, 2, and 3 form a clique.

In terms of collective memory, the concept of cliques may be used to identify subsets of idea units that are strongly interconnected within the knowledge domain. For example in Figure 9e, there are two different ''concept cliques'' containing nodes that are interconnected through at least two people, ABC and BCD. These might represent core elements of the knowledge domain, shared experience, or belief system, that is, fundamental elements that are most widely held or strongly represented among the group members.

2. *Collective Memory of a Story: A Test of Some Network Analysis Concepts*

There are potentially many types of data one can gather to see whether the concepts of network analysis provide an interesting way to look at collective memory, and whether there is anything about memory that can be captured by looking beyond customary aggregate measures such as percent correct. As a first effort, I used the recalls of ''War of the Ghosts'' obtained by Weldon and Bellinger (Weldon & Bellinger, 1997). During the retrieval phase participants were required to recall the story twice. In one condition people first recalled the story individually and then again individually (Ind-Ind), whereas in another condition people first recalled the story collaboratively and then individually (Coll-Ind). The question of interest is whether the second recalls are different as a function of whether people previously worked in a group or alone. One might predict that working in a group would lead to greater similarity in how the individuals represent and organize the material; in other words, the collective memory of the material might be different for those who had recalled the material collaboratively as compared to those who worked alone. Note that for this analysis I am using only individual recalls, but looking at them as a network of knowledge represented across a small group of three people.

The data from the second recall were scored for each individual using the 42 idea units proposed by Mandler and Johnson (1977). The individual data for each of the three members of each group were then entered into a 3-by-42 incidence matrix with the rows representing participants and the columns representing the idea units. Thus, there were eight incidence

matrices for the Ind-Ind recall condition (one for each of the eight groups tested) and eight for the Coll-Ind condition. Note that this will not generate much power for the statistical comparisons because there are only eight observations in each between-subjects condition, but the data were suitable for this exploratory analysis. Each incidence matrix was decomposed into the column-by-column adjacency matrix (of the sort depicted in Figure 9c), yielding eight 42-by-42 valued nondirectional matrices for each condition, which I have suggested represent the collective memories of each group of three participants. These matrices were each separately analyzed to obtain measures of the mean degree, variance of degree, and the standard deviation of density. These three measures were chosen because they capture network-level properties, but node-level measures can be used when particular nodes are of interest in and of themselves. In addition, an analysis was performed to extract a clique of nodes that were interconnected with a strength equal to 3. Note that 3 is the maximum value any line can take in these matrices because there are only three people in each network. Therefore, in this case, the clique identifies the idea units that all three people recalled in common. The size of this concept clique was computed for each network. The value of each index for each network served as the data point for each group and was entered into a *t*-test comparing the Ind-Ind to the Coll-Ind condition.

The data are presented in Table V. The first column contains the nominal group recall from the original experiment (Weldon & Bellinger, 1997), that is, the pooled nonredundant recall of the three members of each group working independently. In this context, this can be thought of as an index of collective memory because it represents the total number of unique items ''known'' (or recallable in this context) by the three people in each group. These means are not significantly different, $t(14) = .59$, $p = .56$. Therefore, by this index the collective memories in the two conditions look

TABLE V

COLLECTIVE MEMORY FOR ''WAR OF THE GHOSTS''[a]

Recall condition	Nominal group recall	Network properties: Mean degree	Degree variance	Standard deviation of density	Concept-clique size
Individual-**Individual**	.77	50.55	594	.94	14.5
Collaborative-**Individual**	.74	63.06	797	1.06	20.9

[a] Data are based on recall in the second, individual recall. See text for definitions of the measures.

pretty similar, that is, the groups know about the same number of concepts whether they had previously recalled alone or together. The question is whether the network analyses reveal anything different.

The first measure is mean degree, which serves as one index of the connectivity of the network. Mean degree is the average connectedness of each idea unit to all the other idea units in the network. This value is higher in the Coll-Ind condition, indicating that on the average the idea units are more strongly interconnected after the group members have recalled the material together rather than individually. This difference approaches but does not exceed the conventional criterion for significance, $t(14) = 1.83$, $p = .09$, which does suggest that the measure may be able to tap into something unique about how the network of ideas is interconnected across the members of the group.

The second measure, the variance of the degrees, is one of the potentially most interesting properties of the network because it gives a sense of the dispersion of the connectedness of the idea units across the network, that is, whether the ideas are interconnected relatively homogeneously, or whether some ideas are more central than others. Collaboration in recall appears to have produced more heterogeneity in the network than did individual recall, as the variance of the degrees is higher in the Coll-Ind condition. This difference is statistically significant, $t(14) = 4.47$, $p < .01$. The most appropriate interpretation of this measure will require more work with it, but it might suggest that the process of collaboration leads to certain idea units emerging as central, popular, or organizing ideas for the knowledge network at large.

The third index is the standard deviation of the density. Recall that in unvalued graphs density is an index of network cohesion, measuring the proportion of total possible links that are present. For valued graphs it has been defined as the values of all the lines divided by the total possible, but this is difficult to interpret differently from mean degree. However, the standard deviation of this measure of density indicates whether the strength of connections between items (i.e., the values of the links) is relatively homogenous or heterogenous throughout the networks. (Note that this is related to the variance of the degree, but the degree focuses on the nodes, whereas density focuses on the lines.) Table V indicates that the standard deviation of the densities is higher in the Coll-Ind condition, $t(14) = 3.78$, $p < .01$. This may suggest that the knowledge networks that emerge through collaboration exhibit more heterogeneity in the strength of connections between idea units. That is, there are more pronounced differences between strong and weak connections that bind the ideas together across the members of the network. This may be due to an increasing number of strong

links developing during the process of collaboration, a hypothesis supported by the next result.

The final measure was concept clique size, which refers to the number of idea units that are all interconnected by a value of three, meaning that all three individuals recalled the entire set of idea units. This value is greater in the Coll-Ind condition, indicating that after collaboration the subset of units that formed the strongest interconnections among the group members is larger than among those who had worked alone. This difference is statistically significant, $t(14) = 3.024$, $p < .01$. This outcome suggests that the process of collaboration leads to more overlap in the individuals' accounts, and that the set of core idea units is larger in the collective memory of the collaborators.

In summary, I have explored the possibility that the methods of social network analysis can be used to represent collective memory as a network of ideas distributed among individuals. To do this I selected a few network properties that appeared to have psychological meaning when conceptualized in terms of collective memory and compared them in networks obtained under different conditions, one set of networks obtained after collaborative recall and the other after individual recall. The interesting outcome was that even though measures of nominal group recall revealed no difference in the level of collective recall in the two conditions, the network measures suggested that there were differences in how the material was organized within the group. Collaborative recall appeared to produce measurable differences in the collective memory. In particular, after collaborative recall the idea units appeared to be more interconnected and the material more cohesively organized. In addition, the collective memory appeared to contain more variability in the centrality of the idea units; that is, the individual idea units showed more variance in their degrees of connectedness to other idea units, perhaps suggesting that core or central concepts in the collective memory emerge more readily after collaboration. Finally, for the groups who had collaborated, the strongest connections in the network incorporated a larger subset of the idea units, suggesting that there is more information in the core knowledge set. Because these data are based on few observations, and have not yet been replicated, they should be viewed as preliminary hypotheses.

Although I have presented the network properties as dependent variables, it is also interesting to consider their potential as independent variables. That is, what are the behavioral implications for both individual and group behavior when collective memories have different properties? For example, do more cohesive networks lead to more group resistance to persuasion, resistance to updating with new information (i.e., more persistent proactive interference), and so forth? Do they lead to more rapid

decision making, more polarized attitudes, or more groupthink? Can groups with more cohesive collective memories assimilate new related information more effectively?

Collective memory has important social, cultural, and political implications, because it affects peoples' perceptions of individuals, groups, and events, and has important consequences for actions and reactions toward them. Shared memories have a significant impact on the construction and communication of events, group histories, the establishment of government and social policy, and the characterization of groups and institutions. Interesting examples occur regularly in current events. For example, during the war in the Persian Gulf, politicians and military analysts commonly referred to Saddam Hussein as "another Hitler," using the collective memory of the holocaust to win support for particular military strategies in the Persian Gulf. If these processes and their effects can be described in terms of measurable constructs, such as those available from network analytic methods, it might allow us to better relate collective memory to individual and group behavior.

V. Conclusions

Although cognitive psychologists have traditionally studied remembering as a set of processes in the individual mind, some social scientists in other fields of psychology and cognitive science have emphasized the social nature of memory. The goal of this chapter was to explore what it might mean for a cognitive psychologist to take a more social perspective on remembering. The first part of the chapter summarized some of the arguments that have been made as to why memory should be viewed as a social process. The second part then considered what the implications of this stance are for research on memory, focusing on the different kinds of conceptual and theoretical frameworks that can be adopted, as well as different methodological tools that can be employed. Finally, the third section presented two lines of empirical research I have undertaken in an effort to study the social nature of memory, one looking at the process of collaboration, and the other an effort to characterize collective memory.

Many cognitive psychologists believe that most of the value of the standard information-processing model has been exhausted, and the field needs to move in some new directions. One of the most important recent developments in this regard is neuroimaging, which has considerable value for addressing many interesting questions in cognitive psychology. I hope that the ideas presented in this chapter suggest further possibilities for new directions to take in the field, in particular, understanding the social nature

of cognition and memory and reintegrating the mind with the larger sociocultural context in which it evolves and functions.

Acknowledgments

I would like to think Emily Parodi, Jason Riggle, Brian Lathrop, and Qiang Liu, who participated in a fun and stimulating seminar on this topic and collaborated in working through many interesting ideas. I also thank Anthony Pratkanis for his enthusiasm and wealth of ideas and references, and David Smith for his pesky skepticism, unrelenting precision, and infinite support, and Henry L. Roediger III for helpful comments.

References

Academic American Encyclopedia. (1996). Danbury, CT.: Grolier Inc.

Andersson, J., & Rönnberg, J. (1995). Recall suffers from collaboration: Joint recall effects of friendship and task complexity. *Applied Cognitive Psychology, 9,* 199–211.

Andersson, J., & Rönnberg, J. (1997). Cued memory collaboration: Effects of friendship and type of retrieval cue. *European Journal of Cognitive Psychology, 9,* 273–287.

Axel, E. (1997). One developmental line in European Activity Theories. In M. Cole, Y. Engeström, & O. Vasquez (Eds.), *Mind, culture, and activity: Seminal papers from the Laboratory of Comparative Human Cognition* (pp. 128–146). Cambridge, UK: Cambridge University Press.

Baltes, P. B., & Staudinger, U. M. (Eds.). (1996). *Interactive minds: Life-span perspectives on the social foundation of cognition.* Cambridge, UK: Cambridge University Press.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge, UK: Cambridge University Press.

Basden, B. H., Basden, D. R., Bryner, S., & Thomas, R. L., III. (1997). A comparison of group and individual remembering: Does group participation disrupt retrieval? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 23,* 1176–1189.

Basden, D. R., & Basden, B. H. (1995). Some tests of the strategy disruption interpretation of part-list cuing inhibition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21,* 1656–1669.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (1999). *UCINET 5.0 Version 1.00.* Natick: Analytic Technologies.

Brent, E., & Thompson, G. A. (1999). Sociology: Modeling social interaction with autonomous agents. *Social Science Computer Review, 17,* 313–322.

Cartwright, D., & Harary, F. (1956). Structural balance: A generalization of Heider's theory. *Psychological Review, 63,* 277–293.

Clark, N. K., & Stephenson, G. M. (1989). Group remembering. In P. B. Paulus (Ed.), *Psychology of group influence,* 2nd ed (pp. 357–391). Hillsdale, NJ: Erlbaum.

Cole, M., Engeström, Y., & Vasquez, O. (Eds.). (1997). *Mind, culture, and activity: Seminal papers from the laboratory of comparative human cognition.* Cambridge, UK: Cambridge University Press.

Conway, M. A. (1997). The inventory of experience: Memory and identity. In J. W. Pennebaker, D. Paez, & B. Rimé (Eds.), *Collective memory of political events: Social psychological perspectives* (pp. 21–46). Mahwan, NJ: Lawrence Erlbaum Associates, Inc.

DeLoache, J. S., & Brown, A. L. (1997). Looking for Big Bird: Studies of memory in very young children. In M. Cole, Y. Engeström, & O. Vasquez (Eds.), *Mind, culture, and*

activity: Seminal papers from the Laboratory of Comparative Human Cognition (pp. 79–89). Cambridge, UK: Cambridge University Press.

Dixon, R. A., & Gould, O. N. (1996). Adults telling and retelling stories collaboratively. In E. P. B. Baltes, E. U. M. Staudinger, & et al. (Eds.), *Interactive minds: Life-span perspectives on the social foundation of cognition* (pp. 221–241). Cambridge, UK: Cambridge University Press.

Doreian, P., & Fararo, T. J. (1985). Structural equivalence in a journal network. *Journal of the American Society for Information Science, 36,* 28–37.

Durkheim, E. (1915). *Elementary forms of the religious life.* New York: Macmillan.

Durkheim, E. (1985). Primitive classification. In K. Thompson (Ed.), *Readings from Emile Durkheim* (pp. 136–138). London: Tavistock. [Original work published as *De quelques formes primitives de classification, Année Sociologique, 1901–2, 1903].*

Durkheim, E. (1985). The elementary forms of religious life. In K. Thompson (Ed.), *Readings from Emile Durkheim* (pp. 117–135). London: Tavistock. [From *The elementary forms of the religious life,* 1915. (J. W. Swain, Trans.) New York: Free Press].

Finlay, F., Hitch, G., J., & Meudell, P. R. (in press.) Mutual inhibition in collaborative recall: Evidence for a retrieval based account. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Fiske, S. T., & Goodwin, S. A. (1994). Social cognition research and small group research, a West Side Story or . . . ? *Small Group Research, 25(2),* 147–171.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition (2nd ed.):* Mcgraw-Hill Book Company.

Gardner, H. (1985). *The mind's new science.* New York: Basic Books.

Greeno, J. G., Moore, J. L., & Smith, D. R. (1993). Transfer of situated learning. In E. Douglas K. Detterman, E. Robert J. Sternberg, & et al. (Eds.), *Transfer on trial: Intelligence, cognition, and instruction* (pp vi, 296): Ablex Publishing Corp.

Halbwachs, M. (1980). *The collective memory.* (*La mémoire collective.*) (F. J. Ditter, Jr., & V. Y. Ditter, Trans.). New York: Harper Row. [Original work published 1950]

Hastie, R., & Kumar, P. A. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality & Social Psychology, 37,* 25–38.

Heider, F. (1946). Attitudes and cognitive organization. *Journal of Psychology, 21,* 107–112.

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processes. *Psychological Bulletin, 121*(1), 43–64.

Hutchins, E. (1994). *Cognition in the wild.* Cambridge, MA: MIT Press.

Ickes, W., & Gonzalez, R. (1996). ''Social'' cognition and social cognition: From the subjective to the intersubjective. In J. L. Nye & A. M. Brower (Eds.), *What's social about social cognition? Research on socially shared cognition in small groups* (pp. 311–323). Thousand Oaks, CA: SAGE.

Janis, I. L., & Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice, and commitment.* New York: Free Press.

Jost, J. T., Kruglanski, A. W., & Nelson, T. O. (1998). Social metacognition: An expansionist review. *Personality & Social Psychology Review, 2,* 137–154.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65,* 681–706.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science, 18,* 513–549.

Kolers, P. A., & Roediger, H. L. (1984). Procedures of mind. *Journal of Verbal Learning & Verbal Behavior, 23,* 425–449.

Lancy, D. F., & Strathern, A. J. (1981). Making twos: Pairing as an alternative to the taxonomic mode of representation. *American Anthropologist, 83,* 773–795.

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology, 37,* 822–832.

Lave, J. (1988). *Cognition in practice: Mind, mathematics, and culture in everday life.* Cambridge, UK: Cambridge University Press.

Lave, J. (1997). What's special about experiments as contexts for thinking. In M. Cole, Y. Engeström, & O. Vasquez (Eds.), *Mind, culture, and activity: Seminal papers from the Laboratory of Comparative Human Cognition* (pp. 57–69). Cambridge, UK: Cambridge University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge, UK: Cambridge University Press.

Loftus, E. F. (1991). Made in memory: Distortions in recollection after misleading information. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 27, pp. 187–215). San Diego, CA: Academic Press, Inc.

Loftus, E. F. (2000). Suggestion, imagination, and the transformation of reality. In A. A. Stone, J. S. Turkkan, & et al. (Eds.), *The science of self-report: Implications for research and practice* (pp. 201–210). Hillsdale, NJ: Lawrence Erlbaum.

Lorge, I., & Solomon, H. (1961). Group and individual behavior in free recall. In H. S. P. S. J. H. Criswell (Ed.), *Mathematical methods in small group processes* (pp. 221–231). Stanford, CA: Stanford University Press.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology, 9,* 111–151.

Mäntylä, T., & Nilsson, L.-G. (1983). Are my cues better than your cues? Uniqueness and reconstruction as prerequisites for optimal recall of verbal materials. *Scandinavian Journal of Psychology, 24,* 303–312.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: W. H. Freeman.

McCauley, C. (1989). The nature of social influence in groupthink: Compliance and internalization. *Journal of Personality & Social Psychology, 57(2),* 250–260.

Meudell, P. R., Hitch, G. J., & Boyle, M. M. (1995). Collaboration in recall: Do pairs of people cross-cue each other to produce new memories? *Quarterly Journal of Experimental Psychology, 48A,* 141–152.

Meudell, P. R., Hitch, G. J., & Kirby, P. (1992). Are two heads better than one? Experimental investigations of the social facilitation of memory. *Applied Cognitive Psychology, 6,* 525–543.

Minick, N. (1997). The early history of the Vygotskian school: The relationship between mind and activity. In M. Cole, Y. Engeström, & O. Vasquez (Eds.), *Mind, culture, and activity: Seminal papers from the Laboratory of Comparative Human Cognition* (pp. 117–127). Cambridge, UK: Cambridge University Press.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior, 16,* 519–533.

Moscovici, S., & Duveen, G. (1998). The history and actuality of social representations. In U. Flick et al. (Eds.), *The psychology of the social* (pp. 209–247). New York: Cambridge University Press.

Mullins, N. C., Hargens, L. L., Hecht, P. K., & Kick, E. L. (1977). The group structure of cocitation clusters: A comparative study. *American Sociological Review, 42,* 552–562.

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning & Verbal Behavior, 19,* 338–368.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (in press). Culture and systems of thought: Holistic vs. analytic cognition. *Psychological Review.*

Osterreith, P. A. (1944). Le test de copie d'une figure complèxe. *Archives de Psychologie, 30,* 206–356.

Pratkanis, A. R., & Turner, M. E. (1999). Groupthink and preparedness for the Loma Prieta earthquake: A social identity maintenance analysis of causes and preventions. *Research on Managing Groups and Teams, 2,* 115–136.

Prokhorov, A. M. E. (1983). *Great Soviet encyclopedia. Translation of Bolshaia sovetskaia entsiklopediia.* New York: Macmillan.

Resnick, L. B., Levine, J. M., & Teasley, S. D. (Eds.). (1991). *Perspectives on socially shared cognition.* Washington, DC: American Psychological Association.

Roediger, H. L., III, & Neely, J. H. (1982). Retrieval blocks in episodic and semantic memory. *Canadian Journal of Psychology, 36,* 213–242.

Rogoff, B. (1998). Cognition as a collaborative process. In W. Damon, D. Kuhn, & R. S. Siegler (Eds.), *Handbook of child psychology: Vol. 2. Cognition, perception, and language* (pp. 679–744). New York: Wiley.

Rogoff, B., & Mistry, J. J. (1985). Memory development in cultural context. In M. P. C. Brainerd (Ed.), *Cognitive learning and memory in children* (pp. 117–142). New York: Springer-Verlag.

Sack, K. (1999, May 5). Fury on the Plains: The scene. *The New York Times.*

Scott, J. (1991). *Social network analysis: A handbook.* Newbury Park, CA: SAGE.

Siegal, M. (1991). A clash of conversational worlds: Interpreting cognitive development through communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 23–40). Washington, DC: American Psychological Association.

Slamecka, N. J. (1969). Testing for associative storage in multitrial free recall. *Journal of Experimental Psychology, 81,* 557–560.

Srull, T. K. (1981). Person memory: Some tests of associative storage and retrieval models. *Journal of Experimental Psychology: Human Learning & Memory, 7,* 440–463.

Stasser, G., & Titus, W. (1987). Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *Journal of Personality and Social Psychology, 53,* 81–93.

Taylor, D. W., Berry, P. C., & Block, C. H. (1958). Does group participation when using brainstorming facilitate or inhibit creative thinking? *Administrative Science Quarterly, 3,* 23–47.

Tobach, E., Falmagne, R. J., Parlee, M. B., Martin, L. M. W., & Kapelman, A. S. (Eds.). (1997). *Mind and social practice: Selected writings of Sylvia Scribner.* Cambridge, UK: Cambridge University Press.

Tulving, E. (1985). How many memory systems are there? *American Psychologist, 40,* 385–398.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 359–380.

Vygotsky, L. S. (1962). *Thought and language.* Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1978). *Mind in society.* Cambridge, MA: Harvard University Press.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications.* New York: Cambridge University Press.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. R. Goethals (Eds.), *Theories of group behavior* (pp. 185–208). New York: Springer-Verlag.

Wegner, D. M. (1995). A computer network model of human transactive memory. *Social Cognition, 13,* 318–339.

Wegner, D. M., Erber, R., & Raymond, P. (1991). Transactive memory in close relationships. *Journal of Personality & Social Psychology, 61,* 923–929.

Weldon, M. S., & Bellinger, K. D. (1997). Collective memory: Collaborative and individual processes in remembering. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 23,* 1160–1175.

Weldon, M. S., Blair, C., & Huebsch, P. (in press). Group remembering: Does social loafing underline collaborative inhibition? *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Wertsch, J. V. (1991). A sociocultural approach to socially shared cognition. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 85–100). Washington, DC: American Psychological Association.

NEUROCOGNITIVE FOUNDATIONS OF HUMAN MEMORY

Ken A. Paller

Memory research has traditionally been segregated into (a) research on the cognitive organization of memory and (b) research on the brain basis of memory. As a result of this segregation, cognitive theories of memory have generally evolved in isolation from neuroscientific theories of memory, and vice versa. Moreover, cognitive theorists have regularly found that neural evidence is not relevant to their theorizing. Neuroscientists, on the other hand, are centrally concerned with understanding how the brain works, and their work has usually not made contact with efforts to understand how the mind works.

Nonetheless, new connections between cognitive science and brain science are evident in a wide variety of domains. Advances in measuring human brain function, for example, have increased the likelihood that relevant information about how the brain works can be used to discover how the mind works. In memory research, there is now a growing trend to use neural information to inform theories of memory (Gabrieli, 1998; Schacter, Norman, & Koutstaal, 1998). This approach requires an appreciation of both the brain substrates of memory and the cognitive organization of memory, and seeks to develop bridges between theories couched at these different levels.

Here, I will make the case that measures of the electrical activity of the human brain, in particular, can be used as powerful tools for studying human memory. Associations between memory functions and electrophysiological

THE PSYCHOLOGY OF LEARNING AND MOTIVATION, VOL. 40

Copyright © 2001 by Academic Press.
All rights of reproduction in any form reserved.

0079-7421/00 $35.00

measures will be interpreted not merely within the context of studies of brain potentials, but within a broader theoretical perspective that addresses both the cognitive structure (as in a characterization of relevant cognitive processes and their interrelationships) and the neural implementation of memory functions. Therefore, I will begin with a discussion of the general problem of bridging neural and cognitive levels. I will then outline a framework for thinking about the neurocognitive foundations of certain types of memory. Finally, I will explore the prospects for future advances based on measuring the electrical activity of the human brain.

I. What Evidence from Neuroscience Is Relevant for Understanding Cognition?

Contact between cognitive and neuroscientific approaches is important if one is concerned with the goal of improving our understanding of the cognitive functions of the human brain. To achieve this goal, it would be helpful to determine which facts about the brain might be relevant. It is disappointing, then, that there is no widely agreed upon set of criteria for making this determination.

In addressing this issue, it is instructive to first consider an alternative goal. At one extreme, some prefer to assume that no brain facts are relevant. Indeed, a great many cognitive scientists have developed functional models of cognition that make no reference to the brain. Three brief arguments against this approach are as follows. First, unless one holds the view that the mind emanates from something other than the brain, one must accept the view that cognitive functions are inherently functions of the brain. It follows that ignoring the neurophysiological substrates of cognitive functions is detrimental to gaining a comprehensive understanding of cognition. Second, scientific theories should be tested using the most extensive basis of empirical evidence available. Indeed, cognition is not measured directly—it is fundamentally an inference based on behavioral observations. Cognitive theories can thus be improved by including a basis of both behavioral and neural observations. Third, whereas behavioral evidence is obtained only after complex interactions among multiple cognitive processes lead to a response, neural evidence is not subject to this limitation. Online measures of brain activity offer the remarkable possibility of tapping into relevant processes as they occur, perhaps facilitating the individuation of component cognitive processes. In short, the understanding that cognition arises from the brain leads naturally to the position that models of cognition should be based on relevant evidence of both behavioral and neural varieties.

Another extreme position that should also be considered is that *all* neuroscientific information is relevant to cognitive theory. This position is based on the premise that we do not know in advance how cognitive functions are related to the neurophysiology of the brain. Therefore, we cannot know in advance precisely which evidence will or will not ultimately be relevant. Indeed, neuroscientific facts at the level of detail of the molecular structure and genetic coding for various neuroreceptors may benefit a future understanding of some aspects of cognition. In the absence of a fully developed theory of cognition, perhaps all such neural information should be made available to cognitive scientists. Nevertheless, this view is problematic because it conflates two concerns: how future conceptions of the mind will unfold and how the development of cognitive theories can proceed at present.

How can we evaluate the relevance of any given neural evidence for developing our *current* understanding of cognition? Clearly, approaching this question with respect to a specific case, as attempted below, is more feasible than providing a universal answer. How the evaluation process will play out depends on the extent to which a hypothetical description of the neural implementation of the cognitive function in question can already be given. If little is known about the neural implementation of a cognitive function, it is difficult to know where to begin forging brain–cognition connections. This is not the case in memory research. The extremely wide range of neuroscience research on plasticity and learning (e.g., Fuster, 1995; Martinez & Kesner, 1998; Milner, Squire, & Kandel, 1998; Squire & Kandel, 1999) provides a rich set of puzzle pieces that will eventually be assembled in new ways as future theories of memory are developed. One way to consider the co-evolution of neural and cognitive theories of memory is with respect to the levels of computation, algorithm, and implementation as distinguished by Marr (1982). Explanations restricted to one level can be important (Dror & Gallogly, 1999). On the other hand, a complete understanding of memory will require explanation at all levels, along with systematic mappings between levels. No matter how the levels are conceptualized, or how many are distinguished, the levels are inherently interdependent, at least to some extent. An accurate conception of neural plasticity at cellular and molecular levels must ultimately be an important part of a comprehensive understanding of human memory. Yet, it is difficult to bring most of this evidence to bear on cognitive theories of memory at present.

Moreover, quite different criteria apply towards achieving the goal of understanding neural substrates versus the goal of understanding cognitive structure. For example, an empirical connection between a brain region and a specific memory function may constitute a step forward in understanding neural substrates, but may have no impact whatsoever on understanding

cognitive structure. Suppose that two putatively distinct memory functions were implemented in two different brain regions. This finding could be a trivial one for cognitive theorizing if the computations and representations in the two regions were identical, merely situated in two different physical locations for arbitrary reasons. On the other hand, an anatomical separation between two memory functions *may* correspond to a fundamental psychological distinction between them, as described in the next section. In general, empirical brain–behavior associations or dissociations are most useful for understanding cognitive structure when a fairly developed theoretical framework for the evidence is already available. It can be especially useful to relate the proposed function of a network of neurons to compatible evidence about other cognitive roles played by the same network. Furthermore, when a brain region plays a role in multiple cognitive functions, a more complete understanding of the interrelationships between diverse cognitive functions can be sought.

In sum, cognitive theories can be reinforced and extended by combining them with empirically supported characterizations of the neural implementation of relevant processes. This statement is compatible with a rationale for combining cognitive and neuroscientific approaches enunciated by philosopher Owen Flanagan (1992). First, he postulated that behavioral facts alone will not provide sufficient empirical constraints on cognitive theories, and that ''psychological explanations need to be constrained by knowledge about the brain'' (Flanagan, 1992, p. 12). But he also suggested that neuroscience cannot generate progress in understanding the mind without adequate emphasis on the psychological phenomena themselves. The goal of understanding cognition is too lofty for either cognitive science or neuroscience alone; it requires both. Sound cognitive theorizing is required if attempts to understand the neural implementation of cognitive functions are to be successful. And a sound understanding of neural implementations is required for cognitive theorizing to be successful in the long term. Therefore, neural and cognitive advances can best be pursued together.

II. The Neural Implementation of Declarative Memory

Mental disorders due to brain damage often provide striking demonstrations of the interrelations between mind and brain. Neuropsychological studies of patients with amnesia have been especially useful for charting connections between memory and the brain. This usefulness is a fortuitous by-product of the fact that the amnesic deficit can be highly selective. An amnesic deficit is selective when it occurs together with a vast array of preserved cognitive functions, including some memory functions. Accord-

ingly, evidence from amnesia has been widely used to develop hypotheses about the fundamental structure of memory. In particular, *declarative memory* has been defined behaviorally as the type of memory required for recalling and recognizing facts and events and for experiencing conscious recollection under such circumstances (Squire, 1987). Ample neuropsychological evidence supports the classification of declarative memory as distinct from other types of memory. Nevertheless, the precise nature of declarative memory remains to be elucidated.

The key empirical support for distinguishing between declarative and nondeclarative memory consists of a set of dissociations, wherein amnesic patients demonstrate poor memory when tested by recall or recognition but not when tested using various *implicit memory tests,* which are memory tests that make no reference to prior learning episodes. These two classes of memory phenomena can thus be distinguished:

1. *Conscious recollection*—when one brings to mind some prior event or some factual knowledge, with the awareness of retrieving a memory.
2. *Perceptual priming*—when behavior is changed in certain circumstances pertaining to a specific perceptual event, as the result of prior experience, and with no necessary experience of recollection.

One example of an implicit memory test used to show preserved perceptual priming in amnesia is the word-identification test (Cermak, Talbot, Chandler, & Wolbarst, 1985; Haist, Musen, & Squire, 1991; Hamann, Squire, & Schacter, 1995; Paller, Mayes, McDermott, Pickering, & Meudell, 1991). In this test, subjects attempt to read words presented in a degraded manner. Perceptual priming can be observed when identification is superior for words that also appeared in a study phase preceding the word-identification test. Although amnesic patients generally show normal priming in the word-identification test, they perform poorly when asked to recognize whether those words were presented earlier (Figure 1). Preserved priming in amnesia has also been verified with a wide variety of other implicit memory tests in a large literature on the topic (Moscovitch, Vriezen, & Goshen-Gottstein, 1993; Schacter, Chiu, & Ochsner, 1993; Shimamura, 1986, 1993). Although such dissociations can arise for artifactual reasons, alternative explanations based on these artifacts cannot explain the bulk of the evidence of preserved priming in amnesia (Hamann et al., 1995; Hamann & Squire, 1997; Squire, Hamann, & Schacter, 1996). Other types of memory shown to be preserved in amnesia include motor skills, cognitive skills, simple classical conditioning, habits, artificial grammar learning, category learning, nonassociative learning, and working memory (Gabrieli, 1998; Schacter & Tulving, 1994; Squire, 1992; Squire & Knowlton, 2000; Squire & Paller, 2000).

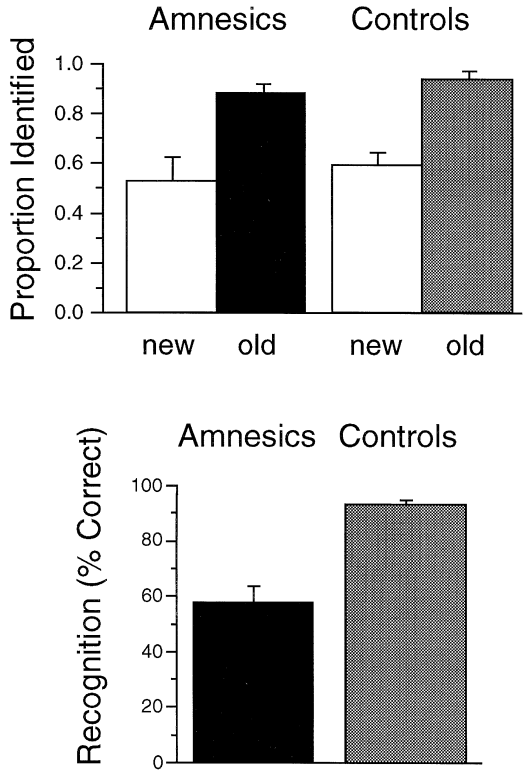


Fig. 1. Memory performance in amnesic patients and age-matched control subjects on implicit and explicit memory tests. Patients with amnesia exhibited normal word-identification priming (top) and impaired three-alternative recognition for the same words (bottom). Figure adapted from Paller et al. (1991).

These are remarkable findings because they support the notion that memory should be conceived of as encompassing a set of functions, not just one ability. Furthermore, the fact that such distinctions are honored by neuroanatomical boundaries bodes well for further neurocognitive exploration. Imagine instead an alternative state of affairs such that declarative and nondeclarative memory depend on exactly the same neuroanatomical regions at some gross level of analysis. There might be no simple way in which declarative memory could be disrupted without also disrupting nondeclarative memory. We could still be dealing with a fundamental distinction, but if the relevant memory functions were implemented in intermingled neural tissue it would be more difficult to clarify the neural basis of the distinction. The discovery that declarative and nondeclarative mem-

ory are anatomically distinct opens the door for further investigations that can both clarify the neural bases of the distinction and elucidate the cognitive processes specifically associated with each type of memory. We can thus envision ways to pursue memory theories that address both cognitive structure and neural implementation.

According to theoretical formulations put forward by countless memory theorists (e.g., see Mayes & Downes, 1997), the neural dysfunction in amnesic patients disrupts declarative memory but leaves other types of memory entirely intact. Exactly why declarative memory has this status remains an issue of continued investigation. Nonetheless, a key generalization is that when memory breaks down in amnesia, recollection is partially disrupted but certain types of priming are preserved. By juxtaposing recollection and priming, we can learn about how they differ from each other, in both their psychological and neural facets, and thus come to a better understanding of each on its own. For example, brain potentials that follow the time course of these two memory phenomena may provide a way to monitor relevant cognitive processes and study their neural substrates. Research of this sort may ultimately support attempts to describe the enigmatic border between conscious and unconscious mental events in neural terms.

A fundamental speculation about declarative memory is that the requisite information storage takes place within neocortical areas dedicated for processing the particular type of information in question (Squire & Paller, 2000). Memories are not all stored in a unitary memory storehouse in the brain. Instead, facial memories are stored in cortical areas where facial information is represented, verbal memories where verbal information is represented, and so on. Memory storage in the cortex generally follows functional specialization in the cortex. Yet, the memory dysfunction of amnesia cuts across all sensory modalities, while at the same time it does not disrupt perceptual abilities. The amnesic impairment is global in that it encompasses memories based on all sorts of information, but it is also focal in that it is restricted to declarative memory. This pattern of memory breakdown provides important clues for understanding the neural and cognitive nature of memory. In fact, there is a substantial consensus among scientists studying memory disorders that the central problem is generally in storing declarative memories rather than in encoding or retrieving them per se.

I will now present one specific conceptualization of this memory storage problem, based on some ideas I proposed earlier (Paller, 1997). First, amnesia can result from a defect in a special sort of *consolidation* process whereby enduring declarative memories are stored in the cerebral cortex. This consolidation process is required because the elements of a declarative memory

are not easily held together, due to their anatomical isolation. Consolidation entails interactions between the cerebral cortex and two key brain areas, the medial temporal region and the medial diencephalon (Figure 2). The brain dysfunction in amnesia most commonly lies in one or the other of these two key areas, as structural neuroimaging, functional neuroimaging, and post mortem histology have shown.

Declarative memories characteristically depend on multiple *neuronal ensembles* (Hebb, 1949) that represent different high-level perceptual, cognitive, and emotional attributes processed in functionally distinct cortical regions. The experience of a fact or event in the present moment, as in immediate memory or working memory, can be achieved when one such set of distributed neuronal ensembles is activated under the control of prefrontal networks. Furthermore, this set of neuronal ensembles can become temporarily connected via cortico-thalamic and cortico-hippocampal networks. This temporary linking function is, in some cases, ultimately

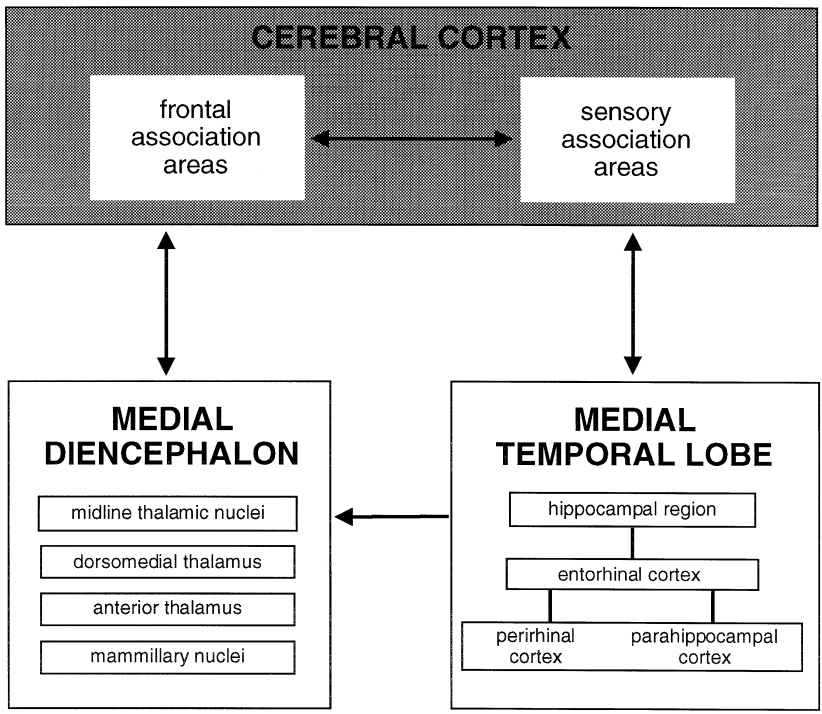


Fig. 2. Neuroanatomical regions involved in the storage of declarative memories. Figure adapted from Squire and Paller (2000).

replaced by new neocortical representations, instantiated by "coherence ensembles" (Paller, 1997), which are located in temporal lobe regions near the hippocampus[1]. The central function of these newly formed neuronal ensembles is to provide coherence to the dispersed neocortical representation. Enduring declarative memories are, by this account, characteristically composed of a set of distributed neocortical ensembles plus associated coherence ensembles (Figure 3).

Consolidation fundamentally entails the repeated activation of this set of neocortical storage sites, thereby mediating memory retrieval, associations with other memories, and the formation of an enduring declarative memory. These several events that comprise consolidation can proceed whether or not the individual is intending to memorize or rehearse the memory, such as during the experience of related events or during sleep. Other aspects of this theoretical formulation have been described in detail elsewhere (Paller, 1997); for present purposes a central implication is that declarative memory and priming can now be given these tentative neurobiological definitions:

1. Declarative memory is a type of neocortical memory in which the relevant plasticity occurs *across* many neocortical zones, and storage requires a special consolidation process that is unique to this type of dispersed neocortical memory.
2. Priming is a type of neocortical memory in which the relevant plasticity occurs *within* a single neocortical zone.

Although priming has most commonly been investigated using single items such as words, sometimes pairs of items have been used instead. In such cases, priming may rely on associations among separate representations. Such new associations might depend on plasticity across multiple neocortical zones rather than merely within-zone plasticity. According to the conceptualization outlined above, this would require consolidation. Indeed, implicit memory tests that tax priming for new associations often reveal impairments in amnesic patients (Cermak, Bleich, & Blackford, 1988; Chun & Phelps, 1999; Mayes & Gooding, 1989; Paller & Mayes, 1994; Schacter & Graf, 1986; Shimamura & Squire, 1989). Exceptions to this generalization (Gabrieli, Keane, Zarella, & Poldrack, 1997; Musen & Squire, 1993) may reflect the use of unitized representations such that

[1] The most severe cases of amnesia are caused by damage to these temporal lobe regions plus the hippocampus. In contrast, damage to these temporal lobe regions alone can lead to semantic dementia, in which remote memories may be disrupted more than recent memories (Hodges & Graham, 1998). Speculatively, damage to a circumscribed portion of these regions with relative sparing of the hippocampus and some adjacent temporal cortex may cause a focal retrograde amnesia (Kapur, 1993; Markowitsch, 1995), such that the patient exhibits excessive retrograde impairments along with a preserved ability to form new declarative memories.

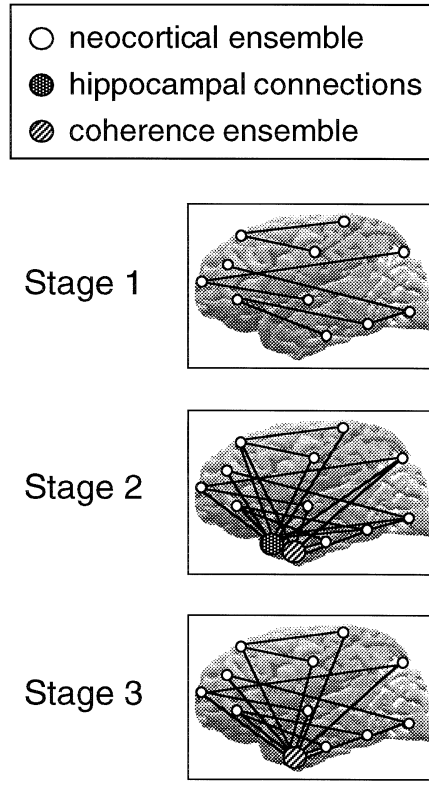


Fig. 3. Figurative depiction of the stages of consolidation of dispersed neocortical memories. In stage 1, a fact or event is encoded via representations that involve multiple cortical regions, based on working memory processes controlled by prefrontal cortex. At about the same time, in stage 2, this dispersed representation begins to make contact with neurons in the medial temporal region. Hippocampal connections rapidly become part of a newly formed network so that the dispersed cortical fragments can remain connected beyond the span of immediate memory. In addition, coherence ensembles are formed in entorhinal, perirhinal, and parahippocampal cortex, and possibly in adjacent cortical regions, and these become part of the network. Connections become strengthened as the memory is reactivated on subsequent occasions, while coherence ensembles develop a central role as they take on gestalt-like aspects of the memory and its relationship to other memories. Coherence ensembles thus function to maintain cohesiveness among the various parts of the declarative memory. In stage 3, coherence ensembles can take part in the reactivation of the dispersed neocortical memory without any necessary contribution from hippocampal connections.

plasticity within single neocortical zones is operative. In short, implicit memory tests do not always show normal priming in amnesia. For normal priming to be observed in amnesia, performance in healthy individuals

must not be mediated by recall or recognition, and performance must not depend on new associations represented across distinct cortical zones.

The dissociation between recollection and priming in amnesia can now be recast in neurobiological terms—recollection is impaired due to consolidation failure following disruption of either cortico-thalamic or cortico-hippocampal networks, whereas priming due to plasticity within single neocortical zones is preserved. In attempting to tie together psychological and biological facets of amnesia, this conceptualization portrays the beginnings of a neurocognitive theory of declarative memory and priming in healthy individuals. Given the rapid growth of cognitive neuroscience in recent years, there is reason to be optimistic that this theory can be developed further, specified in increasingly precise detail, and put to empirical test. But a critical step will be to identify measures of the relevant memory functions in human subjects.

III. Electrophysiological Measures of Recollection

Given these hypotheses about memory derived from neuropsychological research, I will now explore how relevant evidence can be obtained by measuring brain activity in healthy individuals. In particular, I will focus on measures of brain activity called event-related potentials (ERPs), which are extracted from the electroencephalogram (EEG) using signal-averaging methods. Many aspects of this treatment also apply to work with other methodologies that provide evidence about brain activity during cognitive functions, including positron emission tomography (PET), functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), and transcranial magnetic stimulation (TMS). Noninvasive ERP recordings from the scalp reflect the summation of electrical activity generated in various brain regions, activity which summates by virtue of factors such as the spatial alignment of neurons and the synchrony of neural activity (Kutas & Dale, 1997). The present discussion of ERPs and memory has a limited scope (for more extensive reviews of the literature, see Johnson, 1995; Kutas, 1988; Rugg, 1995).

ERPs arguably provide measures of memory processes that can aid the development and testing of theories of human memory in several ways. ERPs have been recorded during various sorts of recognition tests, and results suggest that differential ERP responses to old and new items may be useful for studying retrieval (e.g., Friedman, 1990; Johnson, Pfefferbaum, & Kopell, 1985; Karis, Fabiani, & Donchin, 1984; Neville, Kutas, Chesney, & Schmidt, 1986; Paller, Kutas, & Mayes, 1987; Rubin & McAdam, 1972; Rugg & Nagy, 1989; Sanquist, Rohrbaugh, Syndulko, & Lindsley, 1980;

Warren, 1980). Although not emphasized here, other electroencephalographic phenomena have also been linked to memory functions, including slow potentials (Rösler, Heil, & Roder, 1997) and measures in the frequency domain (Klimesch, 1999).

In ERP experiments with visual words, responses to old words have tended to be more positive than responses to new words about 400 to 800 ms after word onset (Figure 4). ERP studies of recognition have also been conducted with nonverbal stimuli such as simple object drawings (e.g., Friedman & Sutton, 1987) and faces (e.g., Barrett, Rugg, & Perrett, 1988; Smith & Halgren, 1987). In one experiment, ERPs were recorded from subjects while they viewed color slides depicting a variety of people, places, and paintings (Neville, Snyder, Woods, & Galambos, 1982). Following ERP recordings, subjects were shown each slide a second time and asked whether they had recognized it during its initial exposure. An average of 11% of the slides fell into this category, and ERPs to these recognized slides were considerably more positive than ERPs to unrecognized slides, particularly at around 400 ms. This result parallels findings from various other experi-

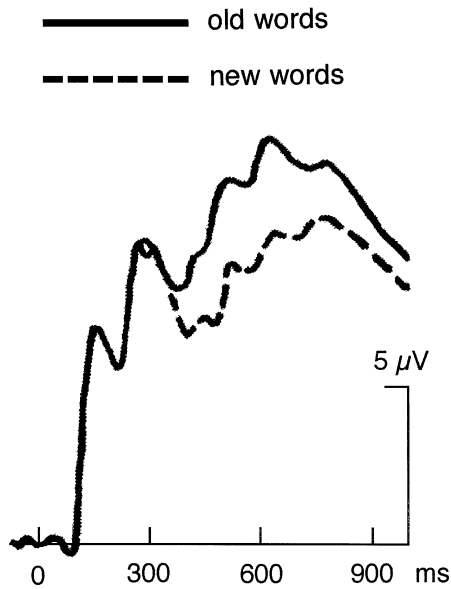


Fig. 4. An example of an old–new ERP difference (also known as an ERP repetition effect) obtained during a yes–no recognition test. The response to old words was more positive than the response to new words, beginning about 400 ms after word onset (from results obtained by Paller et al., 1987). Recordings were made from the midline parietal scalp location, and positive potentials are shown as upward deflections.

ments in which items recognized from a presentation earlier in the experiment elicited ERPs that were more positive than ERPs to new items. These reports of *old-new ERP differences* provide a good foundation for the suggestion that recognition processes are measurable using ERPs.[2]

Before accepting the claim that ERPs tap processes central to recognition, however, we must consider several features of typical recognition paradigms that cloud the issue. First, both the time to make a recognition judgment and the confidence with which it is made may differ considerably between old and new items.[3] Second, the fact that subjects are required to detect recognized items may call into play target-detection operations that differ for old and new items, in that only the former are targets. Third, the subjective probability of old and new items may differ, whether or not actual stimulus probabilities are matched. All of these factors are known to influence ERPs, so these confounding factors are particularly worrisome. In short, it can be problematic to distinguish electrophysiological (or other physiological) effects due to these nonspecific factors from effects due to recognition per se. Likewise, multiple sorts of retrieval-related processes can be difficult to investigate when limited to old–new comparisons (e.g., Rugg & Wilding, 2000). Furthermore, and most important in the present context, simply comparing old and new items in a recognition paradigm does not take into account the idea that both recollection and priming occur in such a situation.

Although several lines of evidence are consistent with the hypothesis that ERPs are sensitive to recognition processes, much of the early evidence fell short of being conclusive. Johnson and colleagues (Johnson et al., 1985; Johnson, Kreiter, Russo, & Zhu, 1998) showed that ERPs correlated with

[2] ERPs have also been used to show that neural events at the time information is first encoded are predictive of whether they will be remembered later. The phenomenon of an ERP *Difference* computed as a function of later *memory* performance, sometimes called *Dm,* has been observed with explicit memory tests such as recall and recognition (Fabiani, Karis, & Donchin, 1986; Karis et al., 1984; Paller, 1990; Paller et al., 1987; Paller, McCarthy, & Wood, 1988; Sanquist et al., 1980) but not with implicit memory tests such as stem completion or word identification (Paller, 1990; Paller & Kutas, 1992). More recently, such effects have been observed with fMRI (Brewer, Zhao, Desmond, Glover, & Gabrieli, 1998; Fernández et al., 1998; Wagner et al., 1998), with event-related EEG synchronization (Klimesch, 1999), with EEG coherence measures (Weiss & Rappelsberger, 2000), and with intracranial ERP recordings in the medial temporal region (Fernández et al., 1999).

[3] In one experiment, for example, ERPs were averaged separately according to confidence measures (an analysis reported by Paller, 1993, of data collected by Paller, Kutas, & Mayes, 1987). ERPs were more positive for words categorized with high confidence than with low confidence. In fact, ERPs in other paradigms are similarly correlated with decision confidence (Hillyard, Squires, Bauer, & Lindsay, 1971; Ruchkin & Sutton, 1978). Confounds with recognition confidence can thus contribute to observed old–new ERP differences, particularly when hits to old items are more confident than correct rejections to new items. In addition, confidence judgments may be insufficiently sensitive to ameliorate this problem.

increases in memory strength as study words were repeated, which would be expected for an ERP correlate of recognition strength. Bentin and colleagues (Bentin & Moscovitch, 1990; Bentin, Moscovitch, & Heth, 1992) recorded old–new ERP differences and proposed that ERPs are sensitive to both implicit and explicit aspects of memory performance. However, the methods used were insufficient for disentangling these different processes. Smith and Halgren (1989) interpreted their ERP results in terms of a distinction between two separate bases for recognition judgments, familiarity based on memory strength and contextual retrieval (Atkinson & Juola, 1973; Jacoby & Dallas, 1981; Mandler, 1980). Lists of 20 abstract words were arranged such that the same 10 words occurred in every list, and subjects were instructed to press a button whenever one of these repeating words was presented. Recognition accuracy increased across the six lists in healthy individuals and in patients with unilateral anterior temporal lobectomies (performed for relief of medically intractable epilepsy). Smith and Halgren (1989) suggested that increases in memory strength were normal in all patients, but that the ability of patients with left-hemisphere excisions to use contextual retrieval was compromised and that this explained attenuated old–new ERP differences that were found in those individuals. In other words, old–new ERP differences were thought to reflect contextual retrieval. This conclusion was not endorsed by the authors of a similar study of epileptic patients (Rugg, Roberts, Potter, Pickles, & Nagy, 1991), who found no relationship between the magnitude of old–new ERP effects and verbal memory performance.

An alternative position was advocated by Rugg and colleagues, who used several lines of evidence to argue that old–new ERP differences reflect *relative familiarity* and not recollection. The central assumption, from two-process models of recognition, was that relative familiarity (i.e., a discrepancy between the level of familiarity cued by an item and the level of familiarity that the individual would expect a priori) can function as a basis for recognition judgments. According to Jacoby and Dallas (1981), for example, familiarity can be based on fluent perceptual processing of repeated items. In one study, injections of the anticholinergic agent scopolamine were found to produce a decrement in recognition performance along with an increase in old–new ERP differences (Potter, Pickles, Roberts, & Rugg, 1992). The authors suggested that the drug had a detrimental effect on recollection that coincided with an increase in the extent to which recognition judgments were based on relative familiarity (i.e., the converse of the hypothesis of Smith and Halgren, 1989). In other experiments, the finding that old–new ERP differences were apparent for low-frequency words but not for high-frequency words was interpreted in relation to the idea that low-frequency words give rise to a higher level of relative familiar-

ity, an idea that may also explain the recognition advantage for such words (Rugg, 1990; Rugg & Doyle, 1992). However, if we grant that recognized low-frequency words also tend to engage more recollection (e.g., Gardiner & Java, 1990), this evidence is equivocal with respect to associating ERPs with familiarity versus recollection.

Conclusively determining whether or not ERPs can be linked to recollection versus other factors required a different empirical approach. In an attempt to isolate brain waves specific to recollection, Paller and Kutas (1992) introduced a procedure based on producing behavioral dissociations between recollection and priming. Instead of only comparing ERPs to old versus new words, ERPs were compared between different types of old words presented during an implicit memory test, the word-identification test. A levels-of-processing manipulation at study was used to produce a memory dissociation of the sort previously demonstrated by Jacoby and Dallas (1981). When subjects study words by focusing on word meaning for one set of words and letter identity for another set of words, the two sets of words are generally associated with different levels of recall performance but similar levels of priming (Richardson-Klavehn & Bjork, 1988; Roediger & McDermott, 1993; but see Bentin, Moscovitch, & Nirhod, 1998; Challis, Velichkovsky, & Craik, 1996; Richardson-Klavehn & Gardiner, 1998). Of course, memory tests cannot be relied on for absolutely pure measures of underlying memory phenomena (i.e., memory tests are not process-pure). Nonetheless, reliable dissociations between certain priming tests and explicit memory tests provide a key ingredient for a convincing interpretation of the ERP findings. The two types of studied words under scrutiny differed in the extent to which recollection was provoked, but they were the same with respect to (a) physical stimulus characteristics, by virtue of the counterbalanced design, (b) the behavioral responses made in the implicit memory test, and (c) the magnitude of priming. Accordingly, we hypothesized that the ERP difference wave was an electrophysiological correlate of recollection, distinct from the influence of priming or other confounding factors. These results constitute the first published evidence to strongly support the hypothesis that the subjective experience of recollection can be monitored via measures of electrical activity of the brain (see also, Düzel, Yonelinas, Mangun, Heinze, & Tulving, 1997; Rugg, Mark, Wall, Schloerscheidt, Birch, & Allan, 1998; Smith, 1993; Smith & Guster, 1993; Wilding, Doyle, & Rugg, 1995).

This hypothesis was supported by results from several follow-up experiments. First, a variation of the original design was used along with a between-subjects manipulation of the extent to which conscious recollection occurred (Paller, Kutas, & McIsaac, 1995). Results substantiated the association between recollection and the ERP difference computed between the

two study tasks, and also showed that it was not specific to the particular circumstances in the initial experiment. Consistent with this conclusion, abnormal ERP responses were observed when the same experimental paradigm was used with elderly participants with mild declarative memory impairments (Joyce, Paller, McIsaac, & Kutas, 1998), and when a related paradigm was used with amnesic patients with moderate to severe declarative memory impairments (Olichney et al., in press). Indeed, Olichney and colleagues found that the magnitude of late positive ERP differences was correlated with the extent of memory impairment in these patients, either measured via memory tests for the words used in the experiment or via standardized neuropsychological tests.

Other experiments showed that ERP correlates of recollection can also be elicited by photographs of faces (Paller, Bozic, Ranganath, Grabowecky, & Yamada, 1999; Paller, Gonsalves, Grabowecky, Bozic, & Yamada, 2000). In these experiments, participants were first asked to memorize a set of faces that were each accompanied by a voice simulating the voice of that individual (e.g., ''I'm Alison and I won the Boston Marathon twice''). In a test phase, ERPs to these faces (''remember faces'') showed reliable differences in comparison to ERPs to new faces as well as to ERPs to faces that had been presented in the study phase without voices and with instructions to forget (''forget faces''). Importantly, results from a separate behavioral experiment showed that priming did not differ between the two types of studied faces, whereas remember faces were later recognized much more accurately. The ERP difference between the two types of studied faces can therefore be taken as a neural correlate of recollection, disentangled from ERP correlates of priming. Of course, it will also be necessary to specify the full range of cognitive processes that contribute to recollective experience in these situations. In this regard, we have speculated that retrieval occurs through interactions between frontal and posterior cortical areas, and this speculation has been supported by combined fMRI and ERP evidence obtained using the same general paradigm (Paller et al., 2000). Furthermore, other ERP results likewise suggest that frontal regions are particularly important for successful episodic retrieval (Ranganath & Paller, 1999a, 1999b, 2000), as do many recent results from neuroimaging and neuropsychology (see Nolde, Johnson, & Raye, 1998; Rugg & Wilding, 2000; Shimamura, 1996).

A strategy analogous to the one described above for visual stimuli has also been used in the auditory modality (Gonsalves & Paller, 2000). Spoken words were again presented under two different study conditions using a levels-of-processing procedure. Priming of lexical decision response time was the same across different study conditions, whereas recognition was not. We interpreted test-phase ERPs that differed due to study conditions

as electrical reflections of the visual imagery engaged when the words were heard. This differential imagery took place both during study and test, even though the test condition did not require it. Indeed, we speculated that our electrophysiological measures reflected the recapitulation of this visual imagery, a prominent factor supporting accurate word recognition.

In sum, studies of old-new ERP differences initially led to divergent hypotheses about relationships between ERPs and memory retrieval, whereas subsequent studies succeeded in isolating ERP correlates of recollective processing. Due to the fact that both recollection and priming tend to occur when studied items are presented in a memory test, conventional analyses of old–new ERP effects are typically equivocal with respect to isolating ERPs associated with one type of memory or the other. Successful demonstrations of ERP correlates of recollection required (a) recording ERPs during recognition as well as in situations wherein subjects were not required to make overt recognition responses; (b) comparing study conditions that differentially influenced declarative memory and priming; and (c) including behavioral measures to confirm the memory dissociations. The idea that the neural events responsible for conscious recollection can be observed as they occur is important because such observations should prove helpful for testing theories about the neurophysiology of memory. Moreover, this approach can potentially enrich our understanding of *both* the cognitive structure and the neural substrates of memory.

IV. Electrophysiological Measures of Perceptual Priming

So far I have focused on processes associated with recollection, but it is also illuminating to consider neural correlates of priming. The same experimental approach has been used, here taking advantage of manipulations that influenced priming more than recollection. In one experiment, words were displayed at study either as complete words or by showing one letter at a time in quick succession (Paller, Kutas, & McIsaac, 1998). Priming was enhanced in the former relative to the latter condition, whereas recognition was relatively unaffected. ERPs corresponding to this differential priming were recorded during the priming test, when all words were displayed as complete words. This ERP correlate of priming was interpreted as a reflection of differential processing of visual word form (i.e., words represented as whole units), given that so many other aspects of word processing were matched in the two study conditions. However, this ERP correlate of priming thus reflects just one subtle memory phenomenon, not all possible consequences of prior experience with a word. The ERP correlate of visual word-form priming recorded by Paller and colleagues (1998) differed from

ERP correlates of recollection cited above in that it occurred slightly earlier and had a focal topography centered at occipital scalp locations. Similar results were found in a subsequent experiment in which priming was manipulated by presenting words either spelled forwards or backwards (Paller & Gross, 1998). Results from these two experiments together suggest that ERPs can provide online measures to monitor priming (see also, Joyce, Paller, Schwartz, & Kutas, 1999; Rugg et al., 1998). In addition, these ERP correlates of priming confirm the necessity of valid experimental procedures to disentangle ERP correlates of implicit and explicit memory.

Perceptual priming is known to occur even when an individual cannot explicitly distinguish between old and new items, and it has been related to the concept of unconscious memory. Speculations about unconscious memory are plentiful—from Freud's explorations of the unconscious to current concerns with repressed memories—but direct measurements of unconscious memory are rare and controversial. Possible electrophysiological correlates of priming have also been revealed in recordings from single neurons in monkey visual cortex (Desimone, 1996). Some neurons in ventral temporal areas, in particular, tend to show reduced responses during stimulus repetitions, or ''repetition suppression.'' Neuroimaging in humans also suggests that priming may result from decreased neural activity following perceptual learning, which may be the essence of efficient perceptual processing (Wiggs & Martin, 1998).

A direct comparison between an electrophysiological correlate of recollection and an electrophysiological correlate of perceptual priming is shown in Figure 5. These measures of brain events underlying recollection and priming provide new empirical footholds for theoretical advances regarding the critical differences between memories that are accessible to consciousness and those that are not—and perhaps between conscious and unconscious events in general.

V. Conclusions: Cognitive Neuroscience and Human Memory

Neuropsychological studies of amnesia have shown which brain areas are essential for various memory functions, but moving ahead to seek a comprehensive understanding of how these functions are implemented in the brain, and of their precise cognitive structure, will require an alliance among multiple empirical approaches, both in healthy and brain-damaged people. The usefulness of electrophysiological and functional neuroimaging techniques in this regard will require experimental designs that take optimal advantage of the spatial and temporal resolution provided by each method. Not only must research in this area take into account the strengths and

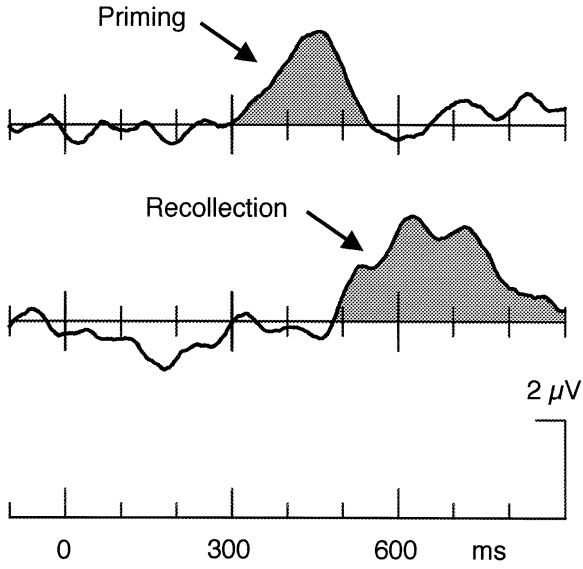


Fig. 5. Brain potentials specifically associated with visual word-form priming and recollection. The upper wave form was computed by subtracting brain potentials elicited by words previously viewed forwards or backwards; priming was greater in the former compared to the latter condition (Paller & Gross, 1998). The lower wave form was computed by subtracting brain potentials elicited by words previously studied in an imagery task versus an orthographic task; recollection was stronger in the former compared to the latter condition (Paller & Kutas, 1992). Recordings were made from the midline parietal scalp location, and positive potentials are shown as upward deflections.

limitations of each technique for measuring brain function, but the successful application of these techniques depends critically on whether subjects' cognitive activities can be adequately controlled.

In several experiments, results have supported the speculation that a particular brain potential provided an objective measure of an unobservable phenomenon, the conscious experience of retrieving a memory. This electrophysiological correlate of conscious recollection contrasts with other findings showing that different brain potentials are associated with perceptual priming of visual word form. The further development of such contrasts between recollection and priming using this methodology should lead to a rich source of evidence pertaining to the neural implementation of these distinct memory functions.

Studies of the neural bases of human memory have the potential for expanding the insights gained from prior cognitive and neuropsychological studies of memory. Contrasts between recollection and priming, in particu-

lar, have launched massive theoretical development. Although more work will be required to adequately test and develop these ideas, a tenable working hypothesis is that recollection and priming can be characterized as different types of neocortical memory dependent on distinct neurophysiological mechanisms. Priming appears to depend on isolated instances of neocortical plasticity such that subsequent processing in one or more cortical regions is altered. In contrast, recollection requires the formation of links among sets of neocortical regions in the service of creating an enduring declarative memory. The aforementioned evidence that ERPs can be used to monitor processes associated with recollection, as well as processes associated with priming, suggests that this approach will be useful for developing more fine-grained hypotheses about the neurophysiology of the two types of memory.

Ultimately, such efforts to understand the neural bases of human memory will shed light not only on the nature and organization of memory processes in the brain, but also on the subjective experience of conscious recollection. However, understanding the neural implementation of memory functions and understanding the cognitive structure of memory should ideally be conceived not as two separate goals. These are fundamentally two parts of the same endeavor. Advances in understanding neural implementations turn on the accuracy of assumptions about cognitive structure. Advances in understanding cognition, as argued above, must ultimately be grounded in an accurate conception of the neural substrates of cognition. Seeking to understand human memory is a colossal challenge and it requires the full force of both cognitive science and brain science.

Acknowledgments

The author's research was supported by grant NS34639 from the National Institute of Neurological Diseases and Stroke and by research funds provided by Northwestern University. Portions of this chapter appeared in a review published in *Behavioural Neurology* (Paller, 2000). Thanks to Dedre Gentner, Marcia Grabowecky, Marta Kutas, Doug Medin, and Larry Squire for their input and encouragement.

References

Atkinson, R. C., & Juola, J. F. (1973). Factors influencing the speed and accuracy of word recognition. In S. Kornblum (Ed.), *Attention and performance IV* (pp. 583–612). New York: Academic.

Barrett, S. E., Rugg, M. D., & Perrett, D. I. (1988). Event-related potentials and the matching of familiar and unfamiliar faces. *Neuropsychologia, 26,* 105–117.

Bentin, S., & Moscovitch, M. (1990). Psychophysiological indices of implicit memory performance. *Bulletin of the Psychonomic Society, 28,* 346–352.

Bentin, S., Moscovitch, M., & Heth, I. (1992). Memory with and without awareness: Performance and electrophysiological evidence of savings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1270–1283.

Bentin, S., Moscovitch, M., & Nirhod, O. (1998). Levels of processing and selective attention effects on encoding in memory. *Acta Psychologica, 98,* 311–341.

Brewer, J. B., Zhao, Z., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. E. (1998). Making memories: Brain activity that predicts how well visual experience will be remembered. *Science, 281,* 1185–1187.

Cermak, L. S., Talbot, N., Chandler, K., & Wolbarst, L. R. (1985). The perceptual priming phenomenon in amnesia. *Neuropsychologia, 23,* 615–622.

Cermak, L. S., Bleich, R. P., & Blackford, S. P. (1988). Deficits in the implicit retention of new associations by alcoholic Korsakoff patients. *Brain and Cognition, 7,* 145–156.

Challis, B. H., Velichkovsky, B. M., & Craik, F. I. M. (1996). Levels-of-processing effects on a variety of memory tasks: New findings and theoretical implications. *Consciousness and Cognition, 5,* 142–164.

Chun, M. M., & Phelps, E. A. (1999). Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nature Neuroscience, 2,* 844–847.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Sciences USA, 93,* 13494–13499.

Dror, I. E., & Gallogly, D. P. (1999). Computational analyses in cognitive neuroscience: In defense of biological implausibility. *Psychonomic Bulletin & Review, 6,* 173–182.

Düzel, E., Yonelinas, A. P., Mangun, G. R., Heinze, H.-J., & Tulving, E. (1997). Event-related brain potential correlates of two states of conscious awareness in memory. *Proceedings of the National Academy of Sciences USA, 94,* 5973–5978.

Fabiani, M., Karis, D., & Donchin, E. (1986). P300 and recall in an incidental memory paradigm. *Psychophysiology, 23,* 298–308.

Fernández, G., Weyerts, H., Schrader-Bölsche, M., Tendolkar, I., Smid, H. G. O. M., Tempelmann, C., Hinrichs, H., Scheich, H., Elger, C. E., Mangun, G. R., & Heinze, H.-J. (1998). Successful verbal encoding into episodic memory engages the posterior hippocampus: A parametrically analyzed function magnetic resonance imaging study. *Journal of Neuroscience, 18,* 1841–1847.

Fernández, G., Grunwald, T., Effern, A., Pezer, N., Lehnertz, J., Dümpelmann, N., & Elger, C. E. (1999). Serial processes of declarative memory formation in the human medial temporal lobe. *NeuroImage, 9,* S921. (Abstract).

Flanagan, O. (1992). *Consciousness reconsidered*. Cambridge, MA: MIT.

Friedman, D. (1990). ERPs during continuous recognition memory for words. *Biological Psychology, 30,* 61–87.

Friedman, D., & Sutton, S. (1987). Event-related potentials during continuous recognition memory. *Electroencephalography and Clinical Neurophysiology, Supplement 40,* 316–321.

Fuster, J. M. (1995). *Memory in the cerebral cortex*. Cambridge, MA: MIT.

Gabrieli, J. D. E. (1998). Cognitive neuroscience of human memory. *Annual Review of Psychology, 49,* 87–115.

Gabrieli, J. D. E., Keane, M. M., Zarella, M., M., & Poldrack, R. A. (1997). Preservation of implicit memory for new associations in global amnesia. *Psychological Science, 8,* 326–329.

Gardiner, J. M., & Java, R. I. (1990). Forgetting in recognition memory with and without recollective experience. *Memory & Cognition, 18,* 23–30.

Gonsalves, B., & Paller, K. A. (2000). Brain potentials associated with recollective processing of spoken words. *Memory & Cognition, 28,* 321–330.

Haist, F., Musen, G., & Squire, L. R. (1991). Intact priming of words and nonwords in amnesia. *Psychobiology, 19,* 275–285.

Hamann, S. B., & Squire, L. R. (1997). Intact perceptual priming in the absence of conscious memory. *Behavioral Neuroscience, 111,* 850–854.

Hamann, S. B., Squire, L. R., & Schacter, D. L. (1995). Perceptual thresholds and priming in amnesia. *Neuropsychology, 9,* 3–15.

Hebb, D. O. (1949). *Organization of Behavior*. New York: Wiley and Sons.

Hillyard, S. A., Squires, K. C., Bauer, J. W., & Lindsay, P. H. (1971). Evoked potential correlates of auditory signal detection. *Science, 172,* 1357–1360.

Hodges, J. R., & Graham, K. S. (1998). A reversal of the temporal gradient for famous person knowledge in semantic dementia: Implications for the neural organisation of long-term memory. *Neuropsychologia, 36,* 803–825.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110,* 306–340.

Johnson, R., Jr. (1995). Event-related potential insights into the neurobiology of memory systems. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 10, pp. 135–163). Amsterdam: Elsevier.

Johnson, R., Jr., Kreiter, K., Russo, B., & Zhu, J. (1998). A spatio-temporal analysis of recognition-related event-related brain potentials. *International Journal of Psychophysiology, 29,* 83–104.

Johnson, R., Jr., Pfefferbaum, A., & Kopell, B. S. (1985). P300 and long-term memory: Latency predicts recognition performance. *Psychophysiology, 22,* 497–507.

Joyce, C. A., Paller, K. A., McIsaac, H. K., & Kutas, M. (1998). Memory changes with normal aging: Behavioral and electrophysiological measures. *Psychophysiology, 35,* 669–678.

Joyce, C. A., Paller, K. A., Schwartz, T. J., & Kutas, M. (1999). An electrophysiological analysis of modality-specific aspects of word repetition. *Psychophysiology, 36,* 655–665.

Kapur, N. (1993). Focal retrograde amnesia in neurological disease: A critical review. *Cortex, 29,* 217–234.

Karis, D., Fabiani, M., & Donchin, E. (1984). ''P300'' and memory: Individual differences in the von Restorff effect. *Cognitive Psychology, 16,* 177–216.

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain Research Reviews, 29,* 169–195.

Kutas, M. (1988). Review of event-related potential studies of memory. In M. S. Gazzaniga (Ed.), *Perspectives in memory research* (pp. 181–218). Cambridge, MA: MIT.

Kutas, M., & Dale, A. (1997). Electrical and magnetic readings of mental functions. In M. D. Rugg (Ed.), *Cognitive neuroscience* (pp. 197–242). Cambridge, MA: MIT.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252–271.

Markowitsch, H. J. (1995). Which brain regions are critically involved in the retrieval of old episodic memory? *Brain Research Reviews, 21,* 117–127.

Marr, D. (1982). *Vision.* San Francisco, CA: Freeman.

Martinez, J. L., Jr., & Kesner, R. P. (Eds.) (1998). *Neurobiology of learning and memory*. San Diego, CA: Academic.

Mayes, A. R., & Downes, J. J. (1997). *Theories of organic amnesia*. East Sussex, UK: Erlbaum.

Mayes, A. R., & Gooding, P. (1989). Enhancement of word completion priming in amnesics by cueing with previously novel associates. *Neuropsychologia, 27,* 1057–1072.

Milner, B., Squire, L. R., & Kandel, E. R. (1998). Cognitive neuroscience and the study of memory. *Neuron, 20,* 445–468.

Moscovitch, M., Vriezen, E., & Goshen-Gottstein, Y. (1993). Implicit tests of memory in patients with focal lesions and degenerative brain disorders. In H. Spinnler & F. Boller (Eds.), *Handbook of neuropsychology* (pp. 133–173). Amsterdam: Elsevier.

Musen, G., & Squire, L. R. (1993). Implicit learning of color–word associations using a Stroop paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 789–798.

Neville, H., Kutas, M., Chesney, G., & Schmidt, A. L. (1986). Event-related brain potentials during initial encoding and recognition memory of congruous and incongruous words. *Journal of Memory and Language, 25,* 75–92.

Neville, H., Snyder, E., Woods, D., & Galambos, R. (1982). Recognition and surprise alter the human visual evoked response. *Proceedings of the National Academy of Sciences USA, 79,* 2121–2123.

Nolde, S. F., Johnson, M. K., & Raye, C. L. (1998). The role of prefrontal cortex during tests of episodic memory. *Trends in Cognitive Sciences, 2,* 399–406.

Olichney, J. M., Van Petten, C., Paller, K. A., Salmon, D. P., Iragui, V. J., & Kutas, M. (in press). Word repetition in amnesia: Electrophysiological measures of impaired and spared memory. *Brain*.

Paller, K. A. (1990). Recall and stem-completion priming have different electrophysiological correlates and are modified differentially by directed forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 1021–1032.

Paller, K. A. (1993). Elektrophysiologische Studien zum Menschlichen Gedächtnis [Electrophysiological studies of human memory]. *Zeitschrift für Elektroenzephalographie, Elektromyographie und verwandte Gebiete, 24,* 24–33.

Paller, K. A. (1997). Consolidating dispersed neocortical memories: The missing link in amnesia. *Memory, 5,* 73–88.

Paller, K. A. (2000). Electrophysiological measures of human memory functions. *Behavioral Neurology,* in press.

Paller, K. A., Bozic, V. S., Ranganath, C., Grabowecky, M., & Yamada, S. (1999). Brain waves following remembered faces index conscious recollection. *Cognitive Brain Research, 7,* 519–531.

Paller, K. A., Gonsalves, B., Grabowecky, M., Bozic, V. S., & Yamada, S. (2000). Electrophysiological correlates of recollecting faces of known and unknown individuals. *NeuroImage, 11,* 98–110.

Paller, K. A., & Gross, M. (1998). Brain potentials associated with perceptual priming versus explicit remembering during the repetition of visual word-form. *Neuropsychologia, 36,* 559–571.

Paller, K. A., & Kutas, M. (1992). Brain potentials during memory retrieval provide neurophysiological support for the distinction between conscious recollection and priming. *Journal of Cognitive Neuroscience, 4,* 375–391.

Paller, K. A., Kutas, M., & Mayes, A. R. (1987). Neural correlates of encoding in an incidental learning paradigm. *Electroencephalography and Clinical Neurophysiology, 67,* 360–371.

Paller, K. A., Kutas, M., & McIsaac, H. K. (1995). Monitoring conscious recollection via the electrical activity of the brain. *Psychological Science, 6,* 107–111.

Paller, K. A., Kutas, M., & McIsaac, H. K. (1998). An electrophysiological measure of priming of visual word-form. *Consciousness and Cognition, 7,* 54–66.

Paller, K. A., & Mayes, A. R. (1994). New-association priming of word identification in normal and amnesic subjects. *Cortex, 30,* 53–73.

Paller, K. A., Mayes, A. R., McDermott, M., Pickering, A. D., & Meudell, P. R. (1991). Indirect measures of memory in a duration-judgement task are normal in amnesic patients. *Neuropsychologia, 29,* 1007–1018.

Paller, K. A., McCarthy, G., & Wood, C. C. (1988). ERPs predictive of subsequent recall and recognition performance. *Biological Psychology, 26,* 269–276.

Paller, K. A., Ranganath, C., LaBar, K. S., Parrish, T. B., Gitelman, D. R., & Mesulam, M.-M. (2000). Neural correlates of memory for faces. Manuscript under review.

Potter, D. D., Pickles, C. D., Roberts, R. C., & Rugg, M. D. (1992). The effects of scopolamine on event-related potentials in a continuous recognition memory task. *Psychophysiology, 29,* 29–37.

Ranganath, C., & Paller, K. A. (1999a). Frontal brain activity during episodic and semantic retrieval: Insights from event-related potentials. *Journal of Cognitive Neuroscience, 11,* 598–609.

Ranganath, C., & Paller, K. A. (1999b). Frontal brain potentials during recognition are modulated by requirements to retrieve perceptual detail. *Neuron, 22,* 605–613.

Ranganath, C., & Paller, K. A. (2000). Neural correlates of memory retrieval and evaluation. *Cognitive Brain Research, 9,* 209–222.

Richardson-Klavehn, A., & Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology, 39,* 475–543.

Richardson-Klavehn, A. & Gardiner, J. M. (1998). Depth-of-processing effects on priming in stem completion: Tests of the voluntary-contamination, conceptual-processing, and lexical-processing hypotheses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 593–609.

Roediger, H. L., III, & McDermott, K. B. (1993). Implicit memory in normal human subjects. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 8, pp. 63–131). Amsterdam: Elsevier.

Rösler, F., Heil, M., & Roder, B. (1997). Slow negative brain potentials as reflections of specific modular resources of cognition. *Biological Psychology, 45,* 109–141.

Rubin, E. H., & McAdam, D. E. (1972). Slow potential concomitants of the retrieval process. *Electroencephalography and Clinical Neurophysiology, 32,* 84–86.

Ruchkin, D. E., & Sutton, S. (1978). Emitted P300 potentials and temporal uncertainty. *Electroencephalography and Clinical Neurophysiology, 45,* 268–277.

Rugg, M. D. (1995). Event-related potential studies of human memory. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 789–801). Cambridge, MA: MIT.

Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high- and low-frequency words. *Memory & Cognition, 18,* 367–379.

Rugg, M. D., & Doyle, M. C. (1992). Event-related potentials and recognition memory for low- and high-frequency words. *Journal of Cognitive Neuroscience, 4,* 69–79.

Rugg, M. D., & Nagy, M. E. (1989). Event-related potentials and recognition memory for words. *Electroencephalography and Clinical Neurophysiology, 72,* 395–406.

Rugg, M. D., Roberts, R. C., Potter, D. D., Pickles, C. D., & Nagy, M. E. (1991). Event-related potentials related to recognition memory: Effects of unilateral temporal lobectomy and temporal lobe epilepsy. *Brain, 114,* 2313–2332.

Rugg, M. D., Mark, R. E., Wall, P., Schloerscheidt, A. M., Birch, C. S., & Allan, K. (1998). Dissociation of the neural correlates of implicit and explicit memory. *Nature, 392,* 595–598.

Rugg, M. D., & Wilding, E. L. (2000). Retrieval processing and episodic memory. *Trends in Cognitive Sciences, 4,* 108–115.

Sanquist, T. F., Rohrbaugh, J. W., Syndulko, K., & Lindsley, D. B. (1980). Electrocortical signs of levels of processing: Perceptual analysis and recognition memory. *Psychophysiology, 17,* 568–576.

Schacter, D. L., Chiu, C. Y. P., & Ochsner, K. N. (1993). Implicit memory: A selective review. *Annual Review of Neuroscience, 16,* 159–182.

Schacter, D. L., & Graf, P. (1986). Preserved learning in amnesic patients: Perspectives from research on direct priming. *Journal of Clinical and Experimental Neuropsychology, 8,* 727–743.

Schacter, D. L. Norman, K. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology, 49,* 289–318.

Schacter, D. L., & Tulving, E. (Eds.). (1994). *Memory systems 1994.* Cambridge, MA: MIT.

Shimamura, A. P. (1986). Priming effects in amnesia: Evidence for a dissociable memory function. *Quarterly Journal of Experimental Psychology, 38A,* 619–644.

Shimamura, A. P. (1993). Neuropsychological analyses of implicit memory: History, methodology, and theoretical interpretations. In P. Graf & M. E. J. Masson (Eds.), *Implicit Memory: New directions in cognition, development, and neuropsychology* (pp. 265–285). Hillsdale, NJ: Erlbaum.

Shimamura, A. P. (1996). The role of prefrontal cortex in monitoring and controlling memory processes. In L. Reder (Ed.), *Implicit memory and metacognition* (pp. 259–274). Mahwah, NJ: Erlbaum.

Shimamura, A. P., & Squire, L. R. (1989). Impaired priming of new associations in amnesia. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 721–728.

Smith, M. E. (1993). Neurophysiological manifestations of recollective experience during recognition memory judgments. *Journal of Cognitive Neuroscience, 5,* 1–13.

Smith, M. E., & Guster, K. (1993). Decomposition of recognition memory event-related potentials yields target, repetition, and retrieval effects. *Electroencephalography and Clinical Neurophysiology, 86,* 335–343.

Smith, M. E., & Halgren, E. (1987). Event-related potentials elicited by familiar and unfamiliar faces. *Electroencephalography and Clinical Neurophysiology, Supplement 40,* 422–426.

Smith, M. E., & Halgren, E. (1989). Dissociation of recognition memory components following temporal lobe lesions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 50–60.

Squire, L. R. (1987). *Memory and brain*. Oxford: Oxford University Press.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99,* 195–231.

Squire, L. R., Hamann, S. B., & Schacter, D. L. (1996). Intact baseline performance and priming in amnesia: Reply to Ostergaard and Jernigan. *Neuropsychology, 10,* 131–135.

Squire, L. R., & Kandel, E. R. (1999). *Memory: From mind to molecule*. New York: Freeman.

Squire, L. R., & Knowlton, B. J. (2000). The medial temporal lobe, the hippocampus, and the memory systems of the brain. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 765–779). Cambridge, MA: MIT Press.

Squire, L. R., & Paller, K. A. (2000). Biology of memory. In B. J. Sadock & V. A. Sadock (Eds.), *Kaplan & Sadock's comprehensive textbook of psychiatry* (7th ed., pp. 425–437). Baltimore, MD: Williams & Wilkins.

Wagner, A. D., Schacter, D. L., Rotte, M., Koutstaal, W., Maril, A., Dale, A. M., Rosen, B. R., & Buckner, R. L. (1998). Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science, 281,* 1188–1191.

Warren, L. R. (1980). Evoked potential correlates of recognition memory. *Biological Psychology, 11,* 21–35.

Weiss, S., & Rappelsberger, P. (2000). Long-range EEG synchronization during word encoding correlates with successful memory performance. *Cognitive Brain Research, 9,* 299–312.

Wiggs, C. L., & Martin, A. (1998). Properties and mechanisms of perceptual priming. *Current Opinion in Neurobiology, 8,* 227–233.

Wilding, E. L., Doyle, M. C., & Rugg, M. D. (1995). Recognition memory with and without retrieval of context: An event-related potential study. *Neuropsychologia, 33,* 743–767.

This Page Intentionally Left Blank

STRUCTURAL INFLUENCES ON IMPLICIT AND EXPLICIT SEQUENCE LEARNING

Tim Curran, Michael D. Smith, Joseph M. DiFranco, and Aaron T. Daggy

I. Introduction

Sequential behavior and learning have been studied from a variety of perspectives. Relevant research has investigated serial list recall (reviewed by Crowder & Greene, 2000), extrapolation of letter and digits series (e.g., Simon & Kotovsky, 1963; Simon & Kotovsky, 1972), and the production of learned movement sequences (e.g., Povel & Collard, 1982; Restle, 1970; Restle & Brown, 1970; Rosenbaum, Kenny, & Derr, 1983). Each of these lines of research has manipulated the structural characteristics of sequences as one way to elucidate the nature of the governing processes and representations. For example, Lashley's (1951) seminal rejection of associative chaining included consideration of the structural features of real-world serial behaviors. Lashley noted that pronunciation of ''right'' and ''tire'' involve the identical motor elements in reverse order (Lashley, 1951). These same motor elements are used to pronounce many other words, so it is unlikely that a serial behavior such as speaking is governed by associative chains with intrinsic orders.

Over the last decade, many sequence learning studies have used Nissen and Bullemer's (1987) serial reaction time (SRT) task. Subjects typically watch a simple stimulus (e.g., an asterisk) move between various visual positions and press corresponding response keys as quickly and accurately

THE PSYCHOLOGY OF LEARNING AND MOTIVATION, VOL. 40

Copyright © 2001 by Academic Press.
All rights of reproduction in any form reserved.
0079-7421/00 $35.00

as possible. Sequential learning is measured via performance improvements that arise from practice with a repeating sequence (usually measured against baseline conditions with randomly appearing stimuli). SRT learning is typically regarded to be "implicit" in two respects (for more thorough definitional considerations, see Frensch, 1997; Shanks & St. John, 1994; Stadler & Roediger, 1997). First, subjects are not informed about the sequence, so they are not intentionally trying to learn it. Second, subjects often demonstrate performance improvements even when they show little or no explicit knowledge of the sequence. The implicit nature of SRT learning may make it particularly relevant to common examples of serial learning and behavior, such as skill learning in which fluent performance is not necessarily accompanied by explicit knowledge of the rules that govern behavior. For example (from Carlson, 1998), how can a bicycle rider avoid falling if she is leaning too far leftward? Cyclists may verbally respond that leaning in the opposite direction would straighten the bike, but in practice this merely would precipitate the fall. Most of these same cyclists would correctly turn into the direction of the fall if they were riding under these conditions. Thus, the biker's performance exhibits a sequential skill (i.e., a series of movements needed to straighten the bike) that presumably was acquired through practice, not a product of explicit instruction, and not explicitly accessible.

Nissen and Bullemer (1987) originally investigated two topics that continue to intrigue SRT researchers: effects of attention and amnesia. Both lines of investigation have benefited from research manipulating sequential structure (A. Cohen, Ivry, & Keele, 1990; Curran, 1997b). Nissen and Bullemer concluded that sequence learning required attention but not awareness because learning was hindered by a distraction task (tone counting). Cohen et al. (1990) found that the attentional requirements of sequence learning interact with sequential structure. They noted that Nissen and Bullemer's sequence (4231324321) contained "ambiguous" pairwise associations—each location (e.g., 4) is followed by more than one other location (e.g., 2 or 3). Thus, a simple associative chain model, for example, would have difficulty learning such a sequence because the elements obey different orders in different parts of the sequence. A. Cohen et al. replicated Nissen and Bullemer's demonstration that ambiguous sequences (e.g., 132312) require attention for learning. However, learning with distraction was possible when the sequence structure included some unique associations (either completely unique: 15243; or a hybrid of unique and ambiguous associations: 143132).

Subsequent studies have shown that ambiguous sequences can be learned without attention, and the relationship between attention and sequence learning continues to be a focus of SRT research (Curran & Keele, 1993; Frensch, Buchner, & Lin, 1994; Hsiao & Reber, 1997; Reed & Johnson,

1994; Stadler, 1995). For the present purposes, A. Cohen et al.'s (1990) study was critically important for emphasizing the analysis of sequence structure and its implications for understanding sequence learning. As Lashley (1951) noted, real-world examples of sequential behavior often contain pairwise ambiguities (e.g., pronouncing tire vs. right), so SRT research is most relevant to everyday skills if it can account for these ambiguities.

Other researchers have identified additional types of nonsequential information that may contribute to SRT performance. For example, if sequence and random conditions do not contain equal event frequencies, subjects may respond faster to a sequence by merely learning that some elements occur more frequently than others (Reed & Johnson, 1994; Shanks & St. John, 1994). Another complication arises from the fact that subjects can learn probabilistic information that aids performance (Stadler, 1992). For example in Nissen and Bullemer's sequence, 4231324321, pairwise associations cannot be used to uniquely predict any element from its predecessor, but some pairwise associations are more probable than others (e.g., 3-2 occurs twice and 3-1 occurs once, Jackson & Jackson, 1992). Thus, the relative frequencies of individual elements and pairwise transitions need to be considered if one is interested in studying the learning of "higher order" (e.g., more than pairwise) sequential information. For a thorough review of structural influences on implicit sequence learning, see Hoffman and Koch (1997).

II. Group Differences Revealed by Structural Manipulations

A. Experiment 1: Sequence Learning in Amnesic Patients (Curran, 1997b)

A number of studies have suggested that sequence learning is normal in patients with anterograde amnesia (Nissen & Bullemer, 1997; Nissen, Willingham, & Hartman, 1989; Reber & Squire, 1994). Such a conclusion fits with the common view that implicit learning and memory do not depend on brain regions damaged in amnesia (Moscovitch, Vriezen, & Goshen-Gottstein, 1993; Schacter, Chiu, & Ochsner, 1993; Squire, Knowlton, & Musen, 1993). However, some theoretical perspectives suggest that amnesic patients may have difficulty with sequence learning (Cohen & Eichenbaum, 1993; Gluck & Myers, 1993; Mayes, Meudell, & Pickering, 1985; McClelland, McNaughton, & O'Reilly, 1995; Sutherland & Rudy, 1989; Wicklegren, 1979). Each of these theories generally suggest that the brain areas damaged in amnesia (hippocampus and adjacent medial temporal cortex) support

associative learning that is more complex than the kind of pairwise associative learning that underlies phenomena such as first-order conditioning. Because Nissen and Bullemer's sequence was used in the first two SRT studies with amnesic patients (Nissen & Bullemer, 1997; Nissen, Willingham, & Hartman, 1989), learning could have been facilitated by probabilistic pairwise associations. Reber and Squire's (1994) amnesic patients learned a sequence in which the relative frequencies of individual elements and pairwise transitions were equated with the random baseline condition. Reber and Squire's results suggested some higher-order sequence learning by amnesic patients, but a nonsignificant trend in the data suggested slightly less learning by amnesic than control subjects.

1. Method

Curran (1997b) re-examined the sequence-learning ability of amnesic patients in an experiment intended to account for learning of pairwise probabilities, lessen the contribution of explicit knowledge, and allow for more direct examination of higher-order sequential learning. Two sequences were tested. The first-order predictive (FOP, 1-2-1-4-2-3-4-3-1-4-2-3) sequence included predictive pairwise transitions. Each location (e.g., 1) was followed by one location twice (e.g., 4), another location once (e.g., 2), and never followed by the third location (e.g., 3). Such pairwise information was uninformative in the second-order predictive (SOP, 1-2-1-4-2-3-4-1-3-2-4-3) sequence because each location was followed once by each other location. Therefore, each SOP element could be predicted only by consideration of at least the two previous elements (e.g., 1-2 predicts 1, see Reed & Johnson, 1994). Each amnesic patient was tested with both sequences in separate sessions (at least 3 months apart). Two age- and education-matched control subjects were tested for each amnesic (1 control per sequence). Individual control subjects were not tested with both sequences because their memory for the first session (including explicit knowledge tests and debriefing) would likely influence performance on a second session. Each of 10 blocks contained 120 trials. The first block was practice with stimulus location randomly determined (with the constraint that the same location was never consecutively repeated). The nine experimental blocks alternated between 12-trial cycles of random (R) and sequential (S) stimuli arranged as R-S-S-R-S-S-R-S-S-R. The within-block alternation of random and sequence cycles (following Stadler, 1993) departs from most SRT experiments in which numerous sequential blocks are followed by a single random comparison block. Alternating random and sequence cycles has the advantages of limiting explicit knowledge and increasing the stability of random RT measures used to measure learning of individual sequence positions (as

described below). Each stimulus remained visible until the subject responded, and the response-to-stimulus interval (RSI) between consecutive stimuli was 200 ms.

2. *Results*

Small accuracy differences between random (96%) and sequence (97%) conditions did not interact with group membership or sequence type. RT results (means of subjects' medians) are summarized in Figure 1. Both groups showed significant learning of both sequences, as indicated by faster RTs to sequence than random cycles. There were no significant differences between groups (main effect nor interactions). Overall, there was a significant condition-(random, sequence) by-sequence-type (FOP, SOP) interaction such that random-sequence differences were greater for the SOP than FOP sequence. These results were generally consistent with previous claims that sequence learning is normal in amnesic subjects. Subjects completed a whole-sequence recognition test and a sequence-fragment recognition test after the SRT task. Neither group showed above-chance discrimination on either test, so explicit knowledge was minimal.

Further analyses concentrated on the learning of individual sequence positions with an emphasis on evaluating higher-order sequential learning. Mean RT was computed for each sequence position across all blocks of trials. Analyses focused on higher-order effects by subtracting each sequence position RT from the random RT elicited by the same pairwise transition. These differences are summarized in Figure 2. For example, the rightmost points in Figure 2 (top) indicate that subjects were about 20 ms faster responding to position 3 when it occurred as the last element of the sequence than when 3 followed 2 within random trials. Any such differences that are significantly greater than zero would reflect learning of more than pairwise information because the pairwise information is equivalent in random and sequence conditions. For the FOP sequence (top of Figure 2), *t*-tests indicated that the random-sequence difference was significant ($p < .05$) at three different positions for amnesic (+ in Figure 2) and control subjects (* in Figure 2). Overall, the patterns were very similar between groups. Both groups showed minimal higher-order learning of the FOP sequence. The three FOP positions showing negative differences (sequence slower than random) were all less probable pairwise transitions. For the SOP sequence (bottom of Figure 2), higher-order learning of six positions was evident for controls ($* p < .05$, $** p < .01$), but only three positions were learned by amnesic patients ($+ p < .05$). Thus, control subjects appeared to learn more higher-order information than amnesic patients—despite similar results when RT was collapsed across all sequence positions. These

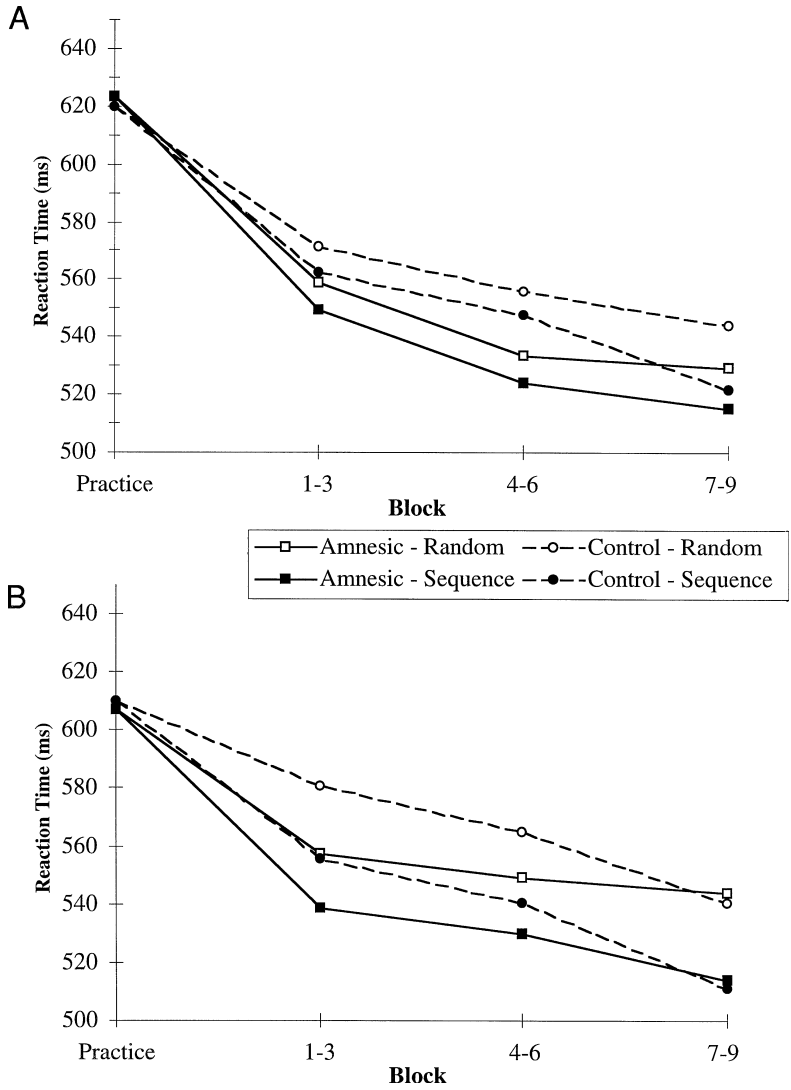


Fig. 1. Reaction time (means of subjects' medians) for amnesic patients and control subjects to random and sequential trials (Experiment 1). (A) First-order predictive sequence (FOP); (B) second-order predictive sequence (SOP). (Reprinted with permission from Curran, 1997b, Figure 1, Copyright 1997, MIT Press, Cambridge, USA.)

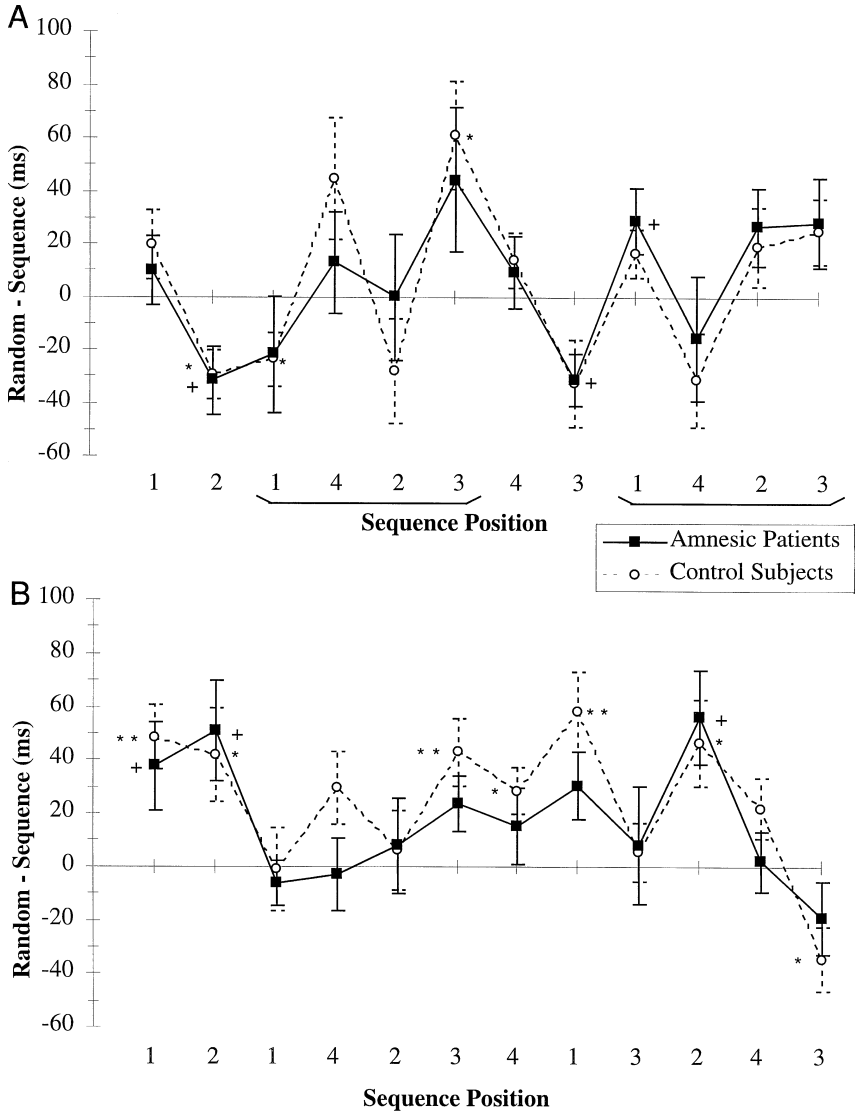


Fig. 2. Differences between reaction times to each sequence element and the same pairwise combinations in random trials (Experiment 1). Reaction times were averaged across all blocks of trials. Positive differences reflect learning of higher-order sequence information. Error bars show the standard error of the mean. *t*-tests (2-tailed) were used to assess the significance of the learning effect for each group and each sequence position separately (control subjects: * $p < .05$, ** $p < .01$; amnesic patients;$^{+}$ $p < .05$). (A) First-order predictive sequence (FOP); (B) second-order predictive sequence (SOP). (Reprinted with permission from Curran, 1997b, Figure 2, Copyright 1997, MIT Press, Cambridge, USA.)

results are consistent with the perspective that subjects can normally learn higher-order sequential information (see also, Cleeremans & McClelland, 1991; A. Cohen et al., 1990; Reed & Johnson, 1994). The control subjects learned twice as many SOP positions as amnesic patients, so amnesia can impair higher-order sequential learning.

B. Experiment 2: Effects of Adult Aging (Curran, 1997a)

1. Primary Results

Older subjects, like patients with amnesia, more often show deficits on explicit than implicit measures of learning and memory (D. Howard & Wiggs, 1993; Kausler, 1994; Light & La Voie, 1993; Mitchell, 1993; Parkin, 1993). Implicit sequence learning was originally reported to be similar for young and older adults (D. Howard & Howard, 1992; J. Howard, Mutter, & Howard, 1992). Jackson and Jackson (1992) noted that previous studies used sequences with predictive pairwise associations. Using a sequence without predictive pairwise associations, young subjects showed significant learning, but middle-aged and older subjects did not (Jackson & Jackson, 1992). A later study, again using a sequence with predictive pairwise associations, reported normal learning in older subjects with ''high ability'' but deficits for subjects with ''low ability'' (based on education, occupation, and verbal ability, Cherry & Stadler, 1995). As a whole, these results suggest that older subjects may be dependent on pairwise information for sequence learning.

Curran (1997a) compared young (undergraduate) and older (mean age = 67 years) subjects ability to learn the previously described FOP and SOP sequences. Comparing RT between random and sequence conditions suggested that young subjects showed similar learning of both FOP and SOP sequences (see Figure 3). For older subjects, the random–sequence conditions interacted with sequence type such that the SOP sequence was better learned than the FOP sequence (replicating the pattern observed in amnesic patients and their controls, averaging 56 years old). Young and older subjects were not directly compared because of large overall RT differences that could artifactually influence random–sequence differences (Chapman, Chapman, Curran, & Miller, 1994). However, interesting group differences were obtained when accuracy was considered (Figure 4). A group-by-condition interaction on accuracy suggested that sequence-random differences indicative of learning were greater for young than older subjects. No sequence type effects approached significance with accuracy as the dependent measure.

Higher-order sequential learning was evaluated by RT differences between each sequence position and corresponding pairwise transitions from

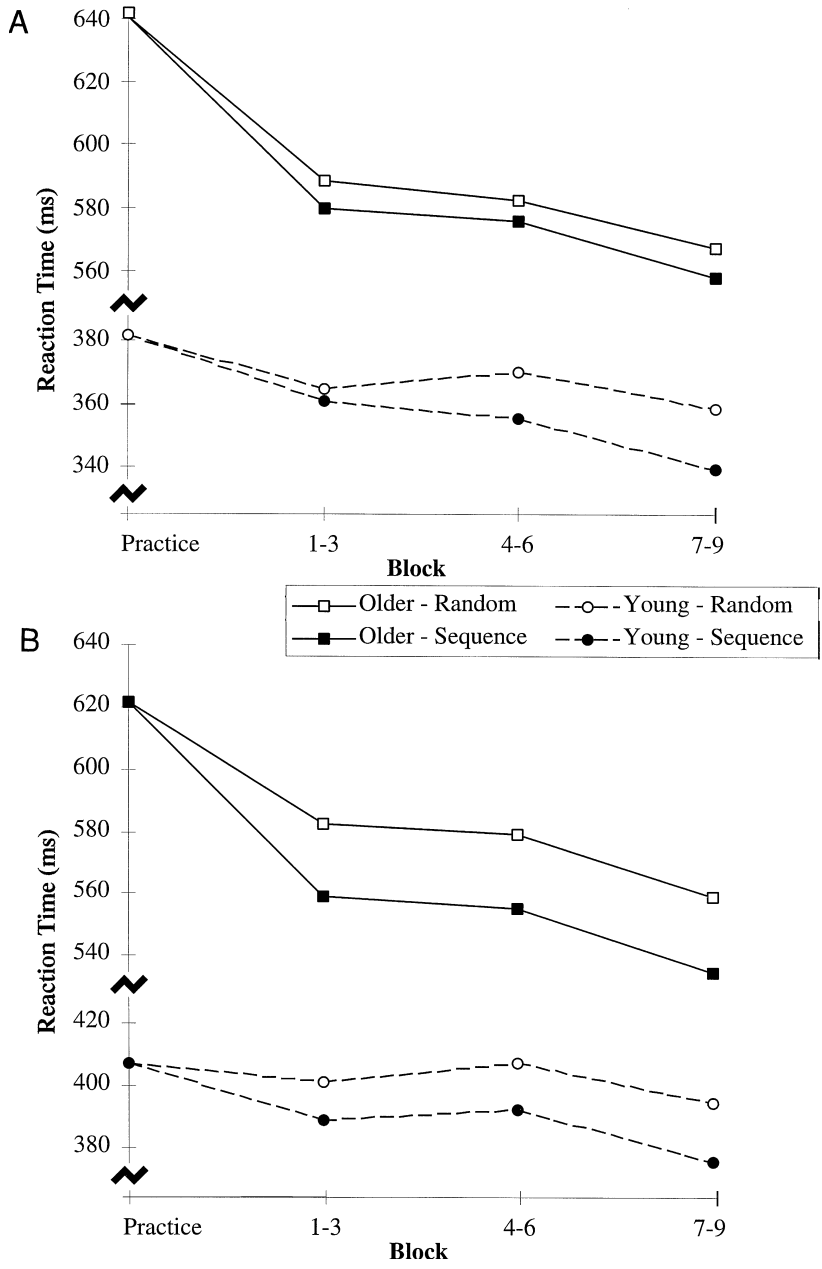


Fig. 3. Reaction time (means of subjects' medians) for older and young subjects to random and sequential trials (Experiment 2). (A) First-order predictive sequence (FOP); (B) second-order predictive sequence (SOP). (Reprinted with permission from Curran, 1997a, Figure 1, Copyright 1997, Springer-Verlag, Germany.)

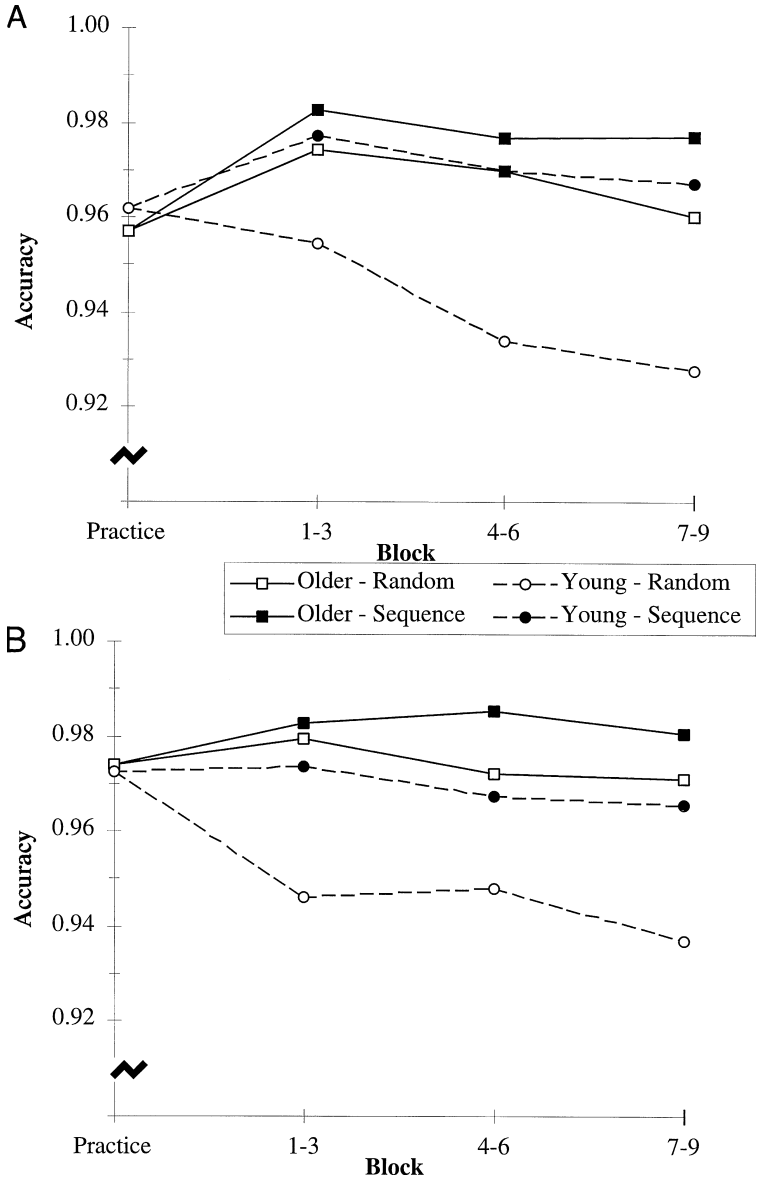


Fig. 4. Accuracy (mean proportion correct) for older and young subjects to random and sequential trials (Experiment 2). (A) First-order predictive sequence (FOP); (B) second-order predictive sequence (SOP). (Reprinted with permission from Curran, 1997a, Figure 2, Copyright 1997, Springer-Verlag, Germany.)

random trials (Figure 5). Learning of the FOP sequence was evident only within a twice-presented subsequence (1-4-2-3, underlined in Figure 5A), and young subjects learned more elements than older subjects. Learning profiles were more similar between young and older subjects for the SOP sequence (Figure 5B). Analogous differences were computed from accuracy at each sequence position. The accuracy pattern indicated that young learned more elements of both sequences than did older subjects (Figure 6).

Recognition tests suggested that explicit knowledge did not account for group differences. Recognition was more accurate for FOP than SOP sequences. FOP recognition was above chance for older subjects ($p < .05$) but only marginally above chance for young subjects ($p = .09$). These recognition patterns contrast with the performance measures of implicit learning: Showing better learning of SOP than FOP sequences (RT) and better learning by young than older subjects (accuracy).

The original prediction was that older subjects would learn FOP sequences better than SOP sequences because they could capitalize on predictive pairwise information. Contrary to this prediction, older subjects learned SOP sequences better than FOP sequences. Although first-order associations were more predictive for FOP than SOP sequences, only the SOP sequence could be completely learned with second-order associations (e.g., 2-1 always predicts 4). Higher-than-second-order information was necessary to learn some FOP segments (e.g., **2**-1-4-2-3 predicts 4, but **3**-1-4-2-3 predicts 1). In summary, older subjects showed an overall learning deficit when accuracy was the dependent measure. RT differences between FOP and SOP sequences suggest that second-order learning could be normal in older subjects, but learning was especially impaired when higher-than-second-order associations were required. This pattern contrasts with amnesic patients who specifically showed less sensitivity to second-order associations in the SOP sequence.

2. *Lag Effects*

Follow-up analyses revealed strong positive correlations between learning of each position (Figure 5 and 6) and the lag between repetitions of the same location [all $r \geq .50$, both groups, both sequences, both accuracy and RT). Clegg (1998) also observed that sequential RT was correlated with repetition lag. Learning differences were most likely to be negative for reversals in which repetitions were only separated by one other stimulus (e.g., 1-2-*1* and 3-4-*3* in Figure 5A). Learning effects were stronger for positions with longer lags between repetitions. A similar pattern was observed in Experiment 1 (Figure 2, Curran, 1997b). Lag effects also were observed in an SRT experiment without a repeating sequence, but with

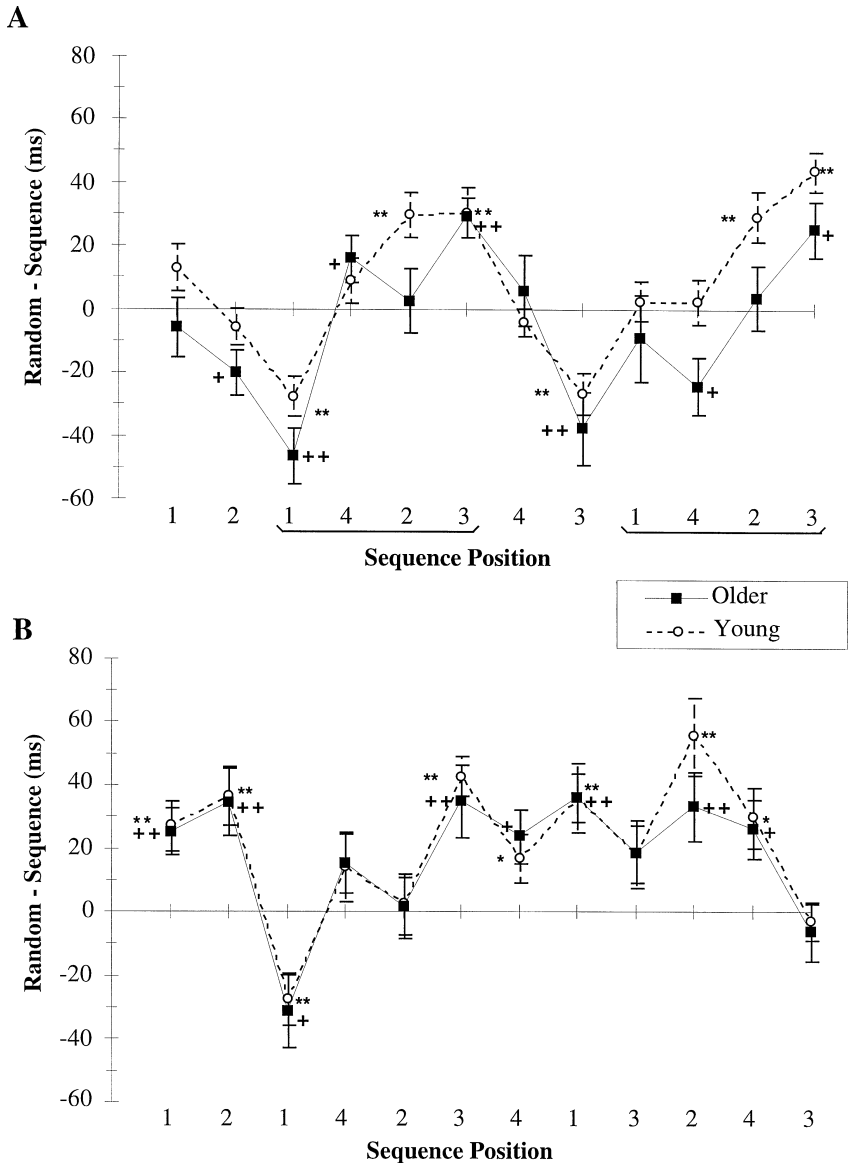


Fig. 5. Differences between reaction times to each sequence element and the same pairwise combinations in random trials (Experiment 2). Reaction times were averaged across all blocks of trials. Positive differences reflect learning of higher-order sequence information. Error bars show the standard error of the mean. *t*-tests (2-tailed) were used to assess the significance of the learning effect for each group and each sequence position separately (young subjects: * $p < .05$, ** $p < .01$; older subjects:$^{+}$ $p < .05$). (A) First-order predictive sequence (FOP); (B) second-order predictive sequence (SOP). (Reprinted with permission from Curran, 1997a, Figure 3, Copyright 1997, Springer-Verlag, Germany.)

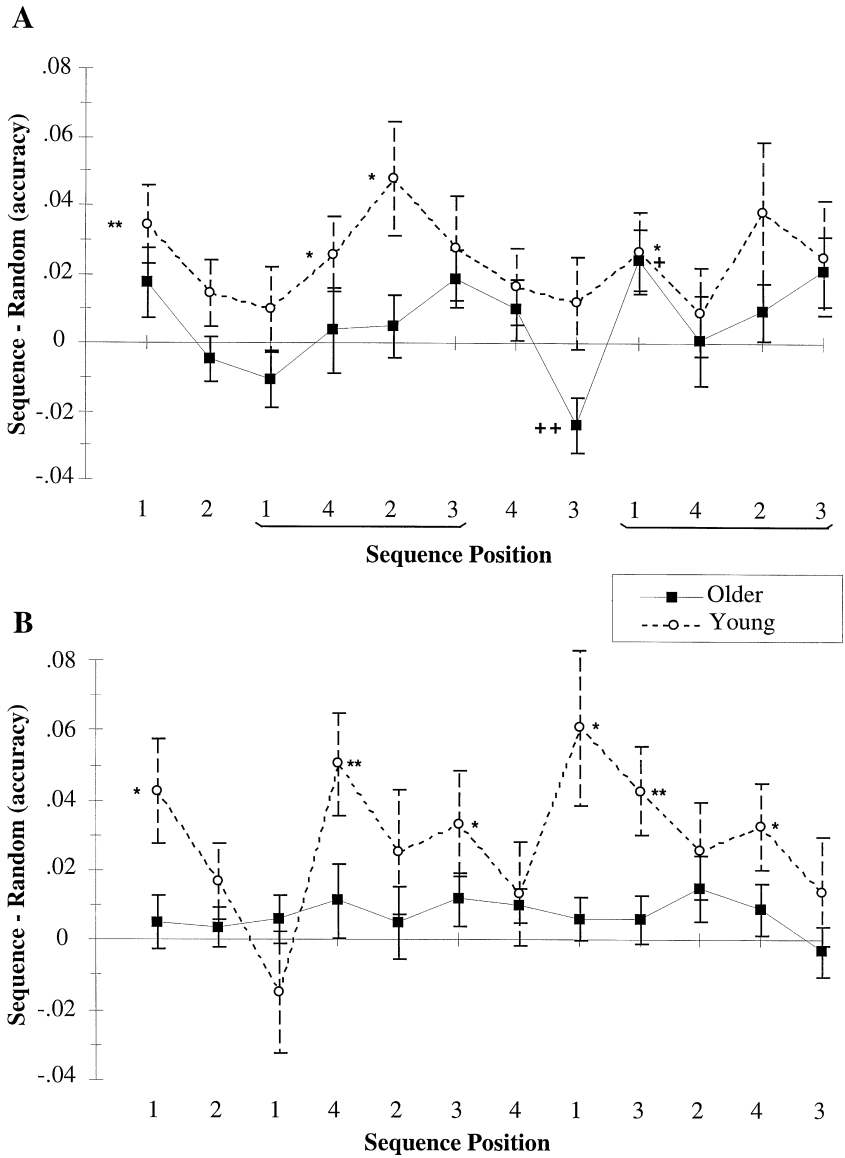


Fig. 6. Differences between accuracy to each sequence element and the same pairwise combinations in random trials (Experiment 2). Accuracy was averaged across all blocks of trials. Positive differences reflect learning of higher-order sequence information. Error bars show the standard error of the mean. *t*-tests (2-tailed) were used to assess the significance of the learning effect for each group and each sequence position separately (young subjects: * $p < .05$, ** $p < .01$; older subjects:$^{+}$ $p < .05$). (A) First-order predictive sequence (FOP); (B) second-order predictive sequence (SOP). (Reprinted with permission from Curran, 1997a, Figure 3, Copyright 1997, Springer-Verlag, Germany.)

stimuli following a rule that each of six locations could only be repeated once in each set of six elements (Boyer, Destrebecqz, & Cleeremans, 1998).

These lag effects may be related to well-studied sequential dependencies in random serial reaction time tasks (reviewed by, Kirby, 1980). In two-choice RT tasks, a ''repetition'' or ''positive recency'' effect is observed with short RSIs ($<$ 500 ms): responses to repetitions (A*A*) are faster than to alternations (B*A*). However, an ''alternation'' or ''negative recency'' effect is observed with longer RSIs: responses to repetitions (A*A*) are slower than to alternations (B*A*). The negative recency effect has been explained in terms of subjects' subjective expectancies (Kirby, 1980; Soetens, 1998). For example, in a serial anticipation task with subjects predicting which of two possible stimuli (A or B) will be presented next within a random sequence, the probability of choosing A decreases with the frequency of immediately prior occurrences of A (Jarvik, 1951). Such negative recency effects are observed when subjects are asked to generate their own random sequences because they believe that AAB is more likely than AAA (Wagenaar, 1972). In other words, subjects' performance in choice tasks is influenced by the gambler's fallacy (Jones, 1971; Kahneman & Tversky, 1972). Similarly, when subjects are asked to freely generate responses they tend to make all possible *n* responses on *n* successive trials (Rabinowitz, 1970). For example, with four response keys subject tend to press all four keys in each cycle of four consecutive trials.

The mean repetition lag was 3.0 for both the FOP and SOP sequences, so learning differences between FOP and SOP sequences shown by older subjects cannot be generally attributable to lag effects. A specific component of the lag effect that may be somewhat of a special case concerns the number of reversals (i.e., repetitions with a lag of 1; e.g., 1-2-1). The FOP sequence contained two reversals (1-2-1, 3-4-3) and the SOP sequence contained one (1-2-1). The higher reversal frequency could underlie the poorer FOP than SOP sequence learning shown by older adults. Moreover, by calculating the mean frequency of reversals per twelve elements (Reed & Johnson, 1994), it can be seen that reversals were more frequent in random trials (.28) than FOP (.167) or SOP (.083) sequences. Differences in reversal frequency, rather than sequence learning, could conceivably underlie performance differences between random and sequence conditions.

One particularly interesting feature of reversals concerns evidence that they can facilitate explicit sequence learning. Restle and colleagues studied explicit serial learning with the method of anticipation. Subjects pressed six response keys corresponding to the positions of six lights. They were told that the lights would follow a repeating sequence, and their accuracy at anticipating the position of the next light was the dependent measure. Restle and Brown (1970) studied the organization of sequences into chunks.

For example, the sequence 212343565 could be organized around runs, 2 1234 356 5; or reversals, 212 343 565 (called ''short trills'' by Restle & Brown, 1970). Such an ambiguous sequence could be contrasted with other sequences that were inherently organized by runs, 6543 5432 4321 (''run only''), or reversals, 565 343 232 121 (''reversal only''). Effects of organization on performance were inferred from systematic declines in anticipation errors within each chunk, and increased errors at the beginning of each new chunk. An ambiguously organized sequence was chunked into runs for subjects pretrained with a run-only sequence, but chunked by reversals for subjects pretrained with a reversal-only sequence (Experiment IV, Restle & Brown, 1970). Organization of an ambiguous sequence also was influenced by presenting an ambiguous pretraining sequence that was chunked into runs or reversals by inserting temporal pauses (Experiment V, Restle & Brown, 1970). Similar results were obtained when subjects completed a tracking task that is more similar to the SRT task, but subjects explicitly were told about the repeating pattern and allowed to observe it prior to the tracking task (Restle & Burnside, 1972). Restle and Brown (1970; Restle, 1970) argued that runs and reversals form subunits at the subordinate level of hierarchical representations (or ''structural trees'').

Similar patterns of organization have been inferred by measuring response latencies to prememorized sequences that are repeatedly tapped (Povel & Collard, 1982). RT is slowest at the beginning of each chunk and becomes faster across later elements within each chunk. Povel and Collard found that an identical sequence was organized by runs when initially presented to subjects as 321234, but organized by reversals when presented as 212343.

Serial list recall also benefits from reversals. Serial recall is generally worse for the second presentation of a repeated item than for the first presentation or nonrepeated items (the ''Ranschburg Effect,'' e.g., Crowder & Greene, 2000; Greene, in press). However, serial recall of repetitions only declines when at least two other items intervene between repetitions (e.g., BD*X*CV*X*FG). Repetitions are recalled better than other items when zero or one item intervenes (e.g., BD*X*C*X*VFG, Crowder, 1968).

The presence of reversals may also have facilitated recognition memory for sequence chunks in the present Experiment 2 (Curran, 1997a). Young and older subjects recognized the FOP sequence (containing two reversals) better than the SOP sequence (containing one reversal). Furthermore, young subjects recognized four-element FOP fragments that contained reversals more often than FOP fragments without reversals. In contrast to these patterns, previously presented SRT performance measures suggest

that reversals hindered implicit sequence learning. Thus, reversal frequency may doubly dissociate implicit and explicit learning.

III. Effects of Reversal Frequency on Implicit and Explicit Sequence Learning (Experiments 3 to 5)

Three experiments investigated the effects of reversal frequency on implicit and explicit learning in the SRT task. Two 12-element, SOP sequences were constructed that differed according to the number of reversals: either 1 (1R) or 3 reversals (3R). All subjects completed an implicit phase followed by an explicit phase with sequence type manipulated between subjects. The implicit phase included nine blocks with mixed random and sequence cycles, followed by recognition and recall tests. In the explicit phase, subjects were asked to try to learn the sequence during four blocks of sequential trials, and then the recall and recognition tests were readministered. It was predicted that implicit learning would be better for 1R than 3R sequences, but explicit learning would be better for 3R than 1R sequences.

A. General Method

Subjects completed a serial reaction time (SRT) task and were randomly assigned to receive a sequence with one reversal (1R) or three reversals (3R). Reversals were cases in which the same location was repeated with one other location intervening (e.g., 1-3-1, 4-2-4, etc.). To minimize idiosyncrasies associated with any one particular stimulus–response sequence, two different 1R and 3R sequences were constructed by assigning different stimulus locations to the sequence positions. The 1R sequences were 1-2-1-3-2-4-3-1-4-2-3-4 and 4-3-4-2-3-1-2-4-1-3-2-1. The 3R sequences were 4-3-4-2-1-4-1-3-2-3-1-2 and 1-2-1-3-4-1-4-2-3-2-4-3. All sequences included four occurrences of each location, and one occurrence of each possible pairwise transition (excluding immediate repetitions). Twelve-trial random cycles also contained each location four times and no immediate repetitions.

Subjects were told that the experiment was simply concerned with speed, accuracy, and stamina. The presence of a repeating sequence was not mentioned until after the implicit learning phase was completed. Subjects were told that the asterisk would move from box to box, and they should press the corresponding key as quickly and accurately as possible (using the first two fingers of each hand). Each asterisk remained on the screen until a response was given, and the next stimulus appeared 200 ms after each response. After each 120-trial block, subjects were given speed and accuracy feedback to increase motivation.

The first block was all random practice trials. The next nine blocks in the implicit phase contained intermixed cycles of random and sequence trials arranged as R-S-S-R-S-S-R-S-S-R (where R is twelve random trials and S is twelve sequence trials). Next, subjects completed four blocks with only the sequence in the explicit phase. The starting position of each sequence was randomly selected on each occurrence (both phases).

Explicit knowledge was assessed following both the implicit and the explicit phases. First, the computer presented two multiple choice questions.

1. While you were doing the task, did you think the location of the asterisk on each trial was
   a. random
   b. predictable
2. If you thought the location of the asterisk on each trial was predictable, in what way was it predictable?
   a. The asterisk appeared more often on one side than on the other.
   b. The asterisk appeared most often in one location.
   c. The asterisk appeared in a repeating pattern of locations.
   d. The location of the asterisk was not predictable.

Following the multiple-choice questions, the subject was told there was a repeating sequence and she or he completed recall and recognition tests.

In the recall task, the subject was asked to press the keys in the order of the repeating sequence. The asterisk followed the response locations. The subject was required to make 15 responses and was asked to guess if uncertain. Accuracy was scored as the largest consecutive recall series that matched any part of the sequence (e.g., Rauch et al., 1995; Reber & Squire, 1994).

In the fragment recognition test (Perruchet & Amorim, 1992) subjects were shown 24 different four-element subsequences (or fragments). Fragment stimuli and responses were the same as the SRT task except subjects were allowed to respond slowly. Each of the 12 sequence fragments started with one of the 12 possible pairwise transitions and ended in accordance with the sequence. Each of the 12 distractor (nonsequence) fragments also started with each of the 12 possible pairwise transitions and then ended with two locations that did not conform to the sequence. Construction of the distractor fragments was semi-random, but constrained so that each location appeared equally often in each fragment position. Thus, sequence and random fragments were matched on beginning pair frequency and frequency of each location at each fragment position. After responding to each fragment subjects answered the following question on a four-point scale: Was that mini sequence part of the repeating sequence? (1) Definitely Not; (2) Probably Not; (3) Probably Was; or (4) Definitely Was.

B. Methodological Differences between Experiments 3, 4, and 5.

Experiments 3 (N = 44 undergraduates) and 4 (N = 48) used asterisks appearing in one of four horizontally arranged boxes as stimuli. Experiment 5 (N = 48) used centrally presented digits (1, 2, 3, or 4) as stimuli. Experiments 4 and 5 controlled the number of reversals appearing within random cycles to match the number of reversals within sequence cycles (1R or 3R). Experiment 3 did not control the reversal frequency within random cycles. Thus, Experiments 3 and 4 differ according to the control of reversal frequency within random cycles. Experiments 4 (spatial) and 5 (digits) differ according to stimulus characteristics.

C. Results of Experiments 3, 4, and 5

1. *Explicit Knowledge*

Results from the explicit knowledge tests will be presented first because they influenced SRT analyses. The explicit knowledge results were qualitatively similar in all three experiments (see Table I). Both recall

TABLE I

Explicit Knowledge Measures from Experiments 3, 4, and 5[a]

			Implicit phase		Explicit phase	
Task	Measure	Experiment	1R	3R	1R	3R
Recognition	Sequence confidence	3	2.71	2.73	2.92	2.91
		4	2.65	2.87	2.75	2.77
		5	2.75	2.67	2.80	2.68
	Distractor confidence	3	2.63	2.65	1.84	2.32
		4	2.60	2.80	2.38	2.32
		5	2.71	2.67	2.12	2.32
	Sequence—distractor	3	0.08	0.08	1.08	0.59
		4	0.05	0.06	0.37	0.45
		5	0.04	0.00	0.68	0.36
Recall	Longest correct recall	3	5.00	4.23	7.64	6.64
		4	4.17	4.17	5.54	5.13
		5	4.50	4.21	7.25	6.71
	n with recall <6	3	15	20		
		4	21	21		
		5	20	20		

[a] Recognition confidence could range from 1 (lowest confidence) to 4 (highest confidence). Recall scores could range from 1 to 12.

and recognition were sensitive to explicit knowledge because scores were significantly higher following the explicit than implicit phase. No accuracy differences between sequence types were significant. Recall appeared to be more sensitive than recognition (replicating Rauch et al., 1995) because above chance performance following the implicit phase was clearly indicated by only the recall test (chance = 3.71, estimated by Rauch et al., 1995). To minimize the contribution of explicit knowledge to SRT performance, subsequent analyses excluded subjects with scores above 5 on the recall test following the implicit phase. This criterion was selected because it was within one standard deviation (Exp 1: SD = 1.77) of chance recall. The last section of Table I shows the numbers of subjects who met this criterion in each experiment. All subsequent Figures (7 to 11) and results were based upon only these subjects with implicit phase recall less than 6.

2. *SRT Performance*

Because SRT performance is influenced by the lag between repetitions, repetition lag was included in analyses by calculating the median RT for trials with lags of 1 (i.e., reversals), 2, 3, 4, or 5. Five was the maximum lag considered because it was the longest lag within the 1R sequence. The 3R sequence also included a lag of 6, but the mean lags were similar between sequences (1R = 2.92, 3R = 3.00). To obtain a sufficient number of random trials for each subject at each lag, the 9 blocks were divided into 3 larger blocks by combining blocks 1–3, 4–6, and 7–9.

RT decreased with increasing lag and increasing blocks in all three experiments. In Experiments 3 and 4, the RT advantage for sequence over random conditions increased across blocks, but in Experiment 5 a marginal difference between sequence and random RT ($p = .07$) did not interact with blocks. When the lags were considered separately, RT differences between sequence and random conditions were only observed for lags of 4 and 5 (see Figure 7). Sequence type interacted with random–sequence conditions and lag in Experiments 4 and 5, such that the 3R sequence showed a larger difference from random than the 1R sequence at lags 4 and 5. The primary difference between Experiments 4 and 5 was that the random–sequence differences at lags 4 and 5 were significant for 1R and 3R in Experiment 4, but only for 3R in Experiment 5.

Accuracy analyses in Experiments 3 and 4 revealed Sequence–random × Block interactions indicative of learning. Significant sequence–random differences in Experiment 5 did not increase across blocks. Three-way Random–sequence × Lag × Sequence type interactions were observed in all three experiments (see Figure 8). 1R accuracy was higher than

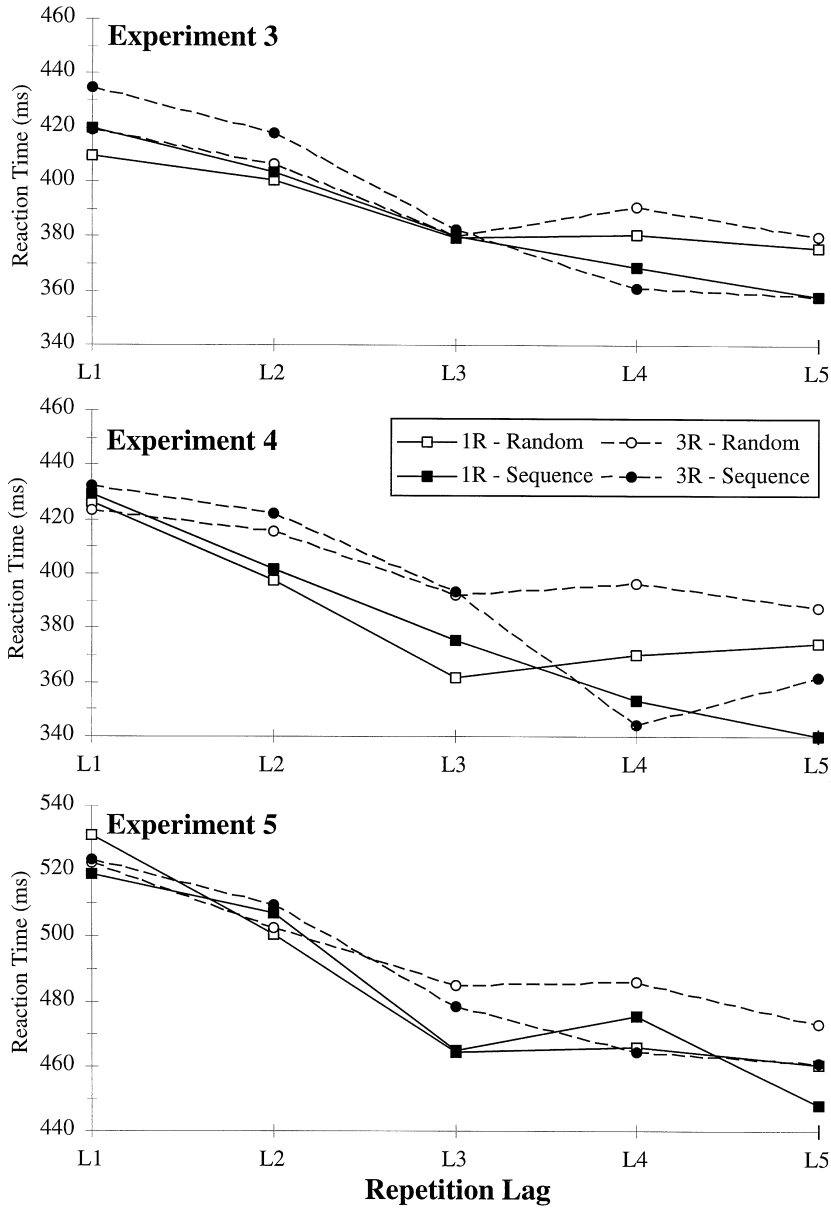


Fig. 7. Implicit phase reaction time (mean of subject's medians) for random and sequence conditions is plotted for each repetition lag from Experiments 3, 4, and 5.

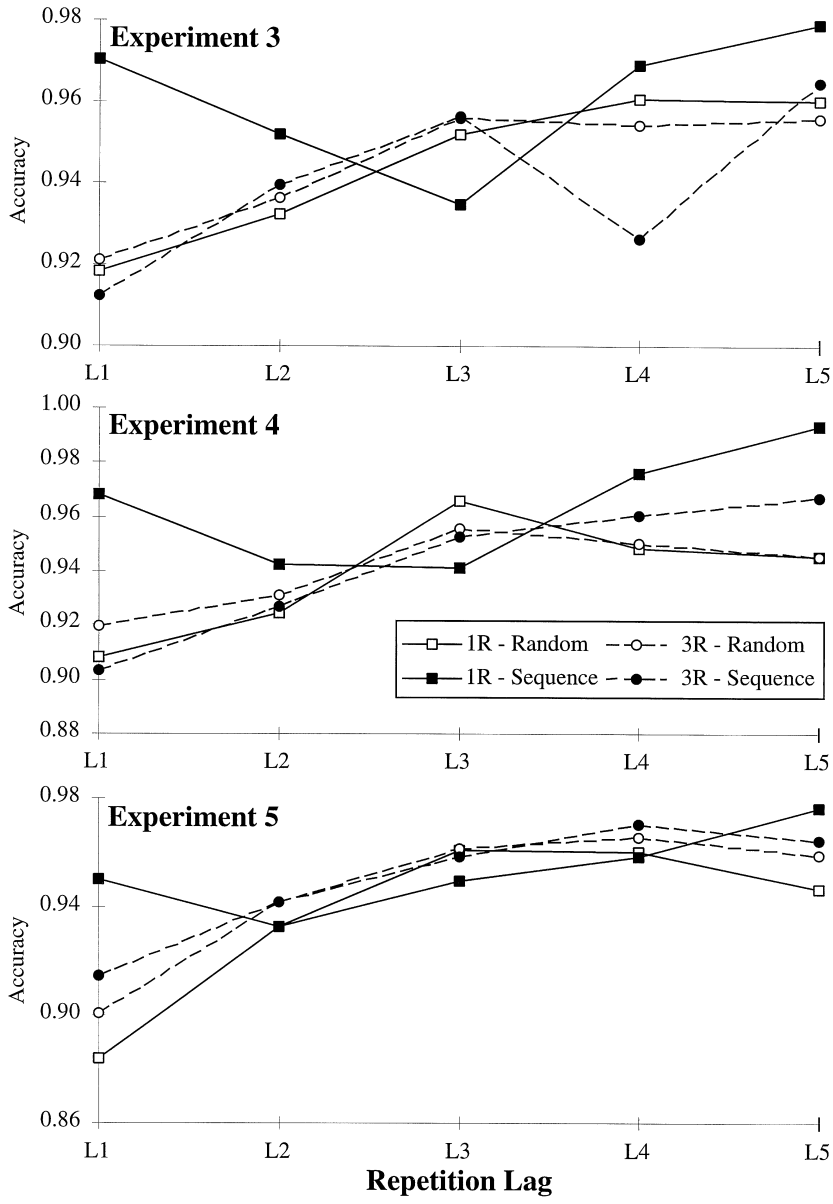


Fig. 8. Implicit phase accuracy (mean proportion correct) for random and sequence conditions is plotted for each repetition lag from Experiments 3, 4, and 5.

random only at lags of 1 and 5. 3R accuracy was not higher than random at any lag.

The explicit phase did not include random trials, so explicit learning was assessed by examining RT decreases across blocks. RT decreases across blocks typically are an imprecise measure of sequence learning because sequence-specific effects are confounded with nonspecific practice effects (e.g., learning of stimulus–response mappings). However, such nonspecific practice effects should primarily occur during the earlier implicit phase of the experiment, so they may not exert a large effect on speed-ups occurring within the explicit phase. Explicit RT decreased as blocks and lags increased (see Figure 9). There were no interactions involving sequence type or lag. Thus, explicit learning appeared across all lags and both sequence types. Explicit accuracy increased with lag, but did not reliably increase over blocks, so learning did not appear to affect accuracy.

3. Sequence Position Effects

Results were further analyzed to assess the pattern of learning across different positions of each sequence. The primary purpose of these analyses was comparing the patterns between implicit and explicit conditions. Median RTs were computed at each sequence position. For random trials median RTs were computed for each pair of locations that matched a particular place in the sequence. To obtain single estimates of random, implicit, and explicit RT at each sequence position, the medians for each block were combined into a single mean. Next, implicit-random and explicit-random difference scores were obtained for each sequence position to estimate the amount of higher-order learning at each position (as in Figures 2 and 5). Such difference scores are problematic when directly comparing implicit and explicit learning because the explicit differences are considerably larger than the implicit differences. Therefore, the qualitative patterns were compared by normalizing the RT differences (z-scores computed across the 12 sequence positions). The z-scores from each subject were entered into a Phase (implicit, explicit) $\times$ Sequence position (1 - 12) analysis of variance (ANOVA) for each sequence type separately.

Normalized RT differences are shown in Figures 10 (1R) and 11 (3R). Interactions between phase and sequence position indicated qualitatively different learning profiles between implicit and explicit phases. These interactions were significant in all three experiments with the 1R sequence, but only for Experiments 3 and 4 with the 3R sequence. At the most general level, implicit learning varied between positions more than explicit

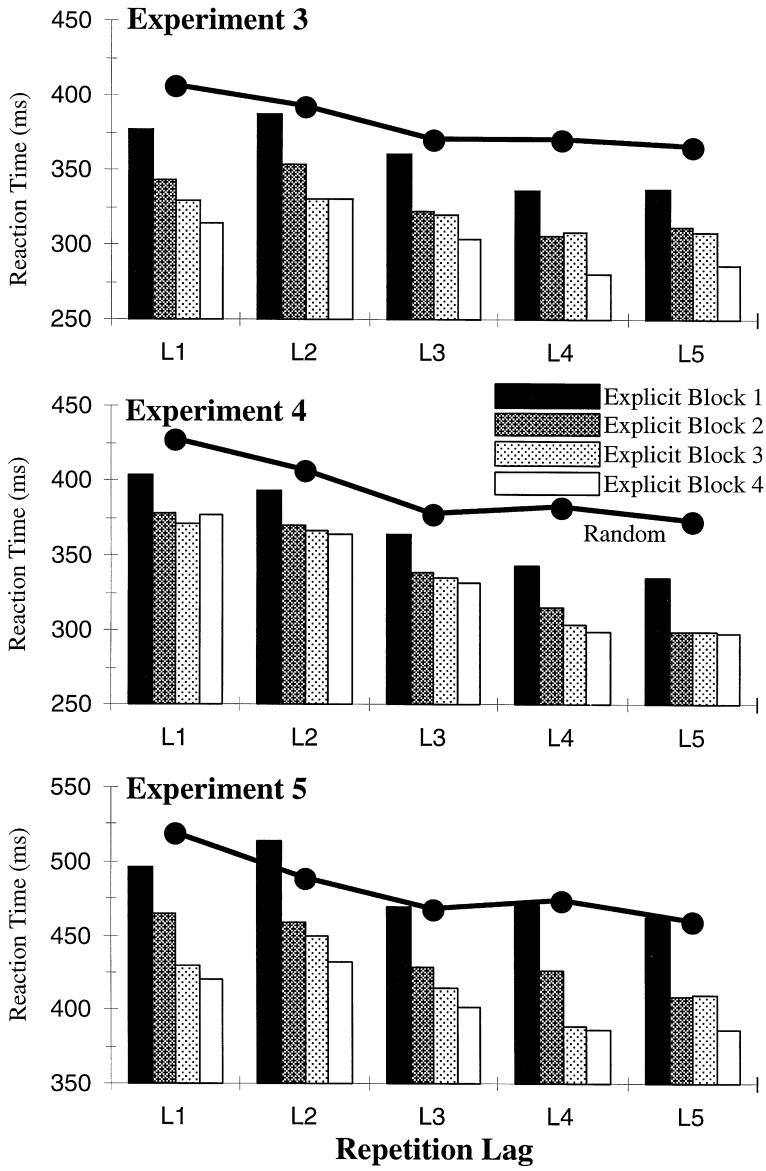


Fig. 9. Explicit phase reaction time (mean of subject's medians, averaged across sequence types) is plotted separately for each block x repetition lag combination (Experiments 3, 4, and 5). Lines with circular markers are random RTs from the last three blocks of the implicit phase.

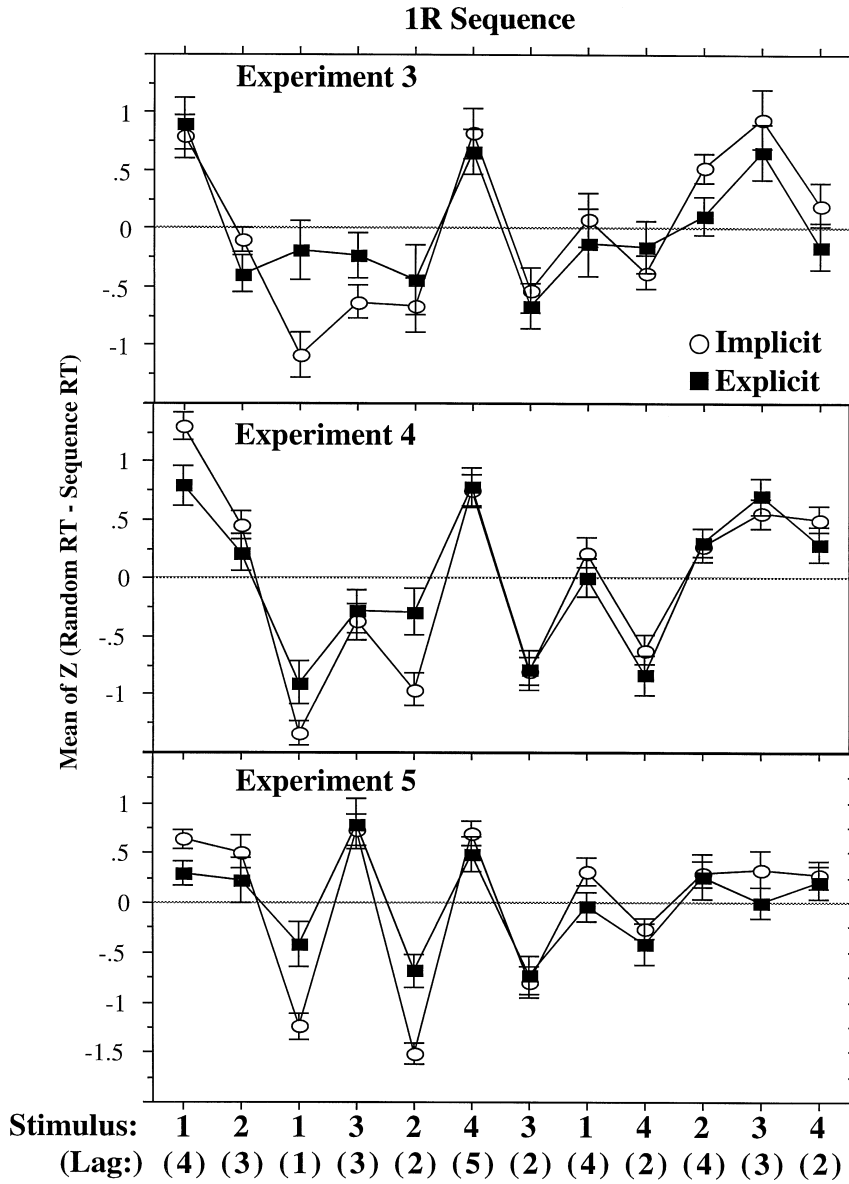


Fig. 10. Comparison of implicit and explicit learning profiles for the 1R sequence (Experiments 3, 4, and 5). Differences between reaction times to each sequence element and the same pairwise combinations in random trials were normalized (z-scores) across sequence positions. Larger differences reflect better learning relative to other positions. Error bars show the standard error of the mean.

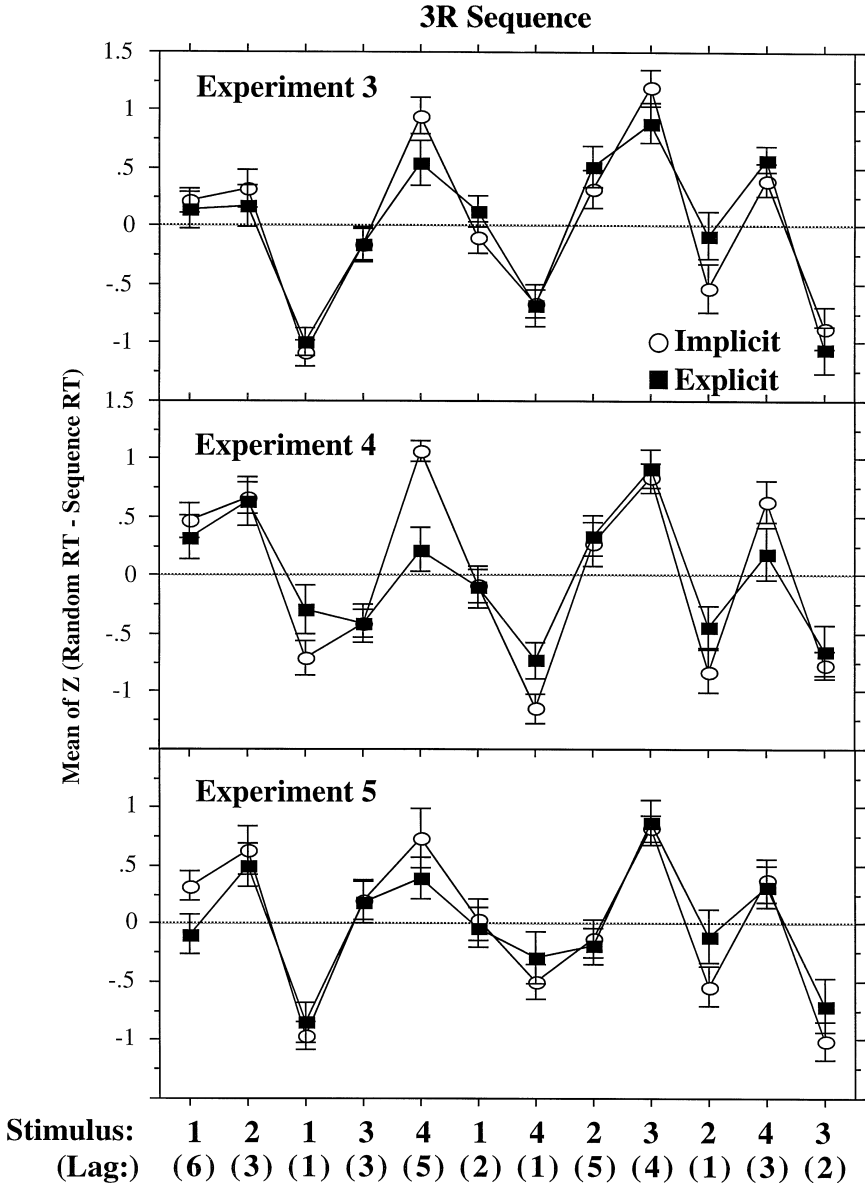


Fig. 11. Comparison of implicit and explicit learning profiles for the 3R sequence (Experiments 3, 4, and 5). Differences between reaction times to each sequence element and the same pairwise combinations in random trials were normalized (z-scores) across sequence positions. Larger differences reflect better learning relative to other positions. Error bars show the standard error of the mean.

learning—as indicated by the greater across-position range of implicit than explicit z-scores. In 1R, for example, the end of the 1-2-1 reversal was associated with more negative learning differences in the implicit than explicit phases. Reversals within the 3R sequence showed smaller differences between implicit and explicit z-scores (most apparent in Experiment 4). It is especially interesting that the 1-2-1 reversal clearly differentiated implicit and explicit learning in 1R, but not in 3R. Thus, the overall frequency of reversals (e.g., 1R vs. 3R) may influence the impact that individual reversals (e.g., 1-2-1) have on learning. The most consistent difference between implicit and explicit learning of the 3R sequence occurred at the fifth 3R position, where the z-score was higher in the implicit than explicit phases. This may be partially attributable to the long repetition lag of 5, but the other lag-5 position (position 8) did not show implicit/explicit differences. The 4-element chunk ending at the fifth position contains symmetric left (2-1) and right-hand (3-4) key presses that may facilitate learning.

Analogous ANOVAs with normalized accuracy differences primarily revealed no differences between implicit and explicit learning except in Experiment 4: For the tenth 3R position, implicit accuracy was higher in random than sequence conditions ($z = -1.05$), but explicit accuracy was similar between conditions ($z = 0.12$).

D. Summary of Experiments 3–5

Following the implicit learning phase of all experiments, subjects demonstrated above-chance explicit recall. Thus, implicit phase analyses focused on subjects with low recall (<6). RT analyses indicated that RT benefits for sequence over random trials were only observed at repetition lags of 4 and 5 (Figure 7). In Experiments 4 and 5, random-sequence differences at these lags were greater for the 3R than 1R sequence (with 1R differences not significant in Experiment 5). This random–sequence-by-1R/3R interaction was not significant in Experiment 3, in which reversal frequency of random and sequence conditions was not equated. Accuracy analyses showed a different pattern, with sequence–random differences only significant for the 1R sequence at lags of 1 and 5 (in all Experiments; Figure 8). Explicit learning was apparent from RT decreases at all lags, and explicit learning did not interact with sequence type (Figure 9). Learning of individual sequence positions showed very similar profiles in the implicit and explicit phases (Figures 10 and 11). Statistically significant differences generally seemed to reflect greater between-position variability in implicit than explicit learning.

IV. General Discussion

A. Amnesia and Aging

Experiments 1 (Curran, 1997b) and 2 (Curran, 1997a) manipulated the predictiveness of pairwise associations. College-aged subjects showed similar learning of sequences with (FOP) and without (SOP) predictive pairwise associations (Figure 3). Older subjects and amnesic patients learned the SOP sequence better than the FOP sequence (Figures 1 and 3) because the SOP sequence could be completely learned with second-order associations, but the FOP sequence could not. Analysis of higher-order learning effects at each sequence position suggested that amnesic subjects did not learn second-order associations as well as their control subjects (Figure 2). Older subjects appeared to learn second-order associations as well as young subjects, but older subjects had difficulty learning sequence positions that required greater-than-second-order associations (Figure 5).

Predominant theories of implicit sequence learning suggest that learning requires more than pairwise associations between adjacent sequence elements. It is generally thought that learning facilitates performance when some combination of previous sequence elements predicts the next element. Possible mechanisms included simple chunking processes (e.g., Frensch & Miner, 1994; Koch & Hoffman, 2000; Servan-Schreiber & Anderson, 1990; Stadler, 1995), contextually sensitive recurrent networks (Cleeremans, 1993; Cleeremans & McClelland, 1991), or recurrent networks that develop chunk-like representations (Dominey, 1998; Keele & Jennings, 1992). From this perspective differences between young subjects, older subjects, and amnesic subjects may reflect chunk size or range of contextual sensitivity. Previous research has shown that college-aged subjects show sensitivity to second-order (1-2 predicts 4, e.g., Jiménez, Méndez, & Cleeremans, 1996) and third-order (1-2-4 predicts 3, e.g., Cleeremans & McClelland, 1991) contingencies. The sensitive range is influenced by structural characteristics of sequences (Jiménez et al., 1996). Older subjects appear to learn second-order associations normally, but are impaired when second-order associations are insufficient. Amnesic subjects show impaired learning of second-order associations.

The pattern shown by older subjects, who learned second-order transitions normally but not higher-order transitions, is consistent with conclusions reached from an alternating SRT task (J. Howard & Howard, 1997). The alternating SRT task presents a sequence (e.g., 1-4-3-2) that alternates with random trials (e.g., 1-r-4-r-3-r-2). Subjects must learn at

least triplets (i.e., second-order associations) for performance on sequence trials to improve over the random trials. Howard and Howard found that older subjects appeared to learn second-order transitions as well as young subjects, but only younger subjects learned higher-than-second-order transitions.

The conclusion that amnesic subjects show less contextually sensitive learning has been supported recently with a much different experimental paradigm (Chun & Phelps, 1999). Chun and Phelps (1999) measured implicit learning in a visual search task. Both amnesic patients and control subjects became faster with practice in searching for a target (rotated T) among distractors (rotated Ls). However, only control subjects showed greater benefit from searching within repeated visual displays (learned contexts) compared with new visual displays (novel contexts). Thus, control subjects showed a contextually sensitive form of implicit learning that was absent in amnesic subjects.

Chun and Phelp's (1999) results were much clearer than those of Experiment 1 (Curran, 1997b), which revealed only subtle learning differences between amnesic patients and control subjects. Of the many procedural differences between these experiments, three features may have increased Chun and Phelps's sensitivity to amnesic deficits. First, Chun and Phelps tested five amnesics with known or likely hippocampal damage. Experiment 1 included 10 amnesics with more heterogeneous injury sites. If the contextually sensitive learning depends specifically on the hippocampus, Chun and Phelps's study was better aimed. Second, Chun and Phelps's subjects were younger (mean = 48 years, range: 41–55) than those in Experiment 1 (mean = 56 years, range: 31–73). Given the documented effects of aging on higher-order associative learning (Curran, 1997a; J. Howard & Howard, 1997), any deficits attributable to amnesia would be more difficult to detect in comparison with an older control group exhibiting age-related deficits. Third, in light of recent theoretical advances in understanding hippocampal function (O'Reilly & Rudy, 1999), Chun and Phelps's task may be more hippocampal-dependent than the SRT task.

Prominent theories of hippocampal function suggest that the hippocampus (possibly in conjunction with adjacent medial temporal cortex) constructs configural representations of the conjunction between multiple stimuli (e.g., N. Cohen & Eichenbaum, 1993; Gluck & Myers, 1993; Mayes et al., 1985; McClelland et al., 1995; Sutherland & Rudy, 1989; Wicklegren, 1979). From this perspective the hippocampus would be critical for learning higher-order sequences as well as contextually sensitive visual search tasks. However, a number of studies have indicated that rats with hippocampal damage can learn stimulus configurations (Alvarado & Rudy, 1995; Bunsey & Eichenbaum, 1996; McDonald et al.,

1997). O'Reilly and Rudy (1999) have suggested that the cortex as well as the hippocampus can accomplish configural learning, but there are other differences between cortical and hippocampal learning. The hippocampus is especially suited for rapid, incidental configural encoding, whereas cortical configural learning is more gradual and error-driven (O'Reilly & Rudy, 1999). From this perspective, learned contextual associations were more incidental to search task performance (Chun & Phelps, 1999) than to SRT performance (Experiment 1), so SRT learning may be less dependent upon the hippocampus.

SRT research with neurologically disordered patients and neuroimaging has primarily implicated the basal ganglia and motor cortical areas (for review see, Curran, 1995, 1997c; Keele, Ivry, Hazeltine, Mayr, & Heuer, 1998). The foregoing discussion has considered the possibility that the hippocampus and/or medial temporal cortex may contribute to implicit sequence learning because of the amnesic results presented in Experiment 1. However, this focus should not be misconstrued as a denial of the importance of other brain areas. More research is needed to better specify the possible functions that each of these brain mechanisms may contribute to sequence learning.

B. Lag Effects

Experiments 1 and 2 suggested that learning increased with the lag between stimulus repetitions. Reversals (e.g., 1-2-1, lag = 1) were especially difficult to learn. This pattern appeared to contrast with previous indications that reversals provide natural chunks that facilitate explicit learning/performance (Povel & Collard, 1982; Restle & Brown, 1970). Experiments 3, 4, and 5 tested the hypothesis that explicit learning would increases with reversal frequency, but implicit learning would decrease. The results contradicted this prediction. Explicit learning was not affected by reversal frequency, nor did it vary across lags. Implicit learning was affected by reversal frequency, but the influence varied across repetition lags and across dependent measures (accuracy or RT).

When RT was the dependent measure, learning increased with reversal frequency (3R > 1R) and was restricted to repetition lags of 4 and 5. When accuracy was the dependent measure, learning was only observed for 1R sequences at lags of 1 and 5. Interpretation of the differences between the dependent measures is uncertain, but it is notable that these differences were replicated across the experiments. Despite these uncertainties, the experiments support the following general points. First, repetition lag can greatly influence implicit learning. Repetition lag confounds could contribute to performance differences between random and sequence conditions

that are usually attributed to learning, so future research should control for these differences. Second, significant learning effects can be observed when analyses do control for repetition lag. Thus, mechanisms that are sensitive to only repetition lag are insufficient to account for SRT performance (e.g., Dominey, Ventre-Dominey, Broussolle, & Jeannerod, 1995). Third, the observed affects of reversal frequency and lag generalized across visuospatial and visual digit stimuli. Generalization to nonvisuospatial targets indicates that oculomotor effects like inhibition of return (Rafal & Henik, 1994; Taylor & Klein, 1998) are not responsible negative recency effects on learning.

Better understanding of lag effects will be important for evaluating models of sequence learning. Cleeremans's serial recurrent network (SRN) model learns the statistical structure of event sequence with contextually sensitive associations (Cleeremans, 1993; Cleeremans & McClelland, 1991). In its original application to the SRT task, the SRN was augmented with a decaying response-activation parameter that would prime recently active responses (Cleeremans & McClelland, 1991). Thus, activation of a given response would be facilitated if the response appeared at the end of a reversal (e.g., 1-2-*1*). Thus, the augmented SRN predicts a positive recency effect, but a negative recency effect was observed in the present experiments.

More recently, the SRN has been able to simulate negative recency effects (Boyer et al., 1998). Boyer et al.'s stimuli followed the rule that each of six locations could only be repeated once in each set of six trials, so long repetition lags were more frequent than short repetition lags. Empirically, a negative recency effect was observed in subjects' responses. Because the effect was present in initial trials and did not increase with training, it was interpreted as a pre-existing bias. The SRN (when not augmented with response priming) learned to simulate the negative recency effect, so Boyer et al. suggested that subjects may have pre-experimentally learned to prepare responses that have not been given recently. The present experiments also indicate that the lag effect was pre-existing (it was evident in the earliest blocks), but the lag effect interacted with the ability to learn the repeating sequences. It is conceivable that a common mechanism (like the SRN) is responsible for both the negative recency effect and for implicit sequence learning, and the interaction of these effects reflects the need to overcome the negative recency bias in order to learn repeating sequences. The RT results show that lag effects were strongest across the first three lags, but were weaker at the fourth and fifth lags that showed the strongest random–sequence learning differences. In other words, sequence learning was most evident at the lags with the least negative recency bias. This

explanation would be consistent with the RT data showing better sequence learning for long than short lags, but not with the accuracy data showing good learning of reversals in the 1R sequence.

A seemingly contradictory aspect of the RT results is that learning tended to be better for the sequence with the highest reversal frequency (3R > 1R), but random-sequence differences were confined to lags of 4 and 5.[1] Thus, reversals may impair learning locally, but may have some global benefit for learning other parts of the sequence. For example, even though subjects' ability to learn that 1-2 predicts 1 is poor, the 1-2-1 chunk may itself provide a parsing point or salient contextual marker for predicting other sequence elements.

Although implicit and explicit learning were differentially influenced by reversal frequency and lag, direct examination of the learning profiles obtained in Experiments 3 through 5 showed only subtle differences between implicit and explicit learning. Learning was especially poor for reversal positions (lag = 1). This contrasts with previous research showing performance benefits for reversal positions occurring within explicitly learned sequences (Povel & Collard, 1982; Restle & Brown, 1970). In conditions most similar to the present experiments, subjects were shown the sequence 2-1-2-3-4-3 as numbers on a computer screen (Povel & Collard, 1982). Next, the subjects practiced tapping the pattern until they could perform it from memory, and then performance was measured while the subject tapped the pattern six times consecutively. Only RTs from completely error-free trials were saved. Subjects were slowest at the first position of each reversal (*2*-1-2-*3*-4-3), indicating the sequence was chunked as 2-1-2 and 3-4-3. This contrasts with the present experiment in which the last reversal position was learned worse than the first two.

There are a number of procedural and analytic differences that could be responsible for these different patterns. First, Povel and Collard's profile was constructed from mean RTs, whereas the present profiles were constructed from sequence–random differences (to focus on learning rather than performance). However, the shapes of the present sequence RT profiles were very similar to the shapes of the RT difference profiles because most of the interposition variability in the differences arose from differences in sequence RT. Grand average RTs at each position were taken for each Experiment × Sequence type × Random/sequence condition. Standard

[1] These differences cannot be attributed to different lag 4 and lag 5 frequencies between 1R and 3R conditions because such frequency differences were not systematically related to learning. The lag-4 frequency was higher for 1R (.25) than random (.15), but the lag-5 frequency was similar for 1R (.08) than random (.09). The lag-4 frequency was lower for 3R (.08) than random (.13), but the lag-5 frequency was higher for 3R (.17) than random (.10). Random values were taken from Experiments 4 and 5.

deviations across the 12 sequence positions (range: 114 to 147 ms) were substantially larger than the corresponding random values (range: 10 to 23). Second, Povel and Collard's subjects were performing a perfectly learned sequence, whereas the present subjects did not completely master the sequence, even in explicit conditions (as documented by recall and recognition errors). There may be a trade-off between learning and performance such that reversal positions are difficult to learn (implicitly or explicitly); but, once learned (as in Povel & Collard, 1982), reversals may lead to performance benefits.

V. Conclusions

Sequence learning interacts with sequence structure. The present experiments have shown that the effects of amnesia, aging, and repetition lag on sequence learning cannot be clearly assessed without a consideration of sequence structure. Aging and amnesia may impair sequence learning by reducing the number of stimuli (chunk size or context) used to predict upcoming elements. Amnesia disrupts learning of second-order associations, and aging disrupts learning of higher-than-second-order associations. Sequence learning is affected by repetition lag such that learning-related RT improvement is confined to sequence positions with longer lags, but these lag effects are influenced by sequence structure. Explicit learning was uniform across all lags.

Future investigations of sequence learning should take these structural influences into consideration. Any conclusions arising from research with a single sequence may have limited generalizability. Furthermore, the present research has shown the usefulness of considering the microstructure of sequence learning rather than merely concentrating on overall performance differences between blocks of trials. Theories of sequence learning should be able to account for sequence position effects as well as the effects of repetition lag.

References

Alvardo, M. C., & Rudy, J. W. (1995). A comparison of kainic acid plus colchicine and ibotenic acid induced hippocampal formation damage on four configural tasks in rats. *Behavioral Neuroscience, 109,* 1052–1062.

Boyer, M., Destrebecqz, A., & Cleeremans, A. (1998). The serial reaction time task: Learning without knowing, or knowing without learning? *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 167–172). Hillsdale, NJ: Erlbaum.

Bunsey, M., & Eichenbaum, H. (1996). Conservation of hippocampal memory funtion in rats and humans. *Nature, 379,* 255.

Carlson, N. R. (1998). *Physiology of behavior* (6 ed.). Boston: Allyn and Bacon.

Chapman, L. J., Chapman, J. P., Curran, T., & Miller, M. B. (1994). Do children and the elderly show heightened semantic priming? How to answer the question. *Developmental Review. 14,* 159–185.

Cherry, K. E., & Stadler, M. A. (1995). Implicit learning of a nonverbal sequence in younger and older adults. *Psychology and Aging. 10,* 379–394.

Chun, M. M., & Phelps, E. A. (1999). Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nature Neuroscience, 2,* 844–847.

Cleeremans, A. (1993). *Mechanisms of implicit learning: Connectionist models of sequence processing.* Cambridge, MA: MIT Press.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General, 120,* 235–253.

Clegg, B. (1998). *The representation and transfer of sequence learning.* Unpublished doctoral dissertation, University of Oregon, Eugene.

Cohen, A., Ivry, R. I., & Keele, S. W. (1990). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 17–30.

Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system.* Cambridge, MA: MIT Press.

Crowder, R. G. (1968). Intraserial repetition effects in immediate memory. *Journal of Verbal Learning and Verbal Behavior. 7,* 446–451.

Crowder, R. G., & Greene, R. L. (2000). Serial learning: Cognition and behavior. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory.* New York: Oxford University Press.

Curran, T. (1995). On the neural mechanisms of sequence learning. [On-line] *Psyche, 2*(12). Available: http://psyche.cs.monash.edu.au/volume2-1/psyche-95-92-12-sequence-91-curran.html

Curran, T. (1997a). Effects of aging on implicit sequence learning: Accounting for sequence structure and explicit knowledge. *Psychological Research/Psychologische Forschung, 60,* 24–41.

Curran, T. (1997b). Higher-order associative learning in amnesia: Evidence from the serial reaction time task. *Journal of Cognitive Neuroscience, 9,* 522–533.

Curran, T. (1997c). Implicit sequence learning from a cognitive neuroscience perspective: What, how, and where? In M. A. Stadler & P. A. Frensch (Eds.), *Handbook of implicit learning* (pp. 365–400). Thousand Oaks, CA: Sage Publications.

Curran, T., & Keele, S. W. (1993). Attentional and nonattentional forms of sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 189–202.

Dominey, P. F. (1998). Influences of temporal organization on sequence learning and transfer: Comments on Stadler (1995) and Curran and Keele (1993). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 234–248.

Dominey, P. F., Ventre-Dominey, J., Broussolle, E., & Jeannerod, M. (1995). Analogical transfer in sequence learning. In J. Grafman, K. J. Holyoak, & F. Boller (Eds.), *Annals of the New York Academy of Sciences* (Vol. 769, pp. 369–373). New York: New York Academy of Sciences.

Frensch, P. A. (1997). One concept, multiple meanings: On how to define the concept of implicit learning. In M. A. Stadler & P. A. Frensch (Eds.), *Handbook of implicit learning* (pp. 47–104). Thousand Oaks, CA: Sage Publications.

Frensch, P. A., Buchner, A., & Lin, J. (1994). Implicit learning of unique and ambiguous serial transactions in the presence and absence of a distractor task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 567–584.

Frensch, P. A., & Miner, C. S. (1994). Effects of presentation rate and individual differences in short-term memory capacity on an indirect measure of serial learning. *Memory and Cognition, 22,* 95–110.

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus, 3,* 491–516.

Greene, R. L. (in press). Repetition effects in immediate memory in the absence of repetition. In H. L. Roediger, J. S. Nairne, I. Neath, & A. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder.* Washington, DC: American Psychological Association.

Hoffman, J., & Koch, I. (1997). Implicit learning of loosely defined structures. In M. A. Stadler & P. A. Frensch (Eds.), *Handbook of implicit learning* (pp. 161–199). Thousand Oaks, CA: Sage Publications.

Howard, D. V., & Howard, J. H. (1992). Adult age differences in the rate of learning serial patterns: Evidence from direct and indirect tests. *Psychology and Aging, 7,* 232–241.

Howard, D. V., & Wiggs, C. L. (1993). Aging and learning: Insights from implicit and explicit tests. In J. Cerella, J. Rybash, W. Hoyer, & M. L. Commons (Eds.), *Adult information processing: Limits on loss* (pp. 511–527). San Diego, CA: Academic Press.

Howard, J. H., Jr., & Howard, D. V. (1997). Age differences in implicit learning of higher-order dependencies in serial patterns. *Psychology and Aging, 12,* 634–656.

Howard, J. H., Mutter, S. A., & Howard, D. V. (1992). Serial pattern learning by event observation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1029–1039.

Hsiao, A. T., & Reber, A. S. (1997). The role of attention in implicit sequence learning: Exploring the limits of the cognitive unconscious. In M. A. Stadler & P. A. Frensch (Eds.), *Handbook of implicit learning* (pp. 471–494). Thousand Oaks, CA: Sage Publications.

Jackson, G., & Jackson, S. (1992). *Sequence structure and sequential learning: The evidence from ageing reconsidered.* (Technical Report 92-9). Institute of Cognitive and Decision Sciences, University of Oregon.

Jarvik, M. E. (1951). Probability learning and a negative recency effect in the serial anticipation of alternative symbols. *Journal of Experimental Psychology, 41,* 291–297.

Jiménez, L., Méndez, C., & Cleeremans, A. (1996). Comparing direct and indirect measures of sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 948–969.

Jones, M. R. (1971). From probability learning to sequential processing: A critical review. *Psychological Bulletin, 76,* 153–185.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430–454.

Kausler, D. H. (1994). *Learning and memory in normal aging.* San Diego: Academic Press.

Keele, S. W., Ivry, R. B., Hazeltine, E., Mayr, U., & Heuer, H. (1998). *The cognitive and neural architecture of sequence representation* (Technical Report 98-03 REVISED): Institute of Cognitive and Decision Sciences, University of Oregon.

Keele, S. W., & Jennings, P. J. (1992). Attention in the representation of sequence: Experiment and theory. *Human Movement Studies, 11,* 125–138.

Kirby, N. (1980). Sequential effects in choice reaction time. In A. T. Welford (Ed.), *Reaction times* (pp. 129–172). London: Academic Press.

Koch, I., & Hoffman, J. (2000). Patterns, chunks, and hierarchies in serial reaction-time tasks. *Psychological Research/Psychologische Forschung, 63,* 22–35.

Lashley, K. S. (1951). The problem of serial order. In L. A. Jeffress (Ed.), *Cerebral mechanisms of behavior* (pp. 112–136). New York: Wiley.

Light, L. L., & La Voie, D. (1993). Direct and indirect measures of memory in old age. In G. Graf & M. E. J. Masson (Eds.), *Implicit memory: New directions in cognition, development, and neuropsychology* (pp. 207–230). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mayes, A. R., Meudell, P. R., & Pickering, A. (1985). Is organic amnesia caused by a selective deficit in remembering contextual information? *Cortex, 21,* 167–202.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complimentary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457.

McDonald, R. J., Murphy, R. A., Guarraci, F. A., Gortler, J. R., White, N. M., & Baker, A. G. (1997). Systematic comparison of the effects of hippocampal and fornix-fimbria lesions on the acquisition of three configural discriminations. *Hippocampus. 7,* 371–388.

Mitchell, D. B. (1993). Implicit and explicit memory for pictures: Multiple views across the lifespan. In P. Graf & M. Masson (Eds.), *Implicit memory: New directions in cognition, development, and neuropsychology* (pp. 171–190). Hillsdale, NJ: Erlbaum.

Moscovitch, M., Vriezen, E., & Goshen-Gottstein, Y. (1993). Implicit tests of memory in patients with focal lesions or degenerative brain disorders. In H. Spinnler & F. Boller (Eds.), *Handbook of neuropsychology* (Vol. 8, pp. 133–173). Amsterdam: Elsevier.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures: *Cognitive Psychology, 19,* 1–32.

Nissen, M. J., Knopman, D. S., & Schacter, D. L. (1987). Neurochemical dissociation of memory systems. *Neurology, 37,* 789–794.

Nissen, M. J., Willingham, D., & Hartman, M. (1989). Explicit and implicit remembering: When is learning preserved in amnesia? *Neuropsychologia, 27,* 341–352.

O'Reilly, R. C., & Rudy, J. W. (1999). *Conjunctive representations in learning and memory: Principles of cortical and hippocampal function* (Technical Report 99–01). Boulder, CO: University of Colorado Institute of Cognitive Science.

Parkin, A. J. (1993). Implicit memory across the lifespan. In P. Graf & M. Masson (Eds.), *Implicit memory: New directions in cognition, development, and neuropsychology* (pp. 191–206). Hillsdale, NJ: Erlbaum.

Perruchet, P., & Amorim, M. (1992). Conscious knowledge and changes in performance in sequence learning: Evidence against dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 785–800.

Povel, D. J., & Collard, R. (1982). Structural factors in patterned finger tapping. *Acta Psychologica, 52,* 107–123.

Rabinowitz, F. M. (1970). Characteristic sequential dependencies in multiple-choice situations. *Psychological Bulletin, 74,* 141–148.

Rafal, R., & Henik, A. (1994). The neurology of inhibition: Integrating controlled and automatic processes. In D. Dagenbach & T. H. Carr (Eds.), *Inhibitory processes in attention, memory, and language* (pp. 1–51). San Diego: Academic Press.

Rauch, S. L., Savage, C. R., Brown, H. D., Curran, T., Alpert, N. M., Kendrick, A., Fischman, A. J., & Kosslyn, S. M. (1995). A PET investigation of implicit and explicit sequence learning. *Human Brain Mapping, 3,* 271–286.

Reber, P. J., & Squire, L. R. (1994). Parallel brain systems for learning with and without awareness. *Learning & Memory, 1,* 217–229.

Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 585–594.

Restle, F. (1970). Theory of serial pattern learning: Structural trees. *Psychological Review, 77,* 481–495.

Restle, F., & Brown, E. (1970). Organization of serial pattern learning, *The psychology of learning and motivation* (Vol. 4, pp. 249–331). New York: Academic Press.

Restle, F., & Burnside, B. L. (1972). Tracking of serial patterns. *Journal of Experimental Psychology, 118,* 219–235.

Rosenbaum, D. A., Kenny, S. B., & Derr, M. A. (1983). Hierarchical control of rapid movement sequences. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 86–102.

Schacter, D. L., Chiu, C. Y. P., & Ochsner, K. N. (1993). Implicit memory: A selective review. *Annual Review of Neuroscience, 16,* 159–182.

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 592–608.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17,* 367–447.

Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review, 70,* 534–546.

Simon, H. A., & Kotovsky, K. (1972). Complexity and the representation of patterned sequences of symbols. *Psychological Review, 79,* 369–382.

Soetens, E. (1998). Localizing sequential effects in serial choice reaction time with the information reduction procedure. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 547–568.

Squire, L. R., Knowlton, B., & Musen, G. (1993). The structure and organization of memory. *Annual Review of Psychology, 44,* 453–495.

Stadler, M. A. (1992). Statistical structure and implicit serial learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 318–327.

Stadler, M. A. (1993). Implicit serial learning: Questions inspired by Hebb (1961). *Memory & Cognition, 21,* 819–827.

Stadler, M. A. (1995). Role of attention in implicit learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 674–685.

Stadler, M. A., & Roediger, H. L. I. (1997). The question of awareness in research on implicit learning. In M. A. Stadler & P. A. Frensch (Eds.), *Handbook of implicit learning* (pp. 47–104). Thousand Oaks, CA: Sage Publications.

Sutherland, R. J., & Rudy, J. W. (1989). Configural association theory: The role of the hippocampal formation in learning, memory, and amnesia. *Psychobiology, 17,* 129–144.

Taylor, T. C., & Klein, R. M. (1998). On the causes and effects of inhibition of return. *Psychological Bulletin and Review, 5,* 625–643.

Wagenaar, W. A. (1972). Generation of random sequence by human subjects: A critical survey of literature. *Psychological Bulletin. 77,* 65–72.

Wicklegren, W. A. (1979). Chunking and consolidation: A theoretical synthesis of semantic networks, configuring in conditioning, S-R versus cognitive learning, normal forgetting, the amnesic syndrome, and the hippocampal arousal system. *Psychological Review, 86,* 44–60.

RECALL PROCESSES IN RECOGNITION MEMORY

Caren M. Rotello

I. Introduction

In recent years, the memory literature and the popular press have been replete with examples of people ''remembering'' events that never occurred. These false memories are sometimes highly elaborated confabulations (e.g., Loftus, 1997), but are frequently just confidently made yet erroneous claims that a person had been met before, or that an event or even a single word had been experienced earlier in some context. How could something as simple as a recognition decision (e.g., ''yes, that word was presented on the list I studied a moment ago'') be so wrong? The processes that contribute to a recognition judgment are quite a bit more varied and more complex than one might expect, as will be described here.

Deciding whether you have seen something before in a given context, or recognition, is often thought of as a judgment process in which the stimulus sends a probe to memory and returns an ''echo'' that reflects the match of the stimulus with the contents of memory. If the strength of that similarity- or familiarity-based response from memory exceeds some decision criterion, then the stimulus is judged to be ''old'' (i.e., it occurred in that context); the stimulus is judged to be ''new'' if the memory strength falls below that criterion. For example, recognition of this sort would allow you to ''know'' that you have previously met someone because she seems very familiar, even if you cannot remember how you know happen to know her.

THE PSYCHOLOGY OF LEARNING
AND MOTIVATION, VOL. 40

Copyright © 2001 by Academic Press.
All rights of reproduction in any form reserved.
0079-7421/00 $35.00

A growing body of research, however, suggests that these memory judgments are more complex, often involving a second, recall-like, process (e.g., Atkinson & Juola, 1973, 1974; Dosher, 1984; Gronlund & Ratcliff, 1989; Hintzman & Curran, 1994; Jacoby & Dallas, 1981; Mandler, 1980; Ratcliff & McKoon, 1982, 1989; Rotello & Heit, in press; see Mandler, 1991, for a brief review). Recognition of this sort might allow you to recognize a person because you remember meeting him at a party last week, or it might allow you to realize that this familiar-seeming person is actually a stranger who merely reminds you of someone else.

The recall-like process in recognition is generally thought to operate more slowly than a familiarity-based process. An early version of such a dual-process theory of recognition memory was proposed by Atkinson and Juola (1973, 1974). They argued that recognition decisions are based on the familiarity of the target item if that familiarity is sufficiently weak or sufficiently strong to "provide the subject with enough information to respond with confidence" (1973, p. 609). If the familiarity of the item cannot support a decision, then a secondary process is used. In their view, the second process is an exhaustive search of the studied items and always produces a correct response for well-learned items. For larger sets of study items, or less well-learned stimuli, one might assume that the search process is neither exhaustive nor infallible.

II. Evidence on the Use of Recall in Recognition

Currently, there are many diverse lines of evidence for a secondary process that suggest it is a recall-like process, ranging from data collected in the controversial *process dissociation* and *remember/know* procedures to data collected in *response signal* and *evoked response potential* (ERP) experiments. I will describe a sampling of the data that tell the clearest story about the secondary process in recognition as a retrieval-based or recollective process.

First, in a classic study on associative recognition in which subjects studied A–B pairs, Humphreys (1978) found that the probability that A–B would be recognized could be predicted from the probability that B was recognized alone and the probability that, if B was not recognized, it could be recalled given A as a retrieval cue. His data strongly implicate a recall component in at least some kinds of recognition judgments. More indirect evidence for the use of recall in associative recognition has been provided by Clark and his colleagues (Clark & Hori, 1995; Clark, Hori, & Callan, 1993). In their experiments, subjects were presented with pairs of words (e.g., A–B, C–D, E–F) to study. Later, they were asked to identify the studied pair

of words among a set of three test pairs. Some of the test triads were NOLAP trials, in which all of the test words were unique (e.g., A–B, C–F, G–I), whereas other triads were OLAP trials, in which one of the words appeared in all three pairs (e.g., A–B, A–D, A–F). After short study lists, recognition was better on the NOLAP trials than on the OLAP trials; with long study lists recognition was equivalent in the two conditions. Clark et al. (1993) argued that the NOLAP trials provided more unique retrieval cues than the OLAP trials (i.e., six unique words rather than four), which could be used in a covert cued recall test to evaluate the test triad. With long study lists, such a cued recall strategy would be rendered ineffective, resulting in a loss of the NOLAP advantage, exactly as occurred (Clark & Hori, 1995).

A second line of evidence for the use of a recall process in recognition memory comes from recent work on neuroimaging and evoked response potentials (ERPs; see Johnson, 1995; Rugg, 1995, for reviews). For example, Curran (1999, in press) has used ERPs to identify unique brain activity markers of a familiarity-based process and of a recollective contribution to recognition of words that differ in plurality from their studied form (e.g., *frog* and *trucks* are studied; *frogs* and *truck* appear on the test list). These markers differ in both their temporal and spatial characteristics. The familiarity process, which allowed discrimination of studied words from completely new words, emerged in the frontal recording regions and somewhat earlier in time than the recollective process. In contrast, the recollective process, which allowed discrimination of studied words (e.g., *frog*) from plurality-changed foils (e.g., *frogs*), appeared in parietal recording sites somewhat later in time. In addition, subjective reports of ''remembering'' the presentation of a stimulus, rather than merely ''knowing'' that it was studied have also been associated with larger effects in parietal sites (e.g., Düzel, Yonelinas, Mangun, Heinze, & Tulving, 1997; Rugg, Schloerscheidt, & Mark, 1998).

A third line of evidence for the use of a recall-like process in recognition memory comes from predictions about the shape of receiver operating characteristic (ROC) curve. A static single-process view of recognition leads to the predictions that recognition ROCs should be curvilinear (and symmetric if the underlying distributions of familiarity are normally distributed and of equal variance), whereas a dual-process view predicts that the ROCs will be asymmetric and increasingly linear as the relative importance of the recall component increases (Yonelinas, 1994). A variety of experiments on both normal and amnesic populations have produced data consistent with the dual-process predictions (e.g., Yonelinas, 1994, 1997; Yonelinas, Kroll, Dobbins, Lazzara, & Knight, 1998). The ROC data will be elaborated upon later.

Finally, the literature on the time-course of recognition decisions also suggests that a faster memory process assesses global aspects of the test stimulus, whereas a slower memory process is affected by the finer details of the stimulus. This has been demonstrated in a number of experiments using a response-signal procedure. For example, Hintzman and Curran (1994; for related work see Dosher, 1984; Ratcliff & McKoon, 1982, 1989) showed that when subjects study nouns that are singular (e.g., *frog*) or plural, they make false alarms to distractor nouns that differ in plurality from the studied words (e.g., *frogs*) during about the first 700 ms of processing. With additional processing time, however, subjects correctly reject *frogs.* In contrast, subjects are able to correctly reject new items that are not similar to presented stimuli. Other experiments have indicated that discrimination between studied and unstudied items that involve retrieval of the specific study episode also requires more processing time. For example, Gronlund, Edwards, and Orht (1997) found that subjects could discriminate between studied and unstudied words more quickly than they could judge whether a studied word had appeared in a particular spatial position (e.g., as the right-hand word in a study-triple). Similarly, Johnson, Kounios, and Reeder (1994), and Hintzman and Caulton (1997) reported that subjects were able to recognize studied items more quickly than they could identify the source of those items (i.e., whether it had been seen or heard). These studies suggest that the familiarity of a stimulus is used as the first basis of response, and that a recall-like process is used for slower or more careful recognition decisions.

III. Evidence on the Specific Nature of the Recall Process

These lines of evidence point to at least two basic mechanisms that underlie recognition decisions: familiarity and recollection. On the assumption that recall contributes to recognition, how does the recall process operate? Rotello and Heit (1999) described a number of possible recollective processes that might contribute to a recognition judgment (see also, Brainerd, Wright, & Mojardin, 1999). For example, a *recall-to-accept* process would facilitate the acceptance of studied items by retrieving a memory trace that matched the studied test probe. By definition, such a process could not affect recognition judgments of distractors, unless the test foil were quite similar to the studied item and the dissimilarities were undetected. An *exhaustive search* process would search through memory for an exact match for the test probe, resulting in a positive judgment if a match is found and a negative decision if no match is found. This type of search process is consistent with several metamemory theories (e.g., Brown, Lewis, & Monk,

1977; Strack & Bless, 1994), but has not been supported empirically (Rotello, 1999; Wixted, 1992).

Despite these alternative possibilities, most researchers have assumed that the recollective component of both item and associative recognition operates as a *recall-to-reject* process in which mismatching information is retrieved from memory and used to reject a test probe (e.g., Clark, 1992, p. 241; Clark & Burchett, 1994, p. 57; Gronlund & Ratcliff, 1989, p. 857; Hintzman & Curran, 1994, p. 14; see Clark & Gronlund, 1996, pp. 56-57 for a review). As a consequence of this assumption in the literature, the majority of the work reviewed here is aimed specifically at evaluating the use of a recall-to-reject process.

Rotello and Heit (1999) described one of the clearest predictions of a recall-to-reject process for recognition performance. Suppose *frog* has been presented on a study list, but *frogs* has not. On the recognition test, subjects are sometimes given very little time in which to make their decision (e.g., less than 700 ms) and on other trials they are given a longer decision period (e.g., 1200 ms or more). Early in processing, familiarity-based responding will cause the ''old'' responses to both *frog* and *frogs* to increase as the familiarity of the item increases, whereas ''old'' responses to unrelated items like *chair* will decline. Consequently, the difference in false alarms to similar (*frogs*) and new foils (*chair*) increases early in processing, when familiarity is the primary basis of response. Later in processing, the recall-to-reject account makes specific predictions when a similar foil such as *frogs* is tested. In that case, the old item, *frog,* is recalled, but the similar item, *frogs,* cannot be recalled. In the face of this negative or mismatching evidence, *frogs* is rejected (called ''new'') despite its high degree of familiarity. The more time that is allowed for the recall process to operate, the more likely it is that a mismatching memory trace will be retrieved to facilitate the rejection of the similar foil. Hence, false alarms to similar foils like *frogs* should decline late in processing, once the recall process has begun to contribute to the judgments (see the upper panel of Figure 1). A recall-to-reject process does not affect recognition of dissimilar items such as *chair* that are not similar to any presented item; those completely new foils would presumably be rejected by the familiarity-based process. Thus, if a recall-to-reject process is used, the difference in the false alarm rates to *frogs* and to *chair* should decrease late in processing, as the second process helps people to reject *frogs* (but does not affect judgments of *chair*). Like the recognition of completely new words, recognition of studied items also would not be influenced by a recall-to-reject process, so subjects' ability to discriminate between studied words and completely new foils should not change late in processing (unless they continue to be influenced by the familiarity-based process).

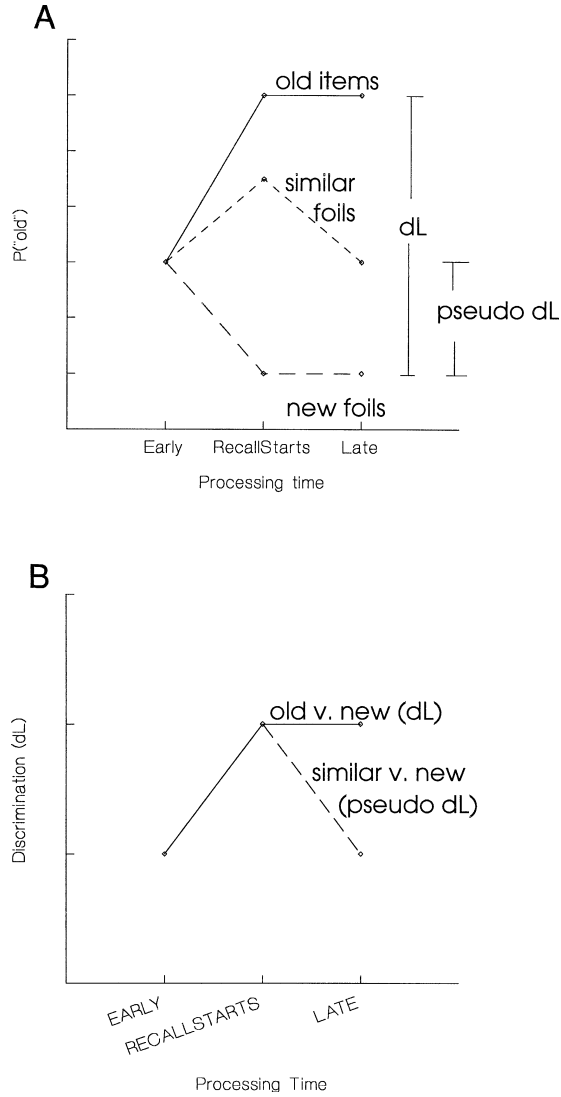


Fig. 1. (A) Hypothetical probability (P) of "old" judgments for studied items, similar foils, and new foils across the time course of judgment. (B) Idealized predictions of the recall-to-reject account in terms of discrimination (d_L). In each case, the "early" stage of processing is assumed to be determined by a familiarity-based process; a recall-to-reject process begins to influence the judgments at the point marked "recall starts" and reaches asymptotic levels "late" in processing.

It is essential that the positive (''old'') responses to studied items, similar foils, and completely new foils are evaluated in *relative* terms. Although it is tempting to simply look at the changes in the rates of positive responding to similar foils across processing time, doing so would fail to consider the possible role of changing response bias. Because subjects may have either very little or a longer period of time in which to make their recognition decision, it seems quite plausible that response bias would change over processing time. Indeed, some preliminary studies of response bias in the response-signal paradigm have revealed increasingly conservative bias over signal lags, when sensitivity (d_L) is held constant (E. Heit, personal communication, November 19, 1999).[1] Therefore, a discrimination measure such as d' or d_L, which is independent of response bias, is a better measure of performance across processing time. In essence, these discrimination scores indicate the standardized distance between the means of two normal (d') or logistic (d_L) distributions of memory strength; larger values of d' or d'_L indicate greater ability of the subjects to discriminate between the two classes of stimuli. The hit and false alarm rate predictions shown in Figure 1A can be translated into hypothetical discrimination predictions, as is shown in Figure 1B. Both the positive responses to studied items (i.e., hits) and to similar foils (i.e., false alarms) are scaled against the positive responses to completely new items. This scaling process yields two discrimination functions: the d_L curve predicts performance on studied items compared to new foils; critically, the *pseudo-*d_L curve predicts performance on similar foils as compared to new foils (e.g., Dosher, 1984). The pseudo-d_L function must be nonmonotonic across time if familiarity and a recall-to-reject process are used: False alarms to similar foils, compared to new foils, increase early in judgment because of the familiarity-based process, but the late-acting recall-to-reject process results in a decreasing false alarm rate to similar foils as compared to new foils.

Rotello and Heit (1999) tested the recall-to-reject account of item recognition in a response-signal paradigm and found no support for such a mechanism. We presented four experiments (including reanalyses of Hintzman & Curran's, 1994, Exps. 2 & 3) in which subjects studied lists of words (e.g., *frog*) or pseudowords (e.g., *PRUMID*) that were shown once or twice each. Subjects were then asked to make recognition judgments of studied items (e.g., *frog, PRUMID*), new items that were similar to studied items (e.g., *frogs, PRAMAR*), and completely new foils that were dissimilar to all studied materials. We computed d_L for the discrimination between studied and new stimuli and between similar foils and completely new foils.

[1] Hintzman and Curran (1997) also reported a measure of bias at each lag in a response signal experiment. However, discrimination (d_L) varied significantly across signal lags, rendering their bias measure difficult to interpret (see Macmillan & Creelman, 1991).

Discrimination data from the second experiment are shown in Figure 2. Those data show clear monotonically increasing d_L and pseudo-d_L functions.

To determine whether there was any significant drop in acceptance of the similar foils at the later response signal lags, the data were fit to two separate functions. The first function represents monotonic growth to a limit (Ratcliff, 1978):

$$d_L(lag) = \frac{\lambda_1}{\sqrt{1 + \nu / (lag - \delta)}}, \tag{1}$$

where *lag* is the response signal lag plus the latency to respond at that lag (i.e., total processing time), λ_1 is the asymptotic level of discrimination, μ is the rate of approach to that asymptote, and δ is the point at which discrimination rises above chance. The second function we fit to our data was a nonmonotonic function of a similar form. It assumes that the data

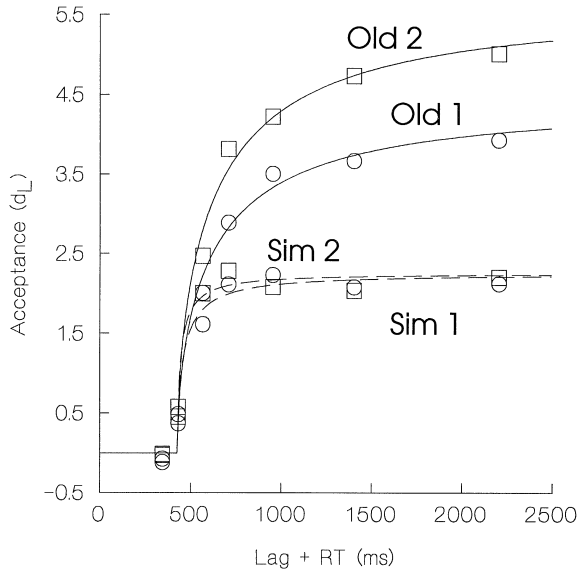


Fig. 2. Positive responses to words studied once (old 1) or twice (old 2) and to foils similar to words studied once (sim 1) or twice (sim 2) scaled against positive responses to completely new foils, from Rotello and Heit (1999). RT, reaction time; d_L, discrimination. (From "Two-process models of recognition memory: Evidence for recall-to-reject?" by C. M. Rotello and E. Heit, 1999, *Journal of Memory and Language, 40,* p. 437. Copyright 1999 by Academic Press. Reprinted with permission.)

conform to Equation 1 until a particular point in processing time, *lag**, at which point discrimination begins to decline to a second (lower) asymptote, λ_2. The nonmonotonic function is given by

$$d_L(lag) = \frac{\lambda_2 + (\lambda_1 - \lambda_2)(lag^* - \delta)/(lag - \delta)}{\sqrt{1 + \nu / (lag - \delta)}} \text{ for lag} \geq \text{lag}^*. \qquad (2)$$

Equations 1 and 2 (and variants) have been widely used (e.g., Dosher, 1984; Gronlund & Ratcliff, 1989; Ratcliff, 1980; Rotello & Heit, 1999, in press; Wickelgren & Corbett, 1977).

For each type of stimulus, distinct rates and asymptotes were estimated, but a single intercept (δ) was used for all the stimuli, and likewise a single *lag** was used for the nonmontonic fits. (Rotello & Heit, in press, found that constraining the intercepts to all be equal did not change the conclusions about the monotonicity or nonmonotonicity of any set of data.) Model fitting was performed using SYSTAT's Simplex estimation method to iteratively adjust parameters so as to minimize the residual sum of squares (RSS). The relative degree of fit of these nested models was evaluated using a technique suggested by Borowiak (1989). He showed that when model A is a nonlinear model with a free parameters estimated using a least-squares criterion, and B is a restricted version of this model with b free parameters, the likelihood ratio statistic is $\lambda = (\text{RSS}_\text{A}/\text{RSS}_\text{B})^{(k/2)}$, where RSS is the residual sum of squares of the model and k is the number of data points to be predicted. Borowiak showed that $-2 \ln (\lambda)$ has a χ^2 distribution with $(a - b)$ degrees of freedom.

Across four experiments conducted in three different laboratories with different subjects, materials, and response signal lags, the consistent finding at both the group and individual levels of analysis was that the nonmonotonic fit was not significantly better than the monotonic fit (the curves in Figure 2 show the best monotonic fit to the group data). Subjects were not increasingly able to discriminate foils that were similar to studied items from completely new distractors. Although the data were inconsistent with the recall-to-reject account of item recognition, they were consistent with the predictions of both a recall-to-accept and an exhaustive search process. In a recall-to-accept process, information that is retrieved from memory (e.g., *frog*) that matches the test probe (*frog*) facilitates acceptance of the probe, over and above its familiarity level. Because no matching information can be retrieved for either a similar foil (*frogs*) or for a completely new word, a recall-to-accept process does not affect the recognition of distractors. Consequently, use of a recall-to-accept process is evidenced by both an increasing ability of participants to discriminate studied test items from new foils across

time, and by a lack of improvement in their ability to reject similar foils relative to completely new foils. So, d_L for old–new discrimination continues to rise late in processing, when the recall process operates, but d_L for similar–new discrimination is constant over the later stages of recognition, because the recall-to-accept process does not assist recognition of either the similar foils or the completely new foils

Despite this failure to find evidence for recall-to-reject processing in simple item recognition tasks, most of us have the strong intuition that we do use recall-to-reject processing in some situations. Indeed, when an item recognition task is made slightly more difficult, by having items appear on one of two different study lists and asking subjects to recognize words from List 1 (say) and to reject lures from List 2 as well as new words, then clear evidence for the use of recall-to-reject emerges (Rotello & Heit, in press, Exp. 3). Other types of source discrimination tasks (e.g., ''seen'' vs. ''heard'') also lead to the use of recall-to-reject processing (McElree, Dolan, & Jacoby, 1999), although those data were not intended to test the specific recall-to-reject predictions.

Several theoretical accounts of associative recognition, in which subjects study and are later tested on pairs of words, assume that a recall-to-reject process is involved (e.g., Humphreys, 1978; Mandler, 1980). Moreover, recall is more likely to operate in associative recognition (e.g., Clark & Gronlund, 1996; Hockley & Consoli, 1999). So, Rotello and Heit (in press) extended their analysis of recall-to-reject processing into the domain of associative recognition. In those experiments, subjects studied word pairs (e.g., *frog–quilt; mug–dress*) and were then asked to decide whether the words in each test pair had been studied together (i.e., in the same pair) on the presentation list. Subjects were tested on intact pairs (e.g., *frog–quilt*), rearranged pairs in which both words had been studied but not in the same pair (e.g., *frog–dress*), and completely new pairs in which neither word had been studied. The critical test probes were the rearranged pairs, which evoke strong familiarity-based responses from memory despite requiring a negative response from the subject. In contrast to the item recognition data, those data collected across the time course of judgment indicate that associative recognition *does* involve a recall-to-reject process: Although the fit of the monotonic function (Equation 1) was good (R^2 adjusted for number of free parameters = .986, RSS = .981), the nonmonotonic fit (Equation 2) was significantly better (R^2 adjusted = .990, RSS = .515; $\chi^2(2) = 7.733$, $p < .05$). See Figure 3 for representative data from Rotello and Heit's (in press) Experiment 1. The conclusions were the same when individual subjects' responses were evaluated.

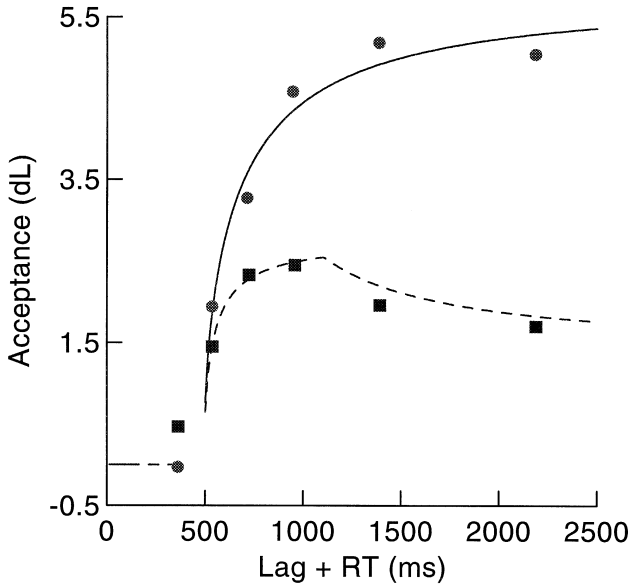


Fig. 3. Positive responses to intact and rearranged pairs, scaled against completely new pairs, from Rotello and Heit (in press, Exp. 1). RT, reaction time; d_L, discrimination. (From "Associative recognition: a case of recall-to-reject processing," by C. M. Rotello and E. Heit, in press, *Memory and Cognition.* Copyright by Psychonomic Society Publications. Reprinted with permission.)

IV. Factors That Influence the Use of Recall Processes in Recognition

So far, we have seen evidence that a recall-to-reject process is used in the recognition of associative and contextual information, but that a recall-to-accept or an exhaustive search process is the best description of the recollective process in simple item recognition tasks. What influences the use of a recall process in recognition? At the most basic level, the same factors that influence free recall have been shown to affect the contribution of a recollective component to recognition. For example, increasing the length of the study list (Clark & Hori, 1995; Clark et al., 1993; Yonelinas, 1997) and increasing the retention interval (Brainerd, Reyna, Wright, & Mojardin, 1999; Hockley & Consoli, 1999) have been shown to decrease the use of recall in a recognition task. Similarly, decreasing the available decision time (Rotello et al., 2000) and weakening the strength of encoding (Yonelinas, 1997) also decrease the recollective component in recognition. Importantly, however, Dobbins, Kroll, Yonelinas, and Liu (1998) have

found dissociations between free recall performance and the recollective component of recognition. Thus, although free recall and recollection are often influenced by the same factors, they are not identical processes.

Of course, one could also ask a more specific question about recollection, namely what factors influence the use of *specific* retrieval processes in recognition. For example, one might assume that the use of an exhaustive search process would decrease as list length increased, for the simple reason that searching the entire list for a memory trace that matched the test probe would become increasingly time-consuming, difficult, and inaccurate. Similarly, one might assume that having mutually exclusive stimuli (e.g., study *frog* or *frogs,* but never both) would facilitate the use of recall-to-reject processing because subjects who recalled a word similar, but not identical, to a test probe could be certain that the similar word was a lure. Although this hypothesis is attractive and mutual exclusivity of the stimuli may indeed be a necessary precondition for the use of recall-to-reject process, the exclusivity of stimuli is not sufficient to allow use of recall-to-reject processing (Rotello & Heit, 1999, Exp. 4).

Other factors that facilitate the use of a recall-to-reject process are motivational or strategic: When subjects are told that using recall-to-reject will help them perform well on the task, they are more likely to actually use the process (Rotello et al., 2000). Also, the ease with which a unique critical feature of the studied event can be retrieved should influence the use of recall-to-reject; easier retrieval of the critical, mismatching feature(s) of the probe and memory trace should facilitate rejection of the similar foil. For example, when subjects are making an associative recognition decision on an A–D rearranged pair, retrieval of any of the features of B (studied with A) or C (studied with D) would allow rejection of the probe. Because the word pairs are commonly generated at random in these experiments, B and D (and A and C) could be expected to differ on a large number of features, making discrimination easy and thus recall-to-reject emerges in the data (Rotello & Heit, in press). In contrast, when the unique identifying features are more difficult to detect, such as when the studied item and similar foil are pseudowords that differ only in terms of one or a few letters, then recall-to-reject is less likely to emerge (Rotello & Heit, 1999).

V. Other Measures of Recall-to-Reject Processing

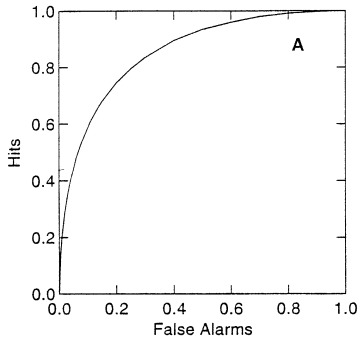
The strongest evidence on the specific nature of the recall process in item and associative recognition has come from response-signal experiments that describe the entire time course of a recognition judgment (Rotello &

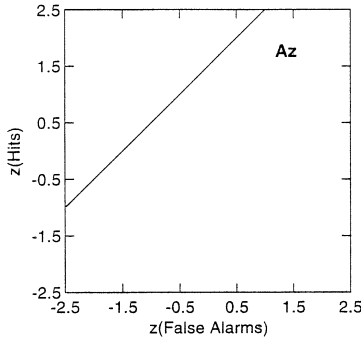
Heit, 1999, in press). However, there are limitations to those data: Finding evidence consistent with a given recall process (e.g., recall-to-reject) necessarily precludes detecting the presence of any other recall processes (e.g., recall-to-accept, exhaustive search). Data consistent with recall-to-reject, for example, are best fit with a nonmonotonic pseudo-d_L function, whereas data consistent with recall-to-accept are best fit by a monotonic pseudo-d_L curve. The two types of data cannot be observed simultaneously. In addition, the response-signal data can provide evidence that recall-to-reject processing is or is not used in a given experimental situation, but it does not quantify that processing (i.e., how much of the time is recall-to-reject used, when appropriate?). Thus, converging evidence from other paradigms is needed to provide important additional evidence on the nature of the recall process(es) in recognition.

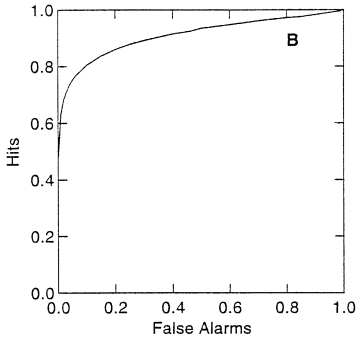
Two converging methods are employed here to evaluate data from a new confidence-rating recognition experiment. First, evidence from ROC curves is used. As we will see, both the shape of the ROC curve and its points of intersection with the *x*- and *y*-axes are informative about the nature of the processes underlying a recognition decision. Second, multinomial modeling, in the form of *conjoint recognition theory* (Brainerd, Reyna, & Mojardin, 1999), is used to provide independent estimates of the contribution of various familiarity- and recall-based processes to the recognition judgments. Before describing the experiment, however, a more detailed discussion of both ROC curves and conjoint recognition theory is required.

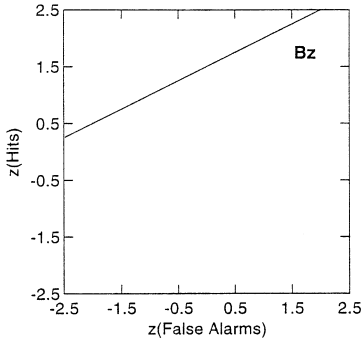
VI. The Receiver Operating Characteristic Curve

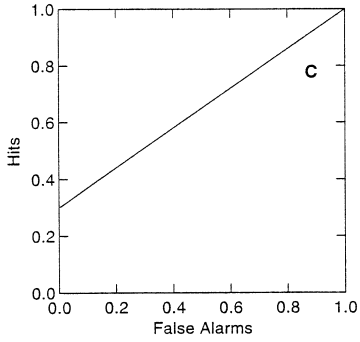
Most recognition experiments, including those employing a response-signal paradigm, ask subjects to simply judge whether a particular test probe had been studied or not. Presumably, subjects in such experiments choose a particular criterion level of familiarity or level of confidence and apply it consistently (Stretch & Wixted, 1998). However, subjects can also judge that one item is ''probably old,'' whereas another falls in the class of ''sure old'' items. If subjects are asked to rate the confidence with which they made each ''old'' or ''new'' judgment, an ROC curve can be generated. An ROC curve is a plot of the hit and false alarm rates at each confidence level, summed with all hits and false alarms for higher confidence levels. In Figure 4A, for example, the left-most point on the curve would represent the hit and false alarms from the highest (i.e., ''sure old'') confidence level and the second point from the left would represent the hit and false alarm rates for the ''sure old'' and ''probably old'' categories combined. Thus, ROC curves are monotonically increasing functions, and recognition judg-

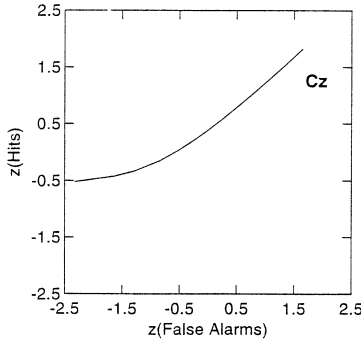

A
Hits
False Alarms
0.0
0.2
0.4
0.6
0.8
1.0

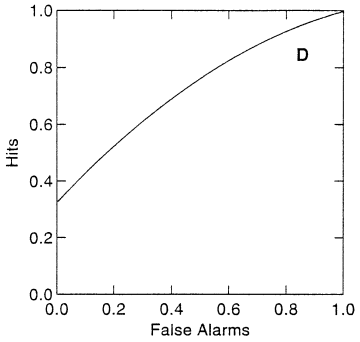

Az
z(Hits)
z(False Alarms)
2.5
1.5
0.5
-0.5
-1.5
-2.5

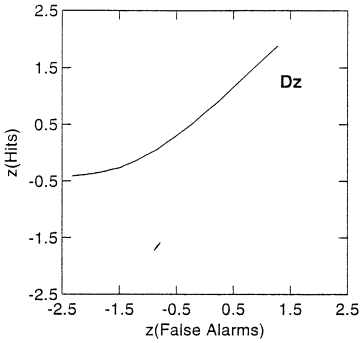

B
Hits
False Alarms
0.0
0.2
0.4
0.6
0.8
1.0


Bz
z(Hits)
z(False Alarms)
2.5
1.5
0.5
-0.5
-1.5
-2.5


C
Hits
False Alarms
0.0
0.2
0.4
0.6
0.8
1.0


Cz
z(Hits)
z(False Alarms)
2.5
1.5
0.5
-0.5
-1.5
-2.5


D
Hits
False Alarms
0.0
0.2
0.4
0.6
0.8
1.0


Dz
z(Hits)
z(False Alarms)
2.5
1.5
0.5
-0.5
-1.5
-2.5

ments that are made on an *N*-point confidence scale produce ROC curves with *N*-1 points (the *N*th point represents the sum of all responses to old and new probes, or [1,1]). Importantly, the shape of the ROC curve can inform us about the underlying processes that contributed to the judgments.

When the underlying familiarity values of old and new test probes are normally distributed with equal variance, the resulting ROC curve will be curvilinear and symmetric about the minor diagonal, as shown in Figure 4A.[2] This can be seen to be true by imagining two normal distributions with means d' units apart: The area under the "old" distribution to the right of each confidence criterion is the hit rate, and the area under the "new" distribution to the right of each criterion is the false alarm rate. Both the hit and false alarm rates will increase smoothly as the criterion is moved left to become less conservative. The symmetry of the ROC occurs because of the assumption of equal variance distributions. Equal variance distributions necessarily result in symmetric ROCs because the pair of distributions itself is symmetric. That symmetry is destroyed when the distributions have unequal variances, leading to the asymmetric ROC curves that are typically found empirically (see Figure 4B). Thus, we can predict that pure familiarity-based recognition based on normally distributed familiarity values (like those assumed by the global matching models) will result in curvilinear and possibly symmetric ROCs.

On the other hand, suppose that recognition judgments were based entirely on the results of a recollective process: If you can recall the test item from the study list, you call it "old" with the highest possible confidence. No other items are placed in that category: Regardless of how familiar they might seem, the test probes that cannot be recalled are always placed in lower confidence categories. Such all-or-none processes have been called single high-threshold models (e.g., Macmillan & Creelman, 1991) because they assume a high threshold for "true" recognition of studied items. The proportion of studied items that fall above that high threshold determines the *y*-intercept of the ROC curve. For items that fail to exceed the threshold,

---

Fig. 4. Hypothetical receiver operating characteristic data showing the use of equal variance normally distributed familiarity values (A), unequal variance familiarity (B), pure recollection (C), or both familiarity and recollection (D). Panels Az–Dz show the same hypothetical data transformed onto normal-normal coordinates. (From "Recall-to-reject in recognition: Evidence from ROC curves," by C. M. Rotello, N. A. Macmillan, and G. Van Tassel, 2000, *Journal of Memory and Language.* Copyright by Academic Press. Reprinted with permission.)

[2] Unless otherwise specified, we will assume that all distributions of familiarity are normal. Other types of distributions would naturally make different predictions about the shape of the resulting ROC curves.

response categories (i.e., the remaining lower levels of confidence) are chosen at random for both the studied items and the distractors. This random assignment of items to responses provides the same increment to both the hit and false alarm rates for each confidence level, which leads to a constant slope. Thus, these single high-threshold models generate linear ROC curves (see Figure 4C), which are rarely observed empirically.

Finally, consider a mixed model of recognition, in which both familiarity and recollection play a role. Such a model predicts ROC curves that are a blend of those predicted by familiarity and recollection, showing a *y*-intercept of greater than zero (caused by a recall-to-accept process in which some studied items are recollected) and a curvilinear shape with a decreasing slope (caused by the normal distributions of familiarity). Figure 4D shows a hypothetical example. Notice that Figure 4D looks somewhat similar to Figure 4B, the unequal-variance familiarity-based model. Theoretically, the two models can be distinguished by plotting the ROC data in normal–normal (z) space: the zROC from a binormal unequal variance model will always be linear with slope equal to the ratio of the standard deviations of the new and old distributions (see Figures 4Az and 4Bz). In contrast, the zROC of a pure recollection model is curvilinear (see Figure 4Cz; Macmillan & Creelman, 1991), and models that assume both familiarity and recollection operate in recognition also display some curvature in the zROC (see Figure 4Dz). Empirically, however, it has proven difficult to distinguish zROCs based on unequal-variance distributions from those based on a combination of familiarity and recollection.

Yonelinas (1994, 1997, 1999) has provided a wealth of data consistent with these predictions: When a recall process is unlikely to be used, such as when the study lists are long and rapidly presented, or when the tested participants are amnesic, curvilinear and symmetric ROCs emerge (e.g., Yonelinas, 1994, 1997; Yonelinas et al., 1998). On the other hand, when recall of the studied materials is made easier, for example by using short study lists, linear ROCs and U-shaped zROCs appear (Yonelinas, 1997, 1999).

ROC curves can also provide useful information on the presence of recall-to-reject processing in recognition. Yonelinas (1997) argued that sometimes participants are able to recall that a test probe is ''new.'' Rather than simply being unfamiliar, such items are actually recollected as ''new'' because of a mismatch with something that *was* studied (i.e., because of a recall-to-reject process). For example, participants might sometimes recall that A and B were studied together, which can allow them to reject the distractor test probe ''A–C'' as ''new.'' Rotello, Macmillan, and Van Tassel (2000) argued that evidence for recall-to-reject can appear in ROC curves when such similar foils are placed in the highest confidence ''new'' category,

and no studied items are given this response.[3] In that case, the point at which the ROC curve first hits the upper-*x*-axis, which we call the *upper-x-intercept,* occurs when the false alarm rate is less than 1.0. It is this reduction of the upper-*x*-intercept that can indicate the presence of recall-to-reject processing. The upper-*x*-intercept does not itself reveal this proportion, because it must typically be estimated by extrapolation. Moreover, the reduction in the upper-*x*-intercept alone is not sufficient to identify the use of recall-to-reject processing.

Remember that recall-to-reject is a very targeted process: It facilitates the rejection of similar foils, but it does not aid in the rejection of completely new foils or in the acceptance of studied items. Thus, a reduced upper-*x*-intercept should not occur for old–new discrimination, as is plotted in a typical ROC curve, because judgments of those stimuli are not affected by recall-to-reject processing. However, recognition of probes that are new but similar to studied stimuli could involve recall-to-reject. Thus, an *old–similar ROC,* which is a plot of hits against false alarms to similar foils, should show a reduced upper-*x*-intercept relative to the standard *old-new ROC* when recall-to-reject processing is used. It is actually the difference in the upper-*x*-intercepts of these two ROC curves that indicates the amount of recall-to-reject processing. Hypothetical old–new and old–similar ROC curves are shown in Figure 5.

Four experiments supported our predictions (Rotello et al., 2000). In two item recognition experiments, subjects were asked to recognize studied words, plurality-changed foils, and completely new words using a six-point confidence scale. When subjects were told that recall-to-reject processing would help their performance, they used it for about 30% of the items. When they were not encouraged to use recall-to-reject, they used it about half as often (14% of the time). Similarly, two associative recognition experiments provided clear ROC evidence for recall-to-reject processing, consistent with the time-course data (Rotello & Heit, in press). In Rotello et al.'s (2000) Experiment 3, subjects studied word pairs and were asked to recognize intact, rearranged, and new pairs. The intact–new ROC curve had an upper-*x*-intercept of 1.0, whereas the intact–rearranged ROC curve had an upper-*x*-intercept of .66 (see Figure 6). Moreover, the reduced intercept for the intact–rearranged ROC curve only occurred when the recognition judgments were made slowly enough for the recall process to

[3] This is a double high-threshold model (Macmillan & Creelman, 1991). Recallable studied items are placed in the highest confidence "old" category, "recallable" foils are placed in the highest confidence "new" category, and the remaining responses are distributed at random over only the intermediate confidence categories. This type of model yields a linear ROC with a *y*-intercept that indicates the proportion of studied items that were recalled and an upper-*x*-intercept that indicates the proportion of new items that were recollected as new.

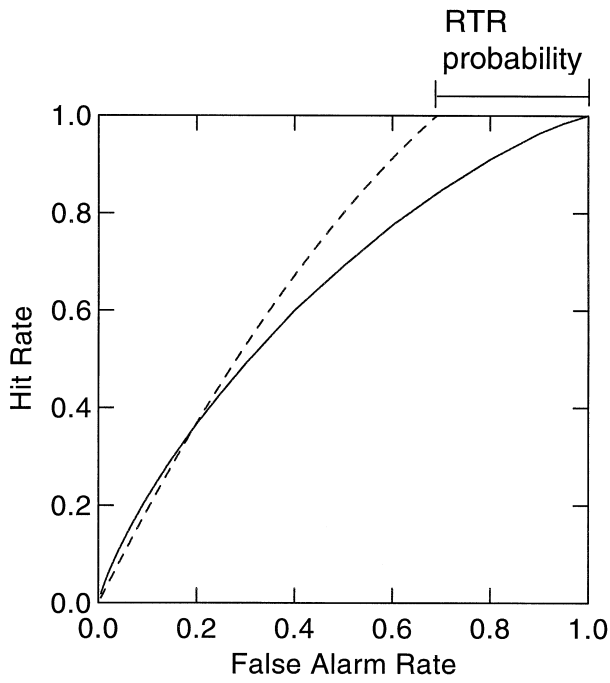


Fig. 5. Hypothetical receiver operating curves (ROC) showing the presence of recall-to-reject (RTR) processing. The solid curve represents ROC data showing discrimination between old and new items, whereas the dashed curve represents ROC data showing discrimination between old items and similar foils.

contribute to the judgments. When the recognition decisions were made after a very brief test interval (i.e., 450 ms), the ROC and intact–rearranged ROC curves both had upper-x-intercepts of 1.0 (Rotello et al., 2000, Exp. 4). In all four experiments, evaluations of the zROC curves supported the interpretation of the data in terms of a two-process model rather than a single-process unequal variance model: The old–similar and intact–rearranged zROC data were significantly U-shaped (see Yonelinas, 1994, 1997).

In each of the four experiments, the y-intercepts of both the intact–new ROC and intact–rearranged ROC curves were above zero (see Figure 6). Those y-intercepts are consistent with the presence of a recall-to-accept process in which the retrieval of a memory trace that matches the test probe facilitates confident acceptance of the probe as a studied item. Thus, Rotello et al.'s data provide an alternative window on the nature of the recall process(es) in recognition. Additionally, this ROC-window provides

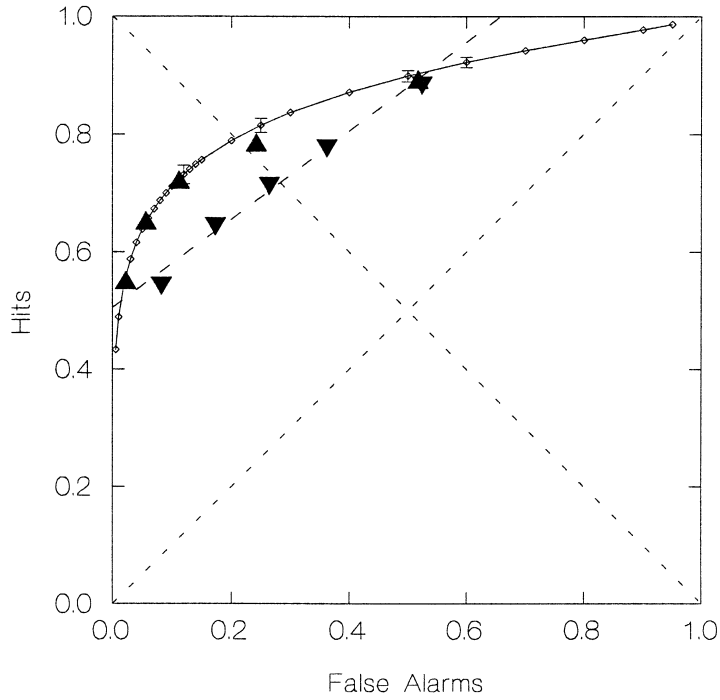


Fig. 6. Intact-new (triangles) and intact-rearranged receiver operating characteristics (ROC) (inverted triangles) data from Rotello et al. (2000, Exp. 3). Curves are the best-fitting functions. The intact-rearranged ROC data indicate the presence of recall-to-reject processing. The upper-*x*-intercept is reduced compared to the intact–new ROC curve. (From "Recall-to-reject in recognition: Evidence from ROC curves," by C. M. Rotello, N. A. Macmillan, and G. Van Tassel, 2000, *Journal of Memory and Language.* Copyright by Academic Press. Reprinted with permission.)

the possibility of observing more than one type of recall process simultaneously (i.e., both recall-to-reject and recall-to-accept processes are evident).

VI. Conjoint Recognition

One limitation in the interpretation of the ROC data presented by Rotello et al. (2000) is that there are no independent measures of the recall-to-reject or recall-to-accept processes. Although the data are consistent with our expectations and the evidence from the time-course literature, we run the risk of being content with data that "look right." The potential pitfalls of this contentment have been made obvious in recent years by problems

that have emerged with the process dissociation procedure, which generates estimates of the contributions of familiarity-based and recollective processes to recognition (Jacoby, 1991). There, although the initial reports showed that the estimates of the processes changed in plausible ways with changes in the experimental task, more careful evaluation of the model revealed serious interpretive difficulties (e.g., Curran & Hintzman, 1995; Ratcliff, Van Zandt, & McKoon, 1995). For example, Ratcliff et al. (1995) showed that even simulated recognition judgments that had been generated from a single-process familiarity-based model (i.e., Search of Associative Memory: e.g., Gillund & Shiffrin, 1984) were decomposed into their familiarity and "recollective" components by the process dissociation procedure. Because only familiarity had been used to generate the data, this was a serious failing of process dissociation.

One central problem with the process dissociation approach is that the two parameters in the model (i.e., the familiarity and recollective components of recognition) are estimated from only two data points (i.e., the positive responses to similar foils when subjects are instructed to accept those foils and when subjects are instructed to reject them). Thus, there are no remaining degrees of freedom with which to assess the parameter values or the overall quality of the model fit: assessment of the process dissociation parameter estimates occurs simply by observing the values across experimental conditions thought to increase either familiarity or recollection.

Fortunately, there have been some important new theoretical developments that allow both the estimation of the contributions of familiarity and recollection to recognition and, critically, the statistical evaluation of the parameter values and overall model fit. The *conjoint recognition model* (Brainerd, Reyna, & Mojardin, 1999) extends the process dissociation model by testing subjects under both inclusion instructions ("say 'old' to studied words and to new words that maintain the gist or meaning of studied words", called *Target+Related* instructions) and two different kinds of exclusion conditions. Under *Target* instructions, subjects are asked to respond positively only to studied words, thus excluding both new and related words; under *Related* instructions, subjects are to respond positively only to unstudied related words (i.e., similar foils), excluding both new words and studied words. These three conditions yield a total of nine data points (P("yes") to targets, similar foils and new words in each condition). The model, as we will see, has seven parameters.

The conjoint recognition model provides a system of equations that represent the probabilities of acceptance or rejection of the three types of stimuli under the three testing conditions. For example, the probability of calling a studied word "old" in the Target condition is

$$p_{t,T} = I_t + (1 - I_t)S_t + (1 - I_t)(1 - S_t)b_T, \quad (3)$$

where I_t is the probability that the studied item is recalled from memory, S_t is the probability that the familiarity of the studied item exceeds a criterion (leading to an "old" judgment), and b_T is a response bias parameter. As in the process dissociation procedure, then, a studied word is called "old" if it is familiar enough or if it is recalled; additionally, it is called "old" if there is a strong enough bias in the absence of both familiarity and retrieval. In the Target+Related condition, a studied probe is accepted under the same set of memory conditions described by Equation 3, except that the bias parameter is different (now called b_{T+R}) to allow for the possibility that a different proportion of test items will called "old." Positive responses to targets are analogous in the Related condition, except that retrieval from memory would lead to rejection of the test probe:

$$p_{t,R} = (1 - I_t)S_t + (1 - I_t)(1 - S_t)b_R \quad (4)$$

The probabilities of a positive response to a similar lure in each of the three conditions can also be simply described. In the Target condition, the similar lures should be rejected, so they are only accepted when the familiarity of the lure (S_r) is great enough to exceed the decision criterion and a recall-to-reject process (N_r) is not used successfully:

$$p_{r,T} = (1 - N_r)S_r + (1 - N_r)(1 - S_r)b_T \quad (5)$$

In contrast, similar lures are supposed to be accepted in the Related condition. Hence, the clearest evidence in favor of accepting similar lures under those instructions is when the studied item to which the lure is similar can be retrieved:

$$p_{r,R} = N_r + (1 - N_r)S_r + (1 - N_r)(1 - S_r)b_R \quad (6)$$

In the Target+Related condition, similar foils are accepted under the same memory conditions as are described by Equation 6, except that the bias parameter (b_{T+R}) is changed to allow for the possibility of a different proportion of test probes being called "old."

Finally, completely new test foils that are unrelated to any studied word will be accepted only because of guessing or bias. Thus, the three bias parameters give the probabilities of false alarms to those lures.

The conjoint recognition system of equations can be fit to data quite easily with the maximum likelihood machinery for fitting multinomial mod-

els. (See Brainerd, Reyna, & Mojardin 1999; programs for fitting generalized processing tree structures are available at Hu's web site: http://xhuoffice.psyc.memphis.edu/gpt/index.htm). Thus, the conjoint recognition framework provides an identifiable set of parameters that describe the contributions of familiarity- and retrieval-based processes to recognition judgments made under a variety of instructions (see Brainerd, Reyna, & Mojardin, 1999). Most relevant for our purposes, the conjoint recognition method will provide an independent measure of recall-to-reject processing: the N_r parameter in Equations 5 and 6 represents the probability that a studied item is retrieved from memory and found to be different from the similar test probe. This is the result of a recall-to-reject process. In the experiment described next, the ROC approach to estimating recall-to-reject processing (Rotello et al., 2000) was combined with the conjoint recognition methodology. The two very different statistical analyses of the data yielded two independent estimates of recall-to-reject processing. To the extent that they agree, we can be confident that the recall-to-reject estimates provided by Rotello et al. are reasonable and accurate. (Note that the conjoint recognition equations will *not* accurately describe the *shape* of the ROC curves generated in a recognition task. In its current mathematical form, the conjoint recognition framework predicts exclusively linear [i.e., high threshold] ROC curves, which are rarely observed empirically [e.g., Yonelinas & Jacoby, 1996].)

VIII. Experiment

A. Materials

1. Subjects

Sixty-one undergraduates at the University of Massachusetts participated in exchange for extra credit in their psychology courses. Subjects were assigned to experimental conditions in order of their appearance in the lab (i.e., every third subject was assigned to the same condition).

2. Stimuli

The stimuli were 300 common nouns (mean frequency 75.7 per million, Kucera & Francis, 1967) and their plurals (e.g., *truck–trucks*). The singular nouns were 3–12 characters in length and were chosen only if their plural form could be created by adding "s." For each subject, one word from each of 144 singular-plurals pairs was selected at random for presentation. The selected words were evenly divided into three study lists of 48 critical

words each. For each list, an additional 4 words were selected from the pool to fill the primacy and recency buffers, thus creating study lists of 52 total words. Presentation order of the critical words was randomized for each list and for each subject.

A test list was constructed for each study list, as follows. Twenty-four of the studied words were selected at random to serve as ''old'' test items, the unstudied alternate form (i.e., singular or plural) of the remaining 24 studied words served as the ''similar'' foils, and 24 additional unstudied words (including both singular and plural forms in a random mixture) were selected from the remaining 144 singular-plural pairs in the pool to serve as completely new foils. The only constraint was that the completely new foils had to have unique roots (i.e., if *chair* was tested then *chairs* was not). Test order was randomized for each subject and for each list.

3. *Procedure*

Subjects studied three lists of 52 words each, at a 3-sec presentation rate. They were asked to learn the words for a memory test, and were told to pay particular attention to the plurality of each studied word. Immediately following each study list, they were asked to recognize words that had been studied on that list, plurality-changed versions of studied words from that list, and completely new words that had not appeared on any list.

In the *Target Condition* ($n = 20$), subjects were asked to respond positively to studied words and to reject both new words and plurality-changed foils. These subjects were told that only the singular or plural form of a word was studied, so that if they could remember *computers,* for example, they could be sure that *computer* had not been studied. The self-paced responses were made on a 6-point confidence rating scale (1 = sure old; 2 = probably old; 3 = maybe old; 4 = maybe new; 5 = probably new; 6 = sure new).

In the *Related Condition* ($n = 21$), subjects were asked to identify the similar lures. They were told to respond positively to the plurality-changed foils, and negatively to both new and studied words. Subjects in this condition were informed that only the singular or plural form of a word was studied, so that if they could remember *computers,* for example, they could be sure that *computer* was a plurality-changed foil. Responses were made on a 6-point confidence rating scale with appropriate anchoring labels (1=sure changed; 2=probably changed; 3=maybe changed; 4=maybe old or new; 5=probably old or new; 6=sure old or new).

In the *Target+Related Condition* ($n = 20$), subjects were asked to respond positively whenever the root form of a studied word appeared on the test list, and to reject completely new words. They were told that if they could

recall studying *computers,* then they should say ''yes'' to either *computers* or *computer.* Again, responses were made on an appropriately labeled 6-point confidence scale in which 1 represented the most positive response category.

Subjects in all conditions were encouraged to ask questions about their task at any point; none did so.

B. Results and Discussion

The ROC data from the three experimental conditions are shown in Figure 7. Panel A shows the probability of responding ''old'' to a studied word in the Target condition plotted against the probability of a false alarm to completely new words or to similar lures. In panel A, the two ROC curves can be labeled the *old–new ROC* and the *old–similar ROC,* respectively. These data are collected in a manner identical to that used by Rotello et al. (2000, Experiment 1). Panel B shows the probability of responding positively to similar foils against positive responses to studied words and completely new items in the Related condition. These *similar–old* and *similar–new* ROC curves are analogous to the curves shown in panel A in that the cognitive processing involved in making the judgments (i.e., *old vs. similar; old* or *similar vs. new*) is similar. Finally, the panel C shows the data from the Target+Related condition: positive responses to studied words and similar foils are plotted on the *y*-axis; positive responses to completely new words are shown on the *x*-axis.

ROC data present an unusual statistical problem: both axes represent dependent variables. For such data, least-squares regression is technically inappropriate because it minimizes the deviations between model and data for only one dependent variable. Nonetheless, Ratcliff, McKoon, and Tindall (1994) found that least-squares fits were generally comparable to fits derived with maximum likelihood estimation techniques. Consequently, we will use a mixture of statistical tools.

1. Target Condition

In the Target condition, subjects were asked to respond positively to studied words and to reject both new words and words that were similar to studied words but that had a different plurality. In considering the results from this condition, the first goal was to use the ROC data to evaluate the types of processes that contributed to the recognition decisions. Remember that recollection-based judgments lead to linear or nearly linear ROC data (see Figure 4C), whereas familiarity-based judgments are strongly curvilinear (Figures 4A and 4B). Consider the data from the Target condition (Figure 7A). The old–new ROC data should be curvilinear because the decision

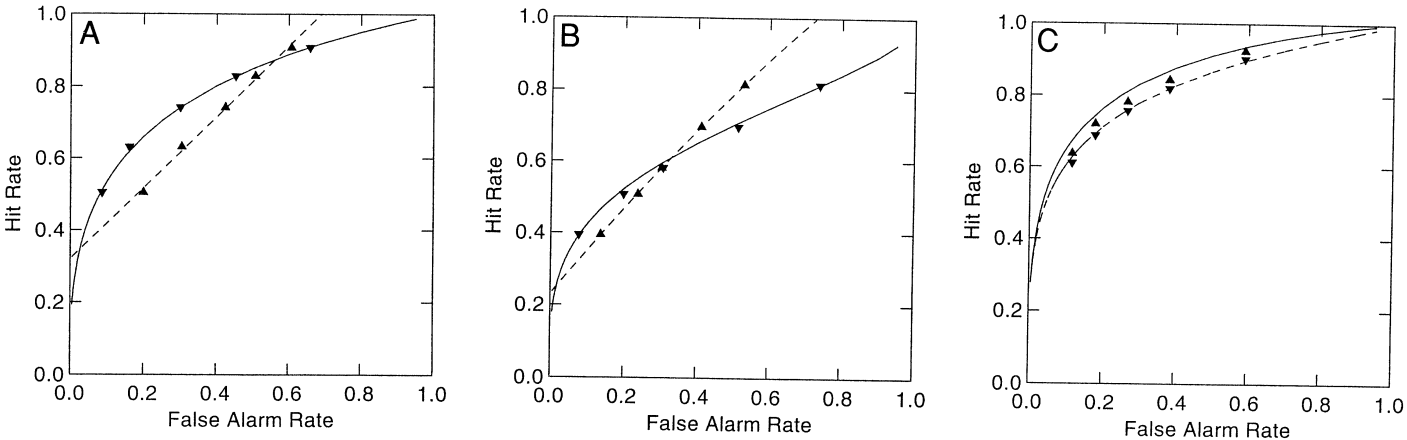


Fig. 7. ROC data from the Target (A), Related (B), and Target+Related (C) conditions. In (A) and (B), the inverted triangles indicate positive responses to the appropriate items (i.e., studied words in the Target condition; similar foils in the Related condition) scaled against positive responses to new items, and the triangles indicate positive responses to appropriate items scaled against familiar lures (i.e., similar foils in the Target condition; studied items in the Related condition). In (C), the inverted triangles are positive responses to similar lures scaled against new items, and the triangles are positive responses to studied items plotted against positive responses to new foils. The curves in each panel are the best-fitting functions (see text).

between old and new stimuli could easily be made on the basis of familiarity alone. In contrast, the old–similar ROC data shows discrimination between two very similar kinds of test probes (they vary only in terms of their plurality); consequently, those data should be essentially linear, as familiarity would not be sufficient to recognize the words. Both of these predictions were supported by the data, as can be seen in Figure 7A: the old–new data were well fit by a binormal ROC curve (fitting accomplished with Metz's, 1998, ROCKIT program, which uses maximum likelihood estimation), whereas the old–similar ROC was too linear to provide a good fit with the assumptions of a binormal signal detection framework. Consequently, the hits to studied items were regressed on the false alarms to similar foils. The best-fitting equation (i.e., P(hit) = .320 + .985 × P(false alarm to similar foils)) was linear and the fit was excellent (R^2 = .994). This function is shown as the dashed line superimposed on the data in Figure 7A.

Next, consider the same ROC data plotted on normal–normal (z) coordinates. In z-space, recognition responses based on normally distributed familiarity values alone will lead to linear zROCs; unequal variance distributions will lead to zROCs that do not have unit slope. In contrast, recognition judgments based only on a recollective (high threshold) process will be U-shaped. Recognition judgments made with both processes will also be U-shaped, although empirically the dual process model is often difficult to dicriminate from the unequal variance familiarity-based model. As can be seen in Figure 8, the zROCs support our conclusions from the ROC data. The data from the Target condition are shown in the left panel, where the old–new zROC is best fit with a linear function (z_{old} = 1.031 + .724 z_{new}, R^2 = .999), showing that old–new discrimination was based on familiarity alone, and the old–similar zROC data are best fit with a quadratic (U-shaped) function (z_{old} = .918 + 1.379 z_{sim} + .390 z_{sim}^2, R^2 = .999), showing that recollection contributed to the discrimination of studied items from similar foils.

Having established that a recollective process was used for the old–similar discriminations, we can now ask more specific questions about the nature of that process. Remember that a recall-to-accept process is indicated by the y-intercept in these data, which shows the proportion of items that exceeded the retrieval threshold and were called ''old.'' In the Target condition, the old–similar y-intercept is .320, indicating that 32% of the studied items were accepted with a recall-to-accept process. That was not the only recall process used: the upper-x-intercept for the old–similar function was .690, which is less than the upper-x-intercept of the old–new ROC (i.e., 1.0). That difference in upper-x-intercepts, .31, indicates that 31% of similar foils were rejected with the use of a recall-to-reject process in the Target condition.

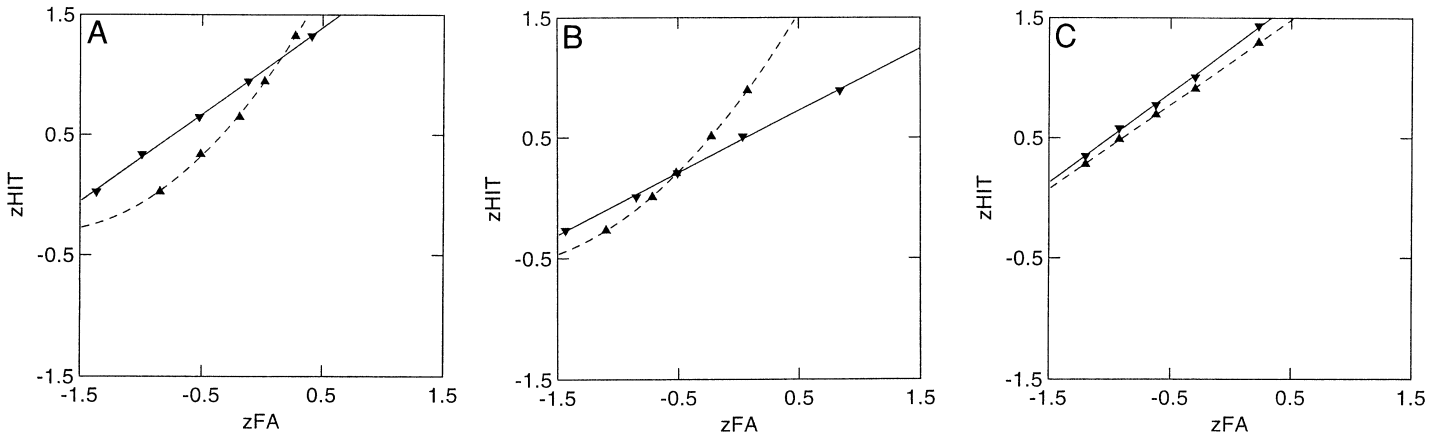


Fig. 8. zROC data from the Target (A), Related (B), and Target+Related conditions (C). In (A) and (B), the inverted triangles indicate positive responses to the appropriate items (i.e., studied words in the Target condition; similar foils in the Related condition) scaled against positive responses to new items, and the triangles indicate positive responses to appropriate items scaled against familiar lures (i.e., similar foils in the Target condition; studied items in the Related condition). In the right panel, the inverted triangles are positive responses to similar lures scaled against new items, and the triangles are positive responses to studied items plotted against positive responses to new foils. The curves in each panel are the best-fitting functions (see text).

2. *Related Condition*

In the Related condition, subjects were asked to respond positively to test probes that were words that had been studied in a different plurality. That is, they were to identify the plurality-changed similar foils and to reject both studied words and new words. As in the analysis of the Target condition, our first goal was to evaluate whether familiarity and/or recollection contributed to the recognition decisions. Because we expected familiarity to be sufficient to allow discrimination of similar foils from completely new words, the similar–new data from the Related condition was predicted to be curvilinear. In contrast, discrimination of similar foils from studied words was expected to involve a recollective process, because both types of stimuli would be very familiar to the subjects. Thus, the similar–old data from the Related condition were expected to be linear.

As can be seen in Figure 7B, the data from the Related condition also supported the predictions. The similar–new data were well fit with a binormal ROC curve, showing that familiarity contributed to those recognition judgments. In contrast, the similar-old ROC data were too linear to be fit with a familiarity-based model alone. Analogous to the analysis in the Target condition, then, the positive responses to similar foils were regressed on the positive responses to studied items. The best-fitting equation (i.e., P(positive response to similar foil) = .230 + 1.236 × P(positive response to studied word) −.247 × P(positive response to studied word)^2) included a quadratic component, indicating that both a recollective process and a familiarity-based process contributed to the judgments. The quadratic equation fit the data perfectly ($R^2 = 1.0$) and is shown as the dashed curve superimposed on the data in Figure 7B.

As in the Target condition, the data from the Related condition were plotted on z coordinates (see Figure 8B). Again, those zROCs supported the conclusions about the contributions of familiarity and recollection to the judgments. The best fitting similar–new zROC function was linear ($z_{sim} = .468 + .518\ z_{new}$, $R^2 = .999$), indicating that those judgments were based on familiarity alone, whereas the similar–old zROC function was best fit by a U-shaped quadratic equation ($z_{sim} = .795 + 1.320\ z_{old} + .319\ z_{old}^2$, $R^2 = 1.0$), supporting the conclusion that discriminating similar foils from studied words involved a recollective process.

Finally, we can evaluate the specific nature of the recall component to these recognition decisions. One might expect that a reduction in the upper-x-intercept of the similar–old ROC curve compared to the upper-x-intercept of the similar–new ROC curve would indicate the use of a recall-to-reject process in the Related condition. It is important to note, however, that the retrieval process that allows rejection of the studied test probes

in this condition is not exactly the same process as recall-to-reject. Recall-to-reject retrieves memory traces of studied items that *mismatch* the test probe, thereby facilitating the *rejection* of the similar (unstudied) test probe. In contrast, rejection of the studied test probe in the Related condition requires that the recollective process retrieve a memory trace that *matches* the test probe and thus lead to a rejection of the studied test probe. In other words, in the Related condition the process than facilitates rejection of studied items is more like a recall-to-*accept* process than a recall-to-reject process in terms of the information that is retrieved from memory. Similarly, the process that allows the similar foils to be called old in the Related condition is more like the recall-to-reject process in the Target condition than like a recall-to-accept process, again because the information retrieved from memory that allows acceptance of the similar foil is a mismatching memory trace. Thus, the best indicator of recall-to-reject processing in the Related condition is the *y*-intercept, and the best indicator of recall-to-accept processing is the upper-*x*-intercept, although neither indicator is measuring exactly the same process as in the Target condition because the decision stage (i.e., accept or reject) differs.

The similar–old ROC function in the Related condition has a *y*-intercept of .23, indicating that about 23% of the studied items were rejected with the use of a recall-to-reject process. It also has an upper-*x*-intercept of .729, indicating that about 27% of the studied items were rejected because their memory traces were recalled. Both of these intercepts are reduced compared to the Target condition.

Because the task is similar in the Target and Related conditions, one would expect the processing estimates should converge on roughly the same value despite being represented on different axes in the two conditions. However, the observed proportions of recall-to-accept and recall-to-reject processing may be a bit smaller in the Related condition than in the Target conditions because the Related condition requires an extra ''translation'' step in the decision process once a memory trace is retrieved. For example, recalling a matching memory trace would typically be construed as evidence in favor of having studied that item, which would lead to a positive response. In the Related condition, though, that positive response must be converted to a negative decision (i.e., ''I have found a matching memory trace, but in this experiment that means I should reject this item''). Errors of translation would tend to reduce the evidence for the recollective process, because errors would appear to be familiarity-based responding (i.e., positive responses to studied items). Thus, the reduced estimates of recall-to-accept and recall-to-reject processing in the Related condition as compared to the Target condition would appear to be a natural consequence of the increased task difficulty in the Related condition.

3. *Target+Related Condition*

Finally, consider the data from the Target+Related condition, shown in Figure 7C. In this condition, we would expect both the old–new and similar–new ROC curves to be curvilinear, because all familiar items (i.e., both old and similar probes) are to be called "old" in the Target+Related condition. The data clearly support this prediction: ROCKIT provided good fits to both curves. In addition, the zROC data from the Target+Related condition, shown in Figure 8C, are consistent with an unequal-variance familiarity-based model of the underlying process: the best-fitting functions to the zROC data are linear with slopes of .749 (old-new) and .705 (similar-new).

4. *Summary of ROC Evidence on Recall-to-Reject Processing*

Together, then, the ROC data from the Target and Related conditions have converged on a mean estimate of about 27% recall-to-reject processing in the recognition of these stimuli. Other ROC data (Rotello et al., 2000) have also indicated that around 30% of similar foils are rejected with recall-to-reject processing, with estimates ranging from 14–34%. The lower estimates have resulted from experiments that either did not encourage subjects to use recall-to-reject processing (estimate: 14%) or involved more complex tasks (i.e., were collected in an associative recognition response-signal paradigm; estimate: 23%). Higher estimates of the contribution of recall-to-reject have emerged from experiments that explicitly encouraged subjects to use such a process (estimates: 30% and 34%). The lower estimate for recall-to-reject in the Related condition (23%) than in the Target condition (31%) is also consistent with this general pattern of more difficult tasks revealing less evidence for recall-to-reject processing. Similarly, recent data reported by Yonelinas (1999) indicated that a source discrimination task (i.e., was this item shown on List 1?) involved 10–40% recall-to-reject processing of items studied on the nontarget list, with lower estimates tending to correspond to task conditions that made the use of a recollective process less necessary (e.g., study lists of distinctly different familarity values).

IX. Conjoint Recognition Estimate of Recall-to-Reject Processing

Next, the conjoint recognition model was used to generate an independent estimate of the amount of recall-to-reject processing in these recognition judgments. To begin, the confidence rating judgments were converted to simple yes–no decisions by collapsing the responses at each of the three

levels of confidence (i.e., "sure," "probably," "maybe") for both the positive and negative judgments to each test probe type in each experimental condition. The probability of an "old" judgment is shown in Table I as a function of instruction condition (Target, Related, Target+Related) and stimulus type (studied, similar foil, new foil).

These response probabilities (e.g., $P_{r,T}$, the probability of a positive response to a similar foil in the Target condition) were fit to Brainerd et al.'s (1999) conjoint recognition equations using Hu's (1998) General Processing Tree software. How well did the model fit the data overall? Table I shows the observed and predicted proportions of "old" judgments in each condition. Although the model makes predictions that are fairly close to most of the observed values, it does tend to overpredict the positive reponses to similar foils in the Related condition, and to underestimate them in the Target+Related condition. These errors, plus the other smaller deviations between model and data, resulted in a goodness-of-fit measure that indicates the conjoint recognition model does *not* fit the data: $\chi^2(2) = 119.34$, $p < .001$.

The poor fit could have occurred for a number of reasons, including two strong assumptions about processing. First, conjoint recognition assumes that familiarity and recollection are stochastically independent; this is an assumption carried over from process dissociation and other multinomial models. As Curran and Hintzman (1995, 1997) and others have shown, violations of this independence assumption can lead to underestimates of the familiarity component of recognition when familiarity and recollection

TABLE I

Observed and Predicted Proportions of "Old" Responses to the Three Classes of Stimuli in Each Test Condition

	Observed proportion "old"	Predicted proportion "old"
Target condition		
Studied words	.741	.770
Similar foils	.424	.424
New foils	.300	.289
Related condition		
Studied words	.290	.290
Similar foils	.594	.669
New foils	.309	.278
Target+related condition		
Studied words	.803	.774
Similar foils	.761	.681
New foils	.256	.303

are positively correlated over subjects and items. However, for small values of the recollective parameter, like those in Table II, the underestimate of familiarity is quite small (e.g., Jacoby, Begg, & Toth, 1997). Thus, violation of the independence assumption, although likely in our paradigm, are probably not the cause of the poor fit.

The second strong assumption made by conjoint recognition seems a more plausible scapegoat: like process dissociation, conjoint recognition assumes that the parameter values are the same in all three instructional conditions. Brainerd, Reyna, and Mojardin (1999) showed that this assumption of parameter invariance holds true for I_t and S_t whenever

$$P_{t,T} = 1 - \frac{(1-p_{u,T})(1-P_{t,T+R})}{(1 - p_{u,T+R})}, \tag{7}$$

and for N_r and S_r whenever

$$P_{r,R} = 1 - \frac{(1-P_{u,R})(1-P_{r,T+R})}{(1 - p_{u,T+R})}. \tag{8}$$

Substitution of the observed proportions of ''old'' responses from Table I suggests that the parameter invariance assumption was not met in this experiment, especially for the N_{r} and S_r parameters. Equation 8 implies that the probability of an ''old'' response to a similar foil in the Related condition should be .778, whereas the observed proportion is only .594.

Although the poor fit of the model could be reason to dismiss conjoint recognition theory as an adequate description of these data, Batchelder and Riefer (1999) argue that the parameter estimates of the underlying processes, not the goodness-of-fit measures, are the essential contribution of models of this sort. After all, there are typically very few (i.e., 1 or 2) degrees of freedom available to fit the data in multinomial approaches so even good fits of model and data are not compelling (i.e., fitting N data points with $N - 1$ or $N - 2$ parameters is not impressive, even if the fit is nearly perfect). Consequently, the focus here will be on the parameter estimates of the underlying processes, shown in Table II. Although the parameters seem to vary somewhat across conditions (see Equations 7 and 8 and surrounding discussion), we will assume that they are reasonable approximations of the underlying processes.

Consider first the bias parameters in the Target (b_T) and Related (b_R) conditions, which indicate subjects' general tendency to respond positively in the absence of either familiarity-based or recollective information. These bias parameters can be thought of as indices of guessing ''yes.'' Because

TABLE II

BEST-FITTING PARAMETER VALUES (WITH STANDARD DEVIATIONS IN PARENTHESES) FOR THE CONJOINT RECOGNITION MODEL

Parameter	Value (*SD*)
I_t	.475 (.015)
N_r	.251 (.017)
S_t	.382 (.019)
S_r	.389 (.015)
b_T	.289 (.011)
b_R	.278 (.011)
b_{T+R}	.303 (.012)

the tasks in the Target and Related conditions were fairly difficult, requiring the discrimination of studied words from similar plurality-changed foils, we might expect that the tendency to guess would largely reflect the proportion of items that were to be called ''old'' (Target condition) or ''plurality-changed'' (Related condition). Indeed, that turned out to be the case: one-third of the test probes in each condition would appropriately be given a positive response, and the bias parameters in each condition revealed that subjects were willing to guess ''yes'' on just under one-third of the new items. In the Target+Related condition, though, the task was easier: subjects merely had to identify words that seemed familiar (i.e., two-thirds of the test stimuli). Consequently, the tendency to guess should be reduced relative to the proportion of items that were to be called ''old,'' and it was. The bias to respond old was .303, which is only half what would be expected if subjects guessed ''old'' based on the proportion of words that seemed familiar.

Next consider the familiarity-based parameters, S_t and S_r, which indicate the probability that a studied word and a similar foil (respectively) seem familiar but cannot be confirmed or disconfirmed as studied items with a recollective process. These two parameters have essentially identical values (.382 and .389). That the two parameter values are so similar is consistent with the idea that familiar-seeming, but nonrecallable, stimuli are those for which the plurality of the studied form is not known (i.e., it has degraded in memory or was never encoded). When the plurality of the studied word (e.g., *frog*-) is unknown, both the studied word (*frog*) and the similar foil (*frogs*) would be equally similar to the memory trace, and thus they would seem equally familiar.

Now consider the two retrieval-based parameters in Table II, I_t and N_r. The probability that a studied word that matches the test probe is retrieved and leads to a positive recognition decision, I_t, is .475. This is an estimate of the amount of recall-to-accept processing; it is higher than that obtained from the ROC curves (.32 in the Target condition; .27 in the Related condition). Finally, the most critical parameter value for current purposes is N_r, which indicates the probability that a similar foil was rejected with a recall-to-reject process. N_r has a value of .251, indicating that about 25% of the similar foils were rejected with a recall-to-reject process. Thus, the conjoint recognition methodology yielded an estimate of recall-to-reject processing for this set of data that was very similar to the overall estimate of recall-to-reject processing obtained from the ROC data (i.e., 27%).

The close agreement between the ROC and the conjoint recognition model predictions of recall-to-reject processing is remarkable given the dramatically different assumptions made in the two approaches. After all, it is clear that theshold-based multinomial models in general, and the conjoint recognition framework in particular, cannot even predict the shape of the ROC curve. In theshold models, the relationship between hits and false alarms in conjoint recognition is always linear, whereas empirically derived ROC data are typically curvilinear (e.g., Yonelinas & Jacoby, 1996). Despite this limitation of the conjoint recognition model, the estimates for a specific underlying process, recall-to-reject, are nearly identical to those from the ROC data.

In summary, this confidence-rating experiment has provided additional evidence on the use of both recall-to-accept and recall-to-reject processes in an item recognition task. The benefits of using ROC analyses and conjoint recognition approaches to the issue are clear: Quantifiable estimates of the amount of both recall-to-accept and recall-to-reject processes (as well as guessing) can be obtained from the same set of data. These results stand in stark contrast to data collected in the response–signal paradigm, which can indicate that either a recall-to-accept or recall-to-reject process contributed to the judgments, but can never reveal both processes at once. Moreover, the response-signal paradigm gives only qualitative information about the type of recollective processing in recognition; it cannot directly indicate the proportion of trials on which recollection aids recognition.

X. Familiarity-Based Explanations of the Data

Recently, familiarity-based accounts of linear ROC data (Kelley & Wixted, 1999) have been proposed. Can the linear ROC curves shown in Figures 6 and 7 be explained with a familiarity-based model using normally distrib-

uted familiarity? If so, then explanations of the data that assume a recollective process may be overly complex. Kelley and Wixted (1999) used simulated data to show that linear ROC data can occur in associative recognition tasks when item information is normally distributed with relatively small variance, and associative information is both normally distributed with much larger variance and is inconsistently used by the ''subjects.'' When the associative information was weakly encoded, as when the word pairs were studied only once, this ''some-or-none'' model of associative information generated linear ROC data, without recourse to a threshold recollective process. The simulation data replicated empirical data quite well: linear associative recognition ROC data emerged when word pairs were studied once by subjects, and the more typically observed asymmetric and curvilinear ROC data emerged (see Figure 4B) in both the simulated and empirical data when the associative information was better learned. Thus, this ''some-or-none'' model of associative recognition challenges the recall-based interpretation of Rotello et al.'s (2000) linear associative recognition ROC data (see Figure 6). However, Kelley and Wixted's model depends upon item information being normally distributed, which will lead to curvilinear item recognition ROC data (see Figures 4A and 4B). The data in Figure 7, as well as in Rotello et al.'s (2000) Experiments 1 and 2, show threshold-based linear item recognition. Thus, whereas Kelley and Wixted's ''some-or-none'' model of linear associative recognition ROC data neatly explains the associative recognition ROC data without calling upon a recollective process, it does not seem to be able to completely explain the item recognition ROC data.

XI. Conclusions

There is now abundant evidence from a variety of paradigms that indicates recognition judgments are based on both a familiarity-based process and a recollective process. Data have been collected in paradigms as diverse as introspective reports of ''remembering'' and ''knowing'' (e.g., Hockely & Consoli, 1999; Yonelinas et al., 1998) and ERP recordings of brain activity throughout the time-course of recognition (e.g., Curran, in press). Data have also been collected in response-signal experiments (e.g., Rotello & Heit, in press), confidence rating tasks (e.g., Yonelinas, 1994), and in simple 'old'–'new' recognition tasks in combination with multinomial models (e.g., Brainerd et al., 1999). In each case, the same take-home message has emerged: the familiarity of a test probe provides the initial basis of recognition, and a recollective process provides information for slower, more accurate recognition judgments.

The conclusion that both familiarity and recollection contribute to recognition is incomplete without an analysis of the type of recollective process involved, however. A number of possible recall process in recognition have recently been described (Rotello & Heit, 1999), and evidence on the use of the most commonly assumed process, the recall-to-reject process, has been collected in a variety of paradigms. One clear message that is developing from these new data is that the recall process is both more flexible and under a greater degree of strategic control than was previously thought. It is no longer appropriate to simply assume that the recollective component of recognition operates in a recall-to-reject fashion.

Acknowledgments

This research was supported by grants to the author from the National Institutes of Health (1 R01 MH60274-01A1), National Science Foundation (SBR-9753136), and the University of Massachusetts. I thank Evan Heit and Neil Macmillan for many stimulating discussions of these issues, Chuck Brainerd for his assistance with the conjoint recognition methodology and for comments on the data, and Jennifer Plaza for collecting the data.

References

Atkinson, R. C., & Juola, J. F. (1973). Factors influencing speed and accuracy of word recognition. In S. Kornblum (Ed.), *Attention and performance IV* (pp. 583–612). New York: Academic Press.

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary development in mathematical psychology: Learning, memory, and thinking* (pp. 243–293). New York: Freeman.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin and Review, 6,* 57–86.

Borowiak, D. S. (1989). *Model discrimination for nonlinear regression models.* New York: Marcel Dekker.

Brainerd, C. J., Reyna, V. F., & Mojardin, A. H. (1999). Conjoint recognition. *Psychological Review, 106,* 160–179.

Brainerd, C. J., Reyna, V. F., Wright, R., & Mojardin, A. H. (1999). *Recollection rejection: The judgment of nonoccurrence.* Manuscript submitted for publication.

Brainerd, C. J., Wright, R., & Mojardin, A. H. (1999). *Beyond recollection and familiarity.* Manuscript submitted for publication.

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *Quarterly Journal of Experimental Psychology, 29,* 461–473.

Clark, S. E. (1992). Word frequency effects in associative and item recognition. *Memory and Cognition, 20,* 231–243.

Clark, S. E., & Burchett, R. E. R. (1994). Word frequency and list composition effects in associative recognition and recall. *Memory and Cognition, 22,* 55–62.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of memory: How the models match the data. *Psychonomic Bulletin and Review, 3,* 37–60.

Clark, S. E., & Hori, A. (1995). List length and overlap effects in forced-choice associative recognition. *Memory and Cognition, 23,* 456–461.

Clark, S. E., Hori, A., & Callan, D. E. (1993). Forced-choice associative recognition: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 19,* 871–881.

Curran, T. (in press). Brain potentials of recollection and familiarity. *Memory and Cognition.*

Curran, T. (1999). The electrophysiology of incidental and intentional retrieval: ERP old/new effects in lexical decision and recognition memory. *Neuropsychologia, 37,* 771–785.

Curran, T., & Hintzman, D. L. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21,* 531–547.

Curran, T., & Hintzman, D. L. (1997). Consequences and causes of correlations in process dissociation. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 23,* 496–504.

Dobbins, I. G., Kroll, N. E. A., Yonelinas, A. P., & Liu, Q. (1998). Distinctiveness in recognition and free recall: The role of recollection in the rejection of the familiar. *Journal of Memory and Language, 38,* 381–400.

Dosher, B. A. (1984). Discriminating preexperimental (semantic) information from learned (episodic) associations: A speed-accuracy study. *Cognitive Psychology, 16,* 519–555.

Düzel, E., Yonelinas, A. P., Mangun, G. R., Heinze, H.-J., & Tulving, E. (1997). Event-related potential correlates of two states of conscious awareness in memory. *Proceedings of the National Academy of Sciences, 94,* 5973–5978.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67.

Gronlund, S. D., Edwards, M. B., & Orht, D. D. (1997). Comparison of the retrieval of item versus spatial position information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1261–1274.

Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 846–858.

Hintzman, D. L., & Caulton, D. A. (1997). Recognition memory and modality judgments: A comparison of retrieval dynamics. *Journal of Memory and Language, 37,* 1–23.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language, 33,* 1–18.

Hintzman, D. L., & Curran, T. (1997). Comparing retrieval dynamics in recognition memory and lexical decision. *Journal of Experimental Psychology: General, 126,* 228–247.

Hockley, W. E., & Consoli, A. (1999). Familiarity and recollection in item and associative recognition. *Memory and Cognition, 27,* 657–664.

Hu, X. (1998). General Processing Tree (Version 1.0) [Computer software]. Memphis, TN: University of Memphis.

Humphreys, M. S. (1978). Item and relational information: A case for context independent retrieval. *Journal of Verbal Learning and Verbal Behavior, 17,* 175–187.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541.

Jacoby, L. L., Begg, I. M., & Toth, J. P. (1997). In defense of functional independence: violations of assumptions underlying the process-dissociation procedure? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 484–495.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110,* 306–340.

Johnson, M. K., Kounios, J., & Reeder, J. A. (1994). Time-course studies of reality monitoring and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1409–1419.

Johnson, R. J. (1995). Event-related potential insights into the neurobiology of memory systems. In F. Boller & J. Garfman (Eds.), *Handbook of neuropsychology* (Vol. 10, pp. 135-163). Amsterdam: Elsevier.

Kelley, R., & Wixted, J. T. (1999, November). *The complexities of associative recognition.* Paper presented at the 40th Annual Meeting of the Psychonomic Society, Los Angeles.

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Loftus, E. F. (1997). Creating childhood memories. *Applied Cognitive Psychology, 11*(Special Issue), S75–S86.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* New York: Cambridge University Press.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252–271.

Mandler, G. (1991). Your face looks familiar but I can't remember your name: A review of dual process theory. In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory to data: Essays on human memory in honor of Bennet B. Murdock* (pp. 207-225). Hillsdale, NJ: Erlbaum.

McElree, B., Dolan, P. O., & Jacoby, L. L. (1999). Isolating the contributions of familiarity and source information to item recognition: A time-course analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 563–582.

Metz, C. E. (1998). ROCKIT computer program. Department of Radiology, University of Chicago. [On-line] Available: http://www.radiology.uchicago.edu/krl/toppage11.htm

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

Ratcliff, R. (1980). A note on modeling accumulation of information when the rate of accumulation changes over time. *Journal of Mathematical Psychology, 21,* 178–184.

Ratcliff, R., & McKoon, G. (1982). Speed and accuracy in the processing of false statements about semantic information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 16–36.

Ratcliff, R., & McKoon, G. (1989). Similarity information versus relational information: Differences in the time course of retrieval. *Cognitive Psychology, 21,* 139–155.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 763–785.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1995). Process dissociation, single-process theories, and recognition memory. *Journal of Experimental Psychology: General, 4,* 352–374.

Rotello, C. M. (1999). Metacognition and memory for nonoccurrence. *Memory, 7,* 43–63.

Rotello, C. M., & Heit, E. (1999). Two-process models of recognition memory: Evidence for recall-to-reject? *Journal of Memory and Language, 40,* 432–453.

Rotello, C. M., & Heit, E. (in press). Associative recognition: A case of recall-to-reject processing. *Memory and Cognition.*

Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language, 43,* 67–88.

Rugg, M. D. (1995). ERP studies of memory. In M. D. Rugg & M. G. H. Coles (Eds.), *Electrophysiology of mind* (pp. 132–170). New York: Oxford University Press.

Rugg, M. D., Schloerscheidt, A. M., & Mark, R. E. (1998). An electrophysiological comparison of two indices of recollections. *Journal of Memory and Language, 39,* 47–69.

Strack, F., & Bless, H. (1994). Memory for nonoccurrences: Metacognitive and presuppositional strategies. *Journal of Memory and Language, 33,* 203–217.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 24,* 1379–1396.

Wickelgren, W. A., & Corbett, A. T. (1977). Associative interference and retrieval dynamics in yes-no recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 189–202.

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 681–690.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1341–1354.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory and Cognition, 25,* 747–763.

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1415–1434.

Yonelinas, A. P., & Jacoby, L. L. (1996). Response bias and the process dissociation procedure. *Journal of Experimental Psychology: General, 125,* 422–434.

Yonelinas, A. P., Kroll, N. E. A., Dobbins, I. G., Lazzara, M., & Knight, R. T. (1998). Recollection and familiarity deficits in amnesia: Convergence of remember/know, process dissociation, and receiver operating characteristic data. *Neuropsychology, 12,* 1–17.

This Page Intentionally Left Blank

REWARD LEARNING:
Reinforcement, Incentives, and Expectations

Kent C. Berridge

How rewards are learned, and how they guide behavior are questions that have occupied psychology since its first days as an experimental science. Many answers have been suggested during the past 100 years. The discussion continues today because the answer to these questions is more complex than it first appears, and also because there may be more than one correct answer.

I. Reinforcement Theories

Certain ideas in intellectual history have emerged occasionally on the scene in so clear and compelling a form that they seem as though they *must* be true (even if they are false). Self-evident to those who hold them, they have a tenacious foothold on intellectual life that is impervious to mere evidence against them. They may be wrong, and may be repeatedly shown to be wrong. Yet they will not die. Among this class of ideas is *reinforcement:* the idea that reward learning consists primarily of a process by which behavior is directly strengthened or weakened by the consequence that follows it.

Reinforcement is alive and strong as a concept in psychology. Covering the last 5 years of the 20th century, the PsycINFO database listed for each year over 500 published articles that include *reinforcement* as a key word.

THE PSYCHOLOGY OF LEARNING
AND MOTIVATION, VOL. 40

Copyright © 2001 by Academic Press.
All rights of reproduction in any form reserved.
0079-7421/00 $35.00

A. Descriptive (Atheoretical) Forms of the Reinforcement Concept: Skinner's Radical Behaviorism

The radical behaviorist sense of reinforcement was the simplest, though not chronologically the first version of the reinforcement concept. Reinforcement in a behaviorist sense, as used by Skinner, was merely a description of the relation between a change in a behavior caused by following it with a stimulus such as food. No explanation was offered. As Skinner put it,

> The Law of Effect is no theory. It simply specifies a procedure for altering the probability of a chosen response. But when we try to say *why* [italics in original] reinforcement has this effect, theories arise. Learning is said to take place because the reinforcement is pleasant, satisfying, tension reducing, and so on. (Skinner, 1950, p. 78).

Skinner took a dim view of theoretical attempts to explain the "why" of reinforcement: "Theories are fun. But it is possible that the most rapid progress toward an understanding of learning may be made by research which is not designed to test theories" (Skinner, 1950, p. 99).

In the first half of the 20th century, such an explicitly antitheoretical account had two forms of appeal. First, it offered a purely objective alternative way of talking about behavior that could be claimed to escape the theoretical disagreements of earlier mentalistic psychology. Second, at the time Skinner wrote it seemed to many possible to believe that his hope for "the most rapid progress" might actually be true. One could be impressed in 1938, when Skinner's (1938) first major book appeared, by the clarity of the early studies of operant performance curves, of the smooth versus scalloped differences in their shapes produced by fixed ratio versus fixed interval reinforcement schedules, and so on. Such precision in behavioral description had never before been achieved, let alone brought under experimental control. One could still hope then that the second half of the 20th century would be filled with many powerful demonstrations of new and useful behaviorist principles for predicting and controlling behavior.

But the Skinnerian hope turned out to be unfulfilled. Very few new principles of use were produced by behaviorist studies over the next 50 years. However objective a behaviorist description might be, it generated very little for predicting or control of behavior beyond the original law of effect, and a few straightforward applications of that law. And as Skinner himself acknowledged, the radical behaviorist concept of reinforcement offered nothing at all for understanding the "why" of reward's effect on behavior. It had no theory and no explanatory power.

Today there remain a few calls for a return to atheoretical behaviorism (Staddon, 1999; Uttal, 2000). They reflect an almost aesthetic preference held by these authors for a purely atheoretical behavioral description, which

avoids any inference about mediating process that could turn out to be wrong, even at the cost of explanatory power. But contemporary behaviorists draw mostly on the old evidence for support, and simply reassert the original Skinnerian faith that good progress might be made in an atheoretical fashion. That faith, once plausible, now has nearly died away from psychology for lack of fruition. Most important to the demise of behaviorism has been a consensus in psychology that the behaviorist fear of theory and psychological constructs was misplaced. It has proven eminently possible to test theories grounded in cognitive, associative, and motivational hypotheses over the past 50 years. Many such theoretical constructs have been disproven, others confirmed, and still others modified on the basis of experimental results. Contemporary psychology and neuroscience of reward learning are both concerned almost entirely with accounts of why reward works. We will therefore turn to theoretical answers to the why of reward and reinforcement.

B. Theoretical Versions of Reinforcement Concept: S–R Habits, Hedonic Reinforcers, and Drive Reduction

1. Thorndike's Original Law of Effect: Satisfiers and Annoyers

The Law of Effect was formulated first by Thorndike (Thorndike, 1898) in remarkably hedonic-sounding language, that is in terms of *satisfaction* or *annoyance*. Thorndike asked regarding animals, What do they *feel?,* a question familiar to today's psychologists of emotion and to affective neuroscience. In Thorndike's words, the initial goal of his experimental program with animals was to "not only tell more accurately *what they do* [italics in original], and give the much-needed information *how they do it,* [italics in original] but also inform us *what they feel* [italics in original] while they act,"(Thorndike, 1998, p. 1126). Regarding such feelings, Thorndike's law of effect stated originally that behavior that was followed by *satisfying* effects would be repeated more often in the future (Galef, 1998). Behavior followed by *annoying* effects, conversely, would be less likely to be repeated. Satisfaction and annoyance are essentially hedonic terms. They specify feelings or affective states of pleasure and displeasure. This use of a hedonic sense of the reinforcement concept soon diminished only to grow again over the intervening century.

Later formulations of the law of effect by Thorndike dropped the explicit postulation of affective states and adopted the more behaviorist formulation that led eventually to Skinner's position (Moss & Thorndike, 1934). In later versions, positive reinforcement became whatever property of a stimulus increased the frequency of the behavior it was contingent upon. No substitute for hedonic satisfaction, however, was offered to replace it until the emergence of drive reduction theories. Hedonic pleasure has always re-

mained the primary candidate, at least implicitly, for a specific psychological state to mediate reinforcement, when such a state is considered at all. As indicated by Skinner's quote above, this is often true even for behaviorists. But as behaviorists no longer tried explicitly to identify the nature of positive reinforcement states, processes of a different nature for ''stamping in'' behavior had to come from other sources, such as from associationist accounts of reinforcement.

2. *Stimulus–Response Association: Reinforcement by Strengthening of S-R Habits*

If reinforcement of a behavioral response means any specific psychological process at all, aside from hedonic pleasure, it must mean strengthening of a stimulus–response (S-R) habit. The idea that the impact of a reward on behavior might be mediated by the strength of a specific association was a stark alternative to hedonic accounts. An S-R association is essentially a *learning* process, not a motivational, emotional, or affectivexs one. Associationist behaviorism grew with the later works of Thorndike, John Watson (1913), and with subsequent generations of associationist behaviorist psychologists who evolved alongside the radical behaviorists, and who gave rise to Thorndike-based concepts of reinforcement that can be found still today in psychology and behavioral neuroscience (McFarland & Ettenberg, 1998; White, 1989).

Associationist behaviorists were not radical or atheoretical behaviorists as was Skinner, because they postulated specific psychological processes in order to explain their observations. In this case, the specific process was the strength of the learned association between two particular events that had been paired together in an individual's history. But they were behaviorists still because they restricted themselves to a few relationships, S-R associations, stimulus–stimulus (S-S) associations, and little else, which dealt entirely with events that could be observed objectively: physical stimuli, behavioral responses, and the measurable relation among them.

Associationist accounts of reinforcement allowed S-R theorists to drop hedonic satisfaction from the arena of postulated psychological events, yet still have a complete explanation for why reinforcers strengthened behavior. The explanation was simply that the occurrence of a reinforcer such as food could cause a direct increase in the strength of S-R association that preceded it. Whatever behavior had just been emitted was the strengthened response. Whatever stimulus had just been present when that behavior occurred was the stimulus that now became associated with the response. The strengthening of the association created something new: a habit. It imbued the stimulus with the ability to elicit again the response, in a direct,

essentially reflexive fashion. When presented a second time, the stimulus was now more likely to elicit the response. If reinforced a second time, the habit would be strengthened further. The apparent simplicity of this explanation of reinforcement was stunning. It allowed one to explain—not merely describe as Skinner could—why a behavior became more and more frequent and persistent when rewards were given contingent upon it. And it allowed explanation without ever for a moment requiring one to invoke more difficult or complex psychological processes.

The explanation was simple. No ''expectation'' of reward was required. The subject didn't have to expect the reward or anticipate or represent its occurrence in any fashion at all. The subject simply responded to the eliciting stimulus with the associated habit. No hedonic impact, no affective or emotional state of any sort, was posited. The subject didn't need to have a satisfaction or any other affective feeling at all. Instead the strengthening of association could be done directly. A three-neuron circuit would suffice. Simply let a neuron representing ''food occurred'' potentiate the connection between a neuron carrying ''stimulus'' to a neuron generating ''response.'' Complicate this circuit further with millions of additional neurons, as real-life brains certainly do, there still appeared no need for more than the three basic functions. For example, many of the basic features of stimulus-response reinforcement appear to be captured by an extremely basic neural arc circuit in the spinal cord, which is capable of elementary operant conditioning of a motor stretch reflex (Wolpaw, 1997).

One might imagine that the S-R habit explanation of reward learning described above is the sort of explanation that only a behaviorist could love. It seems to leave out every sort of psychological process that we experience ourselves regarding reward and substitutes rigid and reflexive S-R habits that are entirely alien to the notion of rewards. But there was actually *evidence* to support the S-R reinforcement theory. A vivid example is the early ''kerplunk'' experiment by the archetype S-R theorist John Watson (Carr & Watson, 1908). In this experiment, rats were trained to run in a maze for a food reward. As the rats learned the maze very well, they began to run very quickly through each length and turn. They became routine and accomplished performers on the task. Then, one day, the length of the runway was changed: either lengthened or shortened. The question was, what would the rats do on their first trial in the longer or shorter runway? The answer was, if the new length were longer, the rats ran smoothly and quickly for their customary distance—as far as they used to in the old runway. Then they paused—even though they had not yet reached the end. If they had reached the spot where food would have been before, they stopped and sniffed at that spot, even if that meant ignoring food further away at the new end. It was as though the rats had learned simply

to run through a motor habit: a certain number of steps, performed almost reflexively once triggered at the beginning. Conversely, what if the runway were suddenly shortened rather than lengthened, so that the food and wall were moved closer to the starting place? On their first trial in the shorter runway, the rats ran full speed—right over the nearer food—and right into the wall at the end, ignoring the clear warning sight of it in front of their faces—kerplunk! Again, it appears that the rats' well-learned performance had become a routine motor habit, a series of running steps linked together as a pure S-R routine.

C. Hullian S-R Reinforcement and Drive Reduction

Findings such as the "kerplunk" phenomenon gave credibility to S-R interpretations that rewards worked simply by strengthening the tendency to emit a particular habitual motor action when in the presence of a particular stimulus. But it was not immediately apparent why different rewards (food, water, drug, or access to a sex partner, social partner, etc.) all should strengthen habits. What did they have in common? Nor was it clear how an individual could learn long sequences of responses, which were unlikely to originally be emitted by chance, and so could not easily be strengthened all at once. Nor again was it clear how individuals learned different responses to obtain different goals. How is it that one learned to go to a drinking fountain for a cold drink when thirsty, but to a food dish when hungry? If these were simply habits, why did't they get mixed up? Why didn't one equally drink when hungry or eat when thirsty?

Hullian learning theory, and similar formulations by other psychologists of the time, represented a principled attempt to explain such complexities within the constraints of an S-R framework (Hull, 1943). First of all, Hull specified the hypothetical property supposed to be shared by all reinforcers that gave them the ability to reinforce behavior: they all *reduced drive*. When we are hungry, Hull argued, we are in a state of drive that activates us. Food, when eaten, reduces that hunger drive. It was the reduction of drive, Hull argued, that was the source of reinforcement. Drive reduction, in essence, was reward. Similarly, when we have been a long time without a drink, or if we are out in the hot sun, thirst activates drive. A sip of water reduces the drive of thirst, and so becomes a reinforcer. Many of the physical stimuli that serve as rewards (food, water, drugs for individuals in drug withdrawal) can be conceived as drive reducers. Social rewards are perhaps a bit more tricky to construe as reducers of drive, but if one grants the possibility of a sexual drive, then the act of engaging in sex can be viewed as possibly reducing that drive. Similarly, if a mother has a maternal drive, then presentation of her infant might reduce that drive by providing

an outlet. If individuals have a social drive to mingle and interact with others, then access to social partners or social approval might in turn reduce the relevant drives. If these were all allowed to be considered as drives, then drive reduction was at least in principle a parsimonious way to account for a wide variety of reinforcers. Behavior was finally a simple consequence of the level of drive at the moment and the ability of stimuli present to trigger S-R habits, which had previously been created by drive reduction. This simple relationship could be expressed as an equation:

$$_{S}E_{R} = {}_{S}H_{R} \times D,$$

where $_{S}E_{R}$ was the actual strength of a behavioral response in the presence of a particular stimulus, $_{S}H_{R}$ was the associative strength of the learned habit within the individual, and D was the level of drive at the moment. Hull added other factors, too, but this equation captures the essence of his model.

Drive reduction, if it was the mechanism for S-R reinforcement, also provided an associationist way to explain why we seek food when hungry but water when thirsty. The answer was the sensations that accompany hunger or thirst were themselves part of the stimulus in a particular S-R link. If a rat felt hungry and perceived a food lever in an operant chamber, then pressing the bar in the presence of that compound stimulus composed of the see-lever-and-feel-hungry combination would produce food, and result in reduction of the drive. But if the rat felt thirsty instead of hungry and saw the same food lever, then the food which a press would produce would *fail* to reduce its thirst drive. Pressing the food lever after seeing the bar while feeling a parched throat, and so on, would be a habit that would not get strengthened. Thus hunger and thirst were not merely drives, but were also cue sensations or discriminative stimuli in themselves, which could combine with drives to determine whether a habit would be triggered. It was the entire stimulus combination that triggered the S-R habit, according to Hullian theory. Specific drives acted as discriminative stimuli, as well as to energize behavior, which served to gate the ability of external stimuli to trigger habits that were related to a particular drive but not to others.

Finally, the Hullian account of reward learning offered one more mechanism to help explain how complex *chains* of responses become learned. The mechanism was Pavlovian associations among pairs of stimuli, and the explanation was that classical conditioning allowed stimuli encountered when distant from the goal to trigger incremental or partial responses. After all, most behavior is not a single act and therefore cannot be reinforced as a simple habit. Behavior is a stream of acts. Complex habits grow out of smaller ones. Even a simple rat that has learned to walk down a T-maze

runway and turn the corner to find food must learn a sequence of responses, not just one. Hullian learning theory suggested that rats did learn a sequence: specifically they learned the sequence backwards, starting with what to do at the food dish, and building gradually back to the beginning of the maze.

For Hull, a hungry rat on its first time in a T-maze must wander aimlessly. It does not know what to do. But eventually, if it wanders enough, it will find the food dish and food within it. If it eats the food, it will have experienced an S-S pairing: the sight of the food dish, followed by the stimulus of food itself. In the parlance of Pavlovian conditioning, the sight of the food dish is a conditioned stimulus (CS). The food itself is an unconditioned stimulus (UCS). So there has been a CS-UCS pairing, and an association between the two stimuli is formed. The next time the rat is put in the maze, it still does not know what to do. But if it wanders down to the end of the runway, it may see the food dish around the corner. Originally the food dish was meaningless: just another sight to see in the maze. But now the sight of the dish is a Pavlovian CS for food, which can elicit at least a weak food-appropriate response. Normally the rat, if it saw real food, would walk toward the food and then eat it. It cannot eat the food dish, especially not from a distance, but it can walk toward the dish just as it would walk toward the food itself. And this is what the rat should do by Hullian theory: emit to the CS whatever fractions it can of the response that it would ordinarily emit to real food. Such *fractional responses* will most often include approach, for various rewards, and so the rat directly approaches the dish. Once it reaches the dish, it can eat the real food, and when it does that, the entire preceding set of responses will be reinforced by drive reduction. The rat has just learned to walk toward the dish as soon as it is seen.

To modern psychologists, the Hullian account of reward learning seems strained, especially given the plausibility of cognitive alternatives. For example, the learning of a sequential maze could be explained by positing that the rat has formed a cognitive representation of the spatial location of food, of its own spatial location, and of the contours of the maze. An alternative explanation is that the rat finds its way to food by following this cognitive map. Several authors have provided excellent discussions of the relation between such cognitive processes and the simpler S-R processes conceived by Hullian learning theory and of evidence relevant to both (Dickinson, 1989; Dickinson & Balleine, 1994; Holland, 1993; Toates, 1986, 1998). My focus here is specifically on the *reward* learning aspect of Hullian theory, that is, the strengthening of associations by *drive reduction,* rather than on the cognitive versus S-R contents of what is learned. I thus wish to restrict our present consideration to the reward concepts of reinforcement and drive reduction versus alternative concepts of reward. Of what

use are they in understanding reward learning, and what evidence has led to the replacement of reinforcement concepts with new concepts of reward? That will be the topic for the rest of this chapter.

D. Role of Drive Reduction Theory in Human Motivation and Social Psychology

The notion that rewards work via mechanisms of reinforcement and that reinforcement is gated by drives proved to have long-lasting consequences in psychology. Reverberations can be found in fields far removed from S-R behaviorism, such as in human social and motivational psychology (McClelland, 1987). Although no longer conceived in terms of S-R habits, of course, reward is sometimes still discussed as working via reinforcement states or goal states that follow performance of the motivated behavior. For example, in a major text on human motivation McClelland (1987) adopted an almost Thorndikian notion of hedonic reinforcement (though he allowed the reinforced response to be as abstract as power seeking) when he wrote,

> This method also led to a working definition of a motive as a recurrent concern for a goal state based on a natural incentive—a concern that energizes, orients, and selects behavior. . . . The fact that the concern is about a goal state has the important implication that the means of getting to the goal is not part of the definition of a motive. The goal state may be defined as the outcome of certain acts. (McClelland, 1987, pp. 590, 591)

McClelland went further in conceiving the reinforcement mechanism to be gated by drive and discussed motives, such as the power motive, as though they were a form of drive. For example, regarding the power motive, he speculates,

> It might be inferred that brain norepinephrine represents a physiological reward system . . . specifically for the power motive. To put it in everyday language, suppose brain norepinephrine turnover . . . represents the extent to which power stimuli ''turn on: subjects, or make them ''feel good''. Clearly, the pictures used in this experiment would have more turn-on value for some subjects than others. So these results mean that subjects who are high in Power will learn power-related materials faster if the materials have turn-on value for them than if they do not; in particular, they will learn the power-related materials faster than subjects low in n Power, for whom the stimuli have no turn on value. The situation is analogous to putting hungry rats in a maze with food at the end that they either like or do not like. The rats who are hungry and like the food will learn fastest, and those who are not hungry and dislike the food will learn slowest. (McClelland, 1987, pp. 279–280)

McClelland's description of how the power motive might interact with reward seems similar to reinforcement and drive concepts we have seen.

Although McClelland's view of human motivational psychology is different in many ways from early learning theories, it shares in common at least a reinforcement mechanism. In this case reinforcement acts through a hedonic state a bit like Thorndike's original satisfying state, which is "turned-on" by appropriate outcomes for individuals who are considered high in a power motive drive state.

E. Role of Drive Reduction Theory in Physiological Psychology

The Hullian idea that reward consisted of reducing drives, and that reward learning was the learning of responses that accomplished drive reduction, was tremendously powerful as well in physiological psychology and behavioral neuroscience. The physiological study of motivation became largely the study of stimuli and brain mechanisms of deficit states, such as blood glucose deficits, liver cues, and neuronal energy metabolism for hunger drive and for hormonal and other physiological stimuli for thirst, sex, aggression, and other drives.

This drive reduction orientation began to change slowly with the discovery by Olds and Milner of electrical brain stimulation reward (Olds & Milner, 1954), the phenomenon in which both animals and humans would learn a response and work in order to deliver depolarizing electrical stimulation via implanted electrodes in their lateral hypothalamus, nucleus accumbens, or associated sites. Brain stimulation reward apparently did not need a special drive to be effective, and James Olds wrote of "pleasure centers in the brain" in terms that were not strictly limited to drive reduction. But it was still possible to conceive of brain stimulation reward as reducing natural ongoing drive states. After all, there is always one drive or another that can be reduced.

Given the predominance of drive reduction theories, this interpretation seemed tenable even after it was discovered that brain stimulation also could *evoke motivated behavior* such as eating. According to drive reduction theory, it was at first expected that sites where stimulation would reduce eating (presumedly by reducing drive) would also be sites where stimulation was rewarding (again presumedly by reducing drive). Conversely, this view led investigators to expect, as Olds wrote later, that an "electrical simulation which caused the animal to respond as if it were very hungry might have been a drive-inducing stimulus and might therefore have been expected to have aversive properties" (Olds, 1973, p. 89). The hope that drive reduction would prevail as the explanation of reward was beautifully captured in retrospect by Neal Miller as he described his own attempts to find the neural basis of drive reduction reward, guided by drive reduction theory:

> If I could find an area of the brain where electrical stimulation had the other properties of normal hunger, would sudden termination of that stimulation function as a reward? If I could find such an area, perhaps recording from it would provide a way of measuring hunger which would allow me to see the effects of a small nibble of food that is large enough to serve as a reward but not large enough to produce complete satiation. Would such a nibble produce a prompt, appreciable reduction in hunger, as demanded by the drive-reduction hypothesis? (Miller, 1973, pp. 54–55)

In other words, if one found a ''hunger drive'' neuron, then one would have found not only hunger but also reward. Increases in the neuron's activity, caused by stimulation, would produce hunger drive and eating behavior. Decreases in the same neuron's activity, caused by a nibble of food, would produce drive reduction and psychological reward. Reductions in such a neuron's activity would be tantamount to reward itself, and would be the neural basis for reward learning regarding food.

Drive reduction was simple and understandable as a theory of reinforcement. Its parsimony made it attractive. Drive reduction *explanations* of reward were seriously damaged, however, on the rocks of unexpected experimental findings that accumulated fast and furious during the late 1960s and early 1970s. A major destructive finding came from observations in brain stimulation reward and electrode-evoked hunger studies themselves. Contrary to expectations, it turned out that the brain sites where stimulation caused eating behavior were almost always the same sites where stimulation was rewarding (Valenstein, 1976; Valenstein, Cox, & Kakolewski, 1970). Rather than be opposite, the two effects were identical, or at least, had identical causes. The realization that drive and reinforcement might reflect the same state, rather than opposites, was a major shock to physiological psychology (and continues to cause aftershocks today). It led in part to the rise of alternative hedonic frameworks of affective neuroscience (for example, Panksepp, 1998).

The reinforcement concept is by no means dead yet in contemporary physiological psychology or behavioral neuroscience. Many behavioral neuroscience investigators still invoke a Thorndikian reinforcement mechanism as a key concept in their accounts of reward learning (Kelley, 1999; McFarland & Ettenberg, 1998; Rolls, 1999; White, 1989). But the *explanation* of reinforcement by *drive reduction* has largely been abandoned, due to the failure of experimental results to confirm the expectations described by Olds and by Miller. Instead, the reinforcement term is used typically today in behavioral neuroscience either in an implicit hedonic sense, or simply as a term to label a behavioral reward effect without requiring any effort by the author to specify a precise mechanism or explanation. Any reward given after successful performance is often called a reinforcer, regardless of the psychological mechanism through which it works to influence behavior.

Despite its longevity, reinforcement may not after all be a very valuable concept for understanding reward. Reinforcement has been largely replaced as an explanatory concept, or at least greatly diminished in importance, by alternative *incentive* concepts of motivation that better explain reward phenomena. Quite a number of motivational theorists in biopsychology have argued that incentive concepts may be much more useful than reinforcement concepts for understanding most reward processes and most motivated behavior (Berridge & Robinson, 1998; Bindra, 1978; Blackburn, Pfaus, & Phillips, 1992; Bolles, 1972; Depue & Collins, 1999; Dickinson & Balleine, 1994; Flaherty, 1996; Panksepp, 1998; Phillips, Blaha, Pfaus, & Blackburn, 1992; Stewart, de Wit, & Eikelboom, 1984; Toates, 1986; Toates, 1994; Tomie, 1996). As they have been quite successful in this argument, in my opinion, I shall focus on incentive processes for the remainder of this chapter.

II. Bases of Modern Incentive Theory: Bolles–Bindra–Toates

A major shift from drive reduction and reinforcement or ''stamping in'' theories began to occur in the early 1970s, as incentive theories of motivation began to rise. Most important to understanding the biopsychology of reward and motivation may be a particular type of incentive mechanism that I will call the Bolles–Bindra–Toates theory, in recognition of the three psychologists who elaborated its basic tenets, Robert Bolles, Dalbir Bindra, and Frederick Toates (Bindra, 1974, 1978; Bolles, 1972; Toates, 1986, 1994). It is worth mentioning that current biopsychological incentive theory also owes much to the work of many other psychologists of learning and motivation, for example, J. Konorski, P. T. Young, R. L. Solomon, R. A. Rescorla, and A. J. Dickinson (Dickinson & Balleine, 1995; Dickinson et al., in press; Konorski, 1967; Rescorla & Solomon, 1967c; Young, 1966). But I think it fair to say that the specific theory of reward and incentive motivation discussed below was developed chiefly by these three theorists.

A. Bolles: S-S Reward, Not S-R Reinforcement

A major attack on drive reduction, together with an alternative approach, was mounted by Robert Bolles in 1972 (Bolles, 1972). Bolles surveyed a number of phenomena, which had been identified in the late 1960s and early 1970s, that defied explanation in terms of drive reduction. These were, first, operant *stereotypies* of behavior that developed in operant conditioning, sometimes called ''misbehavior'' (Breland & Breland, 1961). For example, a raccoon that was being trained to put a coin-like object into the slot

of a piggy bank for food reward obstinately refused to release the coin object from its paw when it had slid into the slot, but instead pulled it out again and made ''washing'' movements with the coin, dipping it and wiping it again, as though the coin were food. Second, *autoshaping* of what looked like operant behavior under conditions where there was actually no reinforcement of responses by drive reduction. For example, a pigeon that received presentations of a signal light followed by free food reward, after many pairings came to peck the signal light whenever it was turned on, even though it had never been explicitly reinforced to do so. The pigeon had ''shaped itself,'' hence the term *autoshaping*. Third, bizarre ''adjunct'' behaviors occurred, such as *schedule-induced polydipsia,* in which animals drank exorbitantly large amounts of water for no apparent reason when put into a training schedule. For example, a hungry rat presented with tiny food rewards interspersed with delays of a minute or so begins to engage in copious amounts of drinking behavior, and may consume up to one-half of its body weight in water in a mere 3-hr session (Falk, 1971).

Bolles argued that these phenomena demonstrated that reward learning could not operate by the response reinforcement and drive reduction principles that had reigned for nearly half a century. First of all, the phenomena were happening in the absence of response reinforcement. Reinforcement, whether drive reduction or hedonic satisfaction or something else, had not stamped in a misbehavior/autoshaped/adjunct response to a stimulus, since none of these responses had been explicitly reinforced by the experimenter. The animals had instead learned something else, and reinforcers were causing some other kind of process to occur. Bolles argued that the animals had learned S-S* associations, that is, associations between a CS (S) and a subsequent hedonic stimulus that caused pleasure (S*). The first S did not elicit a response, he argued. Instead it elicited an expectation, an expectation specifically of the second S*. The animals had also learned an R-S* association, Bolles suggested, namely that their own response was a predictor of the emotionally laden S*. At the moment they engaged in these strange behaviors, such as treating a keylight or a coin as though it were food, he suggested, they did so because they were seized by an overwhelming expectation of food itself (Bolles, 1972).

Bolles's S-S* expectancy account provided the seed of an explanation of why animals sometimes behaved as though they had reward when they did not yet have it. But expectancy by itself was not a full explanation. It is not clear why an individual who expects food should engage in bizarre behavior. Why not simply sit back and enjoy the expectancy, knowing that the real food is about to arrive? More explanation was needed to account for the behavioral phenomena.

B. Bindra: A Conditioned Stimulus Becomes the Incentive

In 1978, Dalbir Bindra provided an extension of the incentive account that helped fill the explanatory gap (Bindra, 1974, 1978). He adopted the S-S* framework of Bolles, but sidestepped the idea that it was an expectancy alone that caused irrational behavior. He suggested instead a specific motivational process was evoked. The CS or S, he suggested, does not simply cause the subject to expect S*, but also evokes a *central motivational state of the hedonic S**, which in turn causes the animal to *perceive* the S as an S*, to the extent that perception could be supported by the stimulus properties of the signal stimulus. The S, in other words, takes on specific motivational properties that normally belong to the S* itself. These motivational properties are incentive properties: they attract the animal and elicit goal-directed behavior and even consumption.

By Bindra's account, food is normally an attractive incentive. Animals seek it out. They eat it if they can. If they are raccoons, they may wash it before they eat it. When they do eat it, they enjoy it. The incentive properties of food allow it to produce all these effects. Bindra made the startling claim that the CS—the S—gained the same properties as a function of its association with the S*. An animal, he suggested, approaches the CS for a reward. It finds the signal attractive. If the CS is food, the animal wants to eat it. If it is an S for a tasty food S*, the animal may enjoy its attempt to eat the CS.

Bindra's assertion that a CS in a sense becomes an incentive to the animal, the same incentive as the S* stimulus, is rather starlingly unintuitive. It seems almost a step back from Bolles's expectancy account. It asserts that signals are taking on motivational and hedonic properties as a consequence of their Pavlovian association with rewards. This is not entirely in congruence with our own sense of how we respond to motives—drive reduction was almost easier to accept by comparison. But what if Bindra were right? The raccoon washes its coin, according to this, because the coin has become ''food-like'' to the animal. The pigeon pecks the keylight signal for food in autoshaping because the light has become a glowing piece of food. Suddenly there is a psychological explanation for these phenomena.

Still at face value, Bindra could not possibly be right, at least not unless something more is added. Critics noted that if CSs simply became incentives, then one would always respond to them, whether hungry or thirsty or not (Gallistel, 1978). There was not a clear way to explain how incentives and CSs would interact with drive states. Yet clearly drive state is important, and, in addition, true expectancy may exist and guide behavior in its own way.

C. Toates: Conditioned Incentives, Hedonic Modulation, and Drive Cues

In a book published in 1986, Frederick Toates adopted the Bolles–Bindra notions of S-S* associations and hedonic transfer of incentive properties to the CS, yet in a new synthesis escaped the rigid predictive difficulties (Toates, 1986). First, he posited that both cognitive expectancy and more basic reward processes might occur simultaneously within an individual, and influence behavior in different ways. Then he addressed how basic reward processes could guide normal behavior via incentive cues and also grab control to produce even irrational behavior.

1. Alliesthesia: Changing Hedonic Value

Toates began with the idea that hedonic incentives were the target of motivation. Hedonic incentives were stimuli that directly produced experiences that felt good. Tasty food, refreshing drinks, sexual partners, drugs that caused pleasant feeling, social rewards were all hedonic incentives. Toates also drew on the work of the psychologist and physiologist Michel Cabanac (Cabanac, 1979, 1992), to note that the ability of these incentives to produce pleasure was themselves in many cases modulated by drive states. Physiological consequences of eating, etc., were posited by Toates to feed back to influence motivation—but not directly. Rather, for Toates, drive states modulate the value of incentive stimuli. Pleasure is not an invariant property of even tasty food, Cabanac had shown. A sweet taste may be delicious when we are hungry; it may be ambrosia if we are starving; yet if we are stuffed after large meal, it may retain only a ghost of its pleasure producing properties. Our palate can become jaded, and no longer take pleasure from what it once did. In laboratory demonstrations, for example, Cabanac showed that human subjects gave higher subjective ratings of pleasure to a sugar solution when they were hungry than when they had recently eaten. The pleasure of the sensation had changed with their physiological state, even though the sensory quality of the sweetness was no more or less than it was before. Cabanac called this changed hedonic quality of the sensory experience ''alliesthesia,'' which means essentially a change in sensation (though to be fully accurate, the phenomenon is a change only in the *pleasure* of the sensation), and argued that alliesthesia is a basic property of most hedonic sensations. A hot bath feels delightful if we are cold, but may seem positively unpleasant on a hot day—when a cold plunge into cool pool seems much more pleasant. In another example, the saltiness of seawater is unpleasantly intense to most individuals. But in a phenomenon known as salt appetite (Schulkin, 1991), the taste of salt

becomes more pleasant if we lose our normal body levels of sodium chloride, either by going on an entirely salt-free diet for several weeks or by taking drugs that fool physiological mechanisms into reducing sodium stores (Beauchamp, Bertino, Burke, & Engelman, 1990).

Alliesthesia shifts in the hedonic impact of stimuli such as sweet or salty tastes can be detected in animals or in human infants too, even without the use of subjective ratings of pleasure, by measuring affective reactions. Certain behavioral affective reactions made by human infants, such as facial expression, reflect the hedonic impact of a taste stimulus (Berridge, 2000; Steiner, 1979; Steiner, Glaser, Hawilo, & Berridge, 1999). If a sweet taste is placed in the mouth of a human infant, a chimpanzee, or a monkey, certain hedonic reaction patterns are elicited, such as rhythmic tongue protrusions or finger sucking (Figure 1). Even in a rat, a sweet taste elicits behavioral affective reactions (e.g., rhythmic tongue protrusions, licking of lips and paws). Alliesthesia can be observed in the change in these hedonic reactions to sweetness; they are increased by hunger and decreased after a meal (Berridge, 1991; Cabanac & Lafrance, 1990). Similarly, the salty

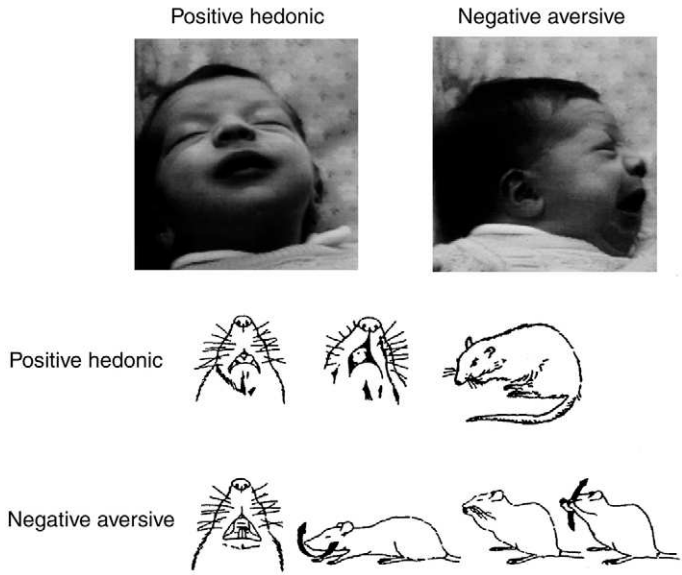


Fig. 1. Hedonic impact or "liking" reflected in behavioral affective reactions elicited by tastes from human infant and by rats. Positive hedonic reaction of a 3-week-old infant to a sweet sucrose solution versus negative aversive reaction to an unpleasantly salty solution is shown at top. Drawing at bottom shows positive hedonic and negative aversive reactions elicited by tastes from rats. (From Berridge, 1996.)

taste of seawater elicits hedonic reactions during physiological states of sodium depletion, but elicits a bitter-like aversive response (e.g., gapes, arm flails, headshakes) during states of normal sodium balance (Berridge, Flynn, Schulkin, & Grill, 1984). Alliesthesia, or shifts in the hedonic impact of incentive stimuli caused by changed physiological states, may thus be a common occurrence in the lives of both humans and other animals.

III. Consequences for the Incentive Model of Motivation and Reward

Toates drew upon phenomena such as alliesthesia to modify the Bolles–Bindra theory, and to allow physiological drive states to play a role. Hunger, thirst, and other drive states, he posited, served to potentiate the hedonic value of primary rewards. Rather than produce reinforcement by being lessened, as drive reduction hypotheses had always assumed, drives interacted with incentive stimuli such as the taste of food to multiply or potentiate the hedonic value of food or of another incentive stimulus. Food tastes better during hunger, water is more delicious during thirst, saltiness becomes more pleasant during sodium deficiency, and so on. Hedonic modulation by physiological drives made incentive stimuli more attractive and so more able to control behavior in a Bindra–Bolles fashion.

But the multiplicative interaction between drives (e.g., hunger) and incentive stimuli (a tasty morsel) works both ways, according to Toates. Not only do drives such as hunger potentiate the incentive value of food, but the reverse is also true. An external incentive stimulus, such as the presentation of a morsel of food to an individual who is hardly at all deprived, can potentiate appetite for more food as effectively as an increase in physiological need. One bite of food can seemingly intensify hunger in an individual who moments before was not at all thinking of dinner. The French expression puts it, *''l'appetit vient en mangeant''* (loosely translated, appetite comes in the act of eating). You may have experienced this as the ''cocktail peanut'' phenomenon: after taking one tidbit without desire and merely to be polite, you suddenly find you want to eat a few more. Cornell, Rodin, and Weingarten (1989) found that human subjects who had been fed to satiety, if asked to take a bite of either pizza or ice cream, would subsequently choose to eat more of whichever of these two foods they had just been given.

In animal laboratory settings the ability of a reward stimulus to increase motivation for itself is known as priming. A rat, for example, may seem reluctant to begin work when placed at the beginning of the day's session back in the chamber where lever presses will earn food, but it can often

be suddenly energized by simply giving it one free food pellet. After ingesting its free reward, the rat is likely to show renewed interest in working for more food. Priming is also well known in studies of brain stimulation reward and of addictive drug reward, where animals earn activation of the stimulation electrode or a small intravenous pulse of rewarding drug by performing a response. Responding may be low in the beginning of a session—unless a free reward is given or until the first reward is finally earned—and then the animal sets to work. This is very much in keeping with the Toates model of incentive motivation.

And how does reward learning play a role according to this Toatesian scheme incentive–drive interaction? Precisely as Bolles and Bindra had specified: CS for primary rewards, such as the sight of a signal for food, acquire the incentive properties of their primary UCS, such as the taste of food, via Pavlovian associations produced by their pairings. So, for example, the mere presentation of an auditory CS that has previously been paired with food reward causes rats that are not physiologically hungry to begin eating again (Weingarten, 1983; Weingarten & Martin, 1989). Priming of appetite can thus be done by a Pavlovian cue as well as by the actual taste of food. Presentation of an auditory CS for sucrose reward appears to elicit a *hedonic* state appropriate to sweetness too (Figure 2). For example, if a tone that has been paired with sucrose is played to a rat while an infusion of water is squirted into its mouth, the rat increases its number of positive hedonic reactions (e.g., lip licking) (Delamater, LoLordo, & Berridge, 1986). Conversely, if a different tone that has been paired with quinine is presented during the oral infusion of water, the rat suppresses its positive hedonic reaction patterns and instead increases its aversive reaction patterns (e.g., gapes).

Toates added the additional stipulation that the incentive properties of such conditioned reward stimuli were modulated by drive states such as hunger in exactly the same way as the incentive properties of food itself. In effect, Toates argued, a CS light that predicts food to a hungry pigeon or rat becomes an attractive, potentially edible, and possibly even tasty food-like object. It elicits approach and even consummatory behavior that would ordinarily be directed to the food itself. But when no longer hungry, the same light for the animal is just a predictive signal, devoid of motivational properties.

The associative link that allows a CS to acquire hedonic value from its unconditioned partner can be quite powerful, according to Toates, powerful enough to overcome any existing hedonic value the signal stimulus may already have. An example of this is taste aversion conditioning. In taste aversion conditioning, a normally palatable food is paired associatively with an unpleasant physiological consequence such as nausea and related upper

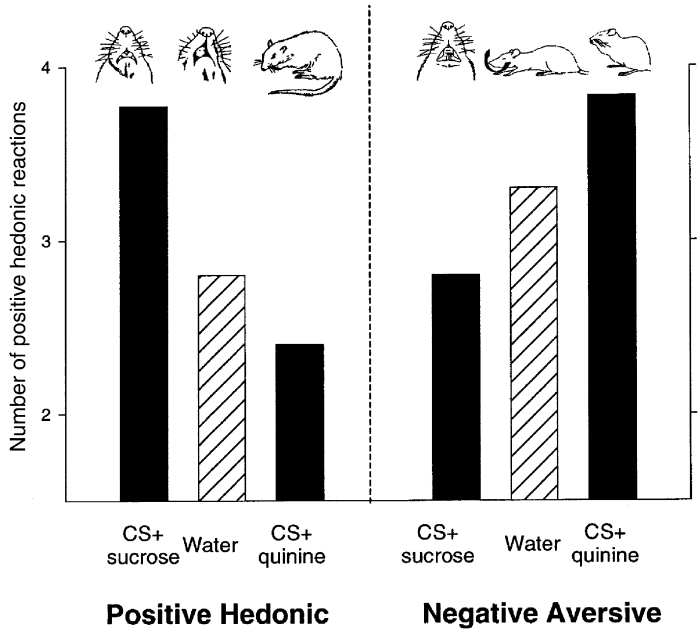


Fig. 2. Conditioned stimuli (CS) elicit hedonic states. The affective reaction pattern elicited from rats by the taste of water is modulated by two tone CS that were either paired with the sweet taste of sucrose (CS+ sucrose) or the bitter taste of quinine (CS+ quinine). Positive hedonic reactions were increased by the sucrose cue and decreased by the quinine cue. Negative aversive reactions were conversely increased by the quinine cue but decreased by the sucrose cue. Each auditory CS appeared to evoke an affective state similar to that of the sweet/bitter taste it signalled. (Figure modified from Delamater et al., 1986.)

gastrointestinal distress of the sort produced by food poisoning or by a drug (or excessive alcohol) that mimics aspects of such illness (Garcia, McGowan, Ervin, & Koelling, 1968; Rozin, 1999). The first time the food is encountered, the illness has not yet occurred. The individual enjoys the palatable food and finds the experience of consuming it pleasant. Then the illness follows, which is of course unpleasant. The taste aversion is not revealed, however, until the food is once again encountered. The important point is that the taste signal does not simply call up a representation of UCS state of illness. It may or may not call up an explicit re-representation of the nausea state. But the taste itself is always perceived as reduced in pleasantness after aversion conditioning—even if one doesn't realize why. The same taste that before seemed so pleasant may evoke quite a different reaction and be perceived as disgusting. It simply tastes bad. Animals too, if they have had a pairing with drug-induced ills after ingesting sweet taste

that evoked positive affective reactions, switch their reaction when the taste is next presented: now they respond with aversive reactions as though the taste were bitter (Berridge, Grill, & Norgren, 1981; Grill & Norgren, 1978). Taste aversion learning obeys all the rules of Pavlovian learning and can be strengthened, weakened, or even prevented by proper application of those rules. For example, one can protect against taste aversion learning via Pavlovian *latent inhibition,* which means that a very familiar CS, one that has been experienced many times before, will resist entering into the new Pavlovian association. In other words, if a food is already familiar when the illness first occurs, then a conditioned aversion is not likely to follow. That is why, if several foods are eaten at a meal hours before a visceral illness occurs, but only one of the foods is new, then the individual may afterwards intensely dislike that new food but not feel differently about the other foods (Rozin, 1999).

The power of S-S* associations was dramatically illustrated in a special demonstration of taste aversion conditioning by Peter Holland, in which an auditory tone S (Pavlovian CS) was made to literally stand in for its food S* (Holland, 1990). In one experiment, Holland presented one tone to rats as they drank a wintergreen-flavored sucrose solution and a different tone as they drank a peppermint sucrose solution. Then later, he presented *one* of the tones by itself (without any sucrose or flavor) to the rats as they were made ill by lithium chloride (injection of LiCl causes nausea). Finally, the rats were allowed to choose between the peppermint sucrose and the wintergreen sucrose (without any tones present). Holland found that the rats drank considerably less of the particular solution whose tone had previously been paired with illness. They had developed a conditioned aversion for the flavor, apparently mediated by the auditory CS. In this case, no food was paired with illness, only its CS tone had been paired. Holland concluded that the tone had evoked a sensory representation of its taste—a taste memory—when it had been paired with the illness, and that the rats had learned an aversion to the *remembered* taste, called to mind by the tone. The CS had fully substituted for its taste in the learning of this new aversive reward value: after S-S* associations, the S can stand in for the S*.

Toates's use of S-S* associations in formulating an extension of Bindra's theory of incentive motivation is remarkable, for it entails several implications, some of which are quite unintuitive (Figure 3A). The theory is all the more interesting and important if these implications are true.

First it means, as we have said already, that the incentive value and even hedonic value of stimuli rises and falls with changes in relevant physiological deficit or drive states. This should be true regarding food reward, drink reward, sex reward, drug reward, brain stimulation reward, and perhaps

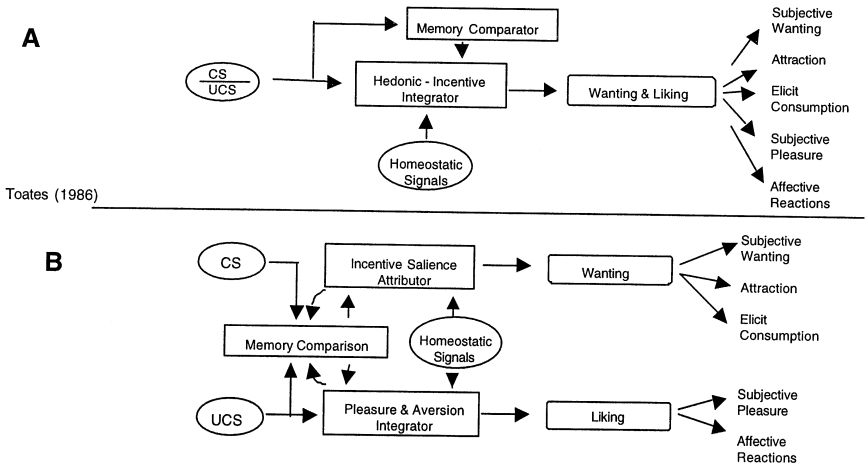


Fig. 3. (A) The Bindra–Toates model of conditioned incentive motivation in which "wanting" and "liking" are identical, based on Toates (1986). (B) Incentive salience model or modified Bindra–Toates model in which "liking" and "wanting" for rewards and Pavlovian reward cues are separate, from Berridge (1996; modified from Robinson & Berridge, 1993). Some brain manipulations, such as impairment of dopamine systems, disrupt Pavlovian incentive salience, selectively suppressing "wanting" for a food reward (but leaving "liking" alone). Other brain manipulations, such as microinjection of amphetamine into the nucleus accumbens, selectively potentiates the ability of a Pavlovian conditioned stimulus (CS) to elicit "wanting" for its food reward (without potentiating "liking" for that food reward). In human drug addiction, too, neural systems of incentive salience may be selectively sensitized, leading addicts to "want" to take drugs even if they don't particularly "like" them. (From Robinson & Berridge, 1993.) UCS = unconditioned stimulus (such as food reward or drug reward).

many other types of reward. A bite of sweetness is not fixed in its hedonic value, and neither is a given dose of cocaine. Both may shift predictably according to this model.

Second, it means that drive states will themselves be often potentiated by a sudden encounter with a relevant incentive or even with a CS for that incentive. The smell of fresh-baked bread or of sizzling meat are potent triggers of appetite as mealtime approaches. Even the mere sight of the food or a realistic image may make the mouth water and arouse appetite.

Third, Toates suggested, the multiplicative outcome of the interaction between external incentive stimuli and internal physiological cues for drive states is not only sufficient to trigger motivation but may be *necessary* for substantial motivation to occur. In a multiplicative relationship, there is no output if either element is zero. Incentive stimuli and drive cues both need the other to create a motivational state. If either category is totally missing, then even relatively strong intensities of the other category may have little

effect on motivation or behavior. In other words, drive reduction, which was so important to earlier theories of reinforcement, is almost powerless by itself to cause reward. Could this counterintuitive assertion possibly be true? The answer appears to be yes.

A. Drive Reduction: Insufficient for Strong Reward

In retrospect, evidence to support Toates can be found going back decades. In an early demonstration of the importance of simultaneous co-occurrence of oral incentive stimuli together with drive reduction, Miller and Kessen (1952) trained rats to run in a T maze for one of two possible rewards. One reward was intragastric milk that was pumped directly into the stomach, which reduced hunger drive alone. The other reward was the chance to lap milk normally, which both reduced drive and provided an oral incentive stimulus comprising taste/tactile/consummatory act components. This experiment was predicated on the demonstration by many earlier studies that rats would run faster down a runway to obtain a larger reward (e.g., more food pellets) than for a smaller reward, and similarly would run faster for any food that they preferred. In other words, running speed during a test was proportional to the ''goodness'' of the reward that had been given during training, and better rewards produce faster running. The rats approached the ''pure drive reduction'' goal only slowly, essentially walking hesitantly toward the intragastric reward. Intragastric milk appeared to be only a very weak reward, as the rats would choose it over nothing, but never eagerly. By contrast, rats ran swiftly toward the natural milk drink reward, which combined oral incentive plus drive reduction, even though the same amount of milk ended up in their stomachs in both cases.

In a similar demonstration, McFarland showed that drive reduction was not a sufficient reward for thirst (McFarland, 1969). McFarland trained doves to peck an illuminated keylight in order to receive a sip of water to drink. Once the doves had learned this task, McFarland switched them to a new task, which came in one of two versions. In one version, the second task was the same as the first: pecking on a new key light simply earned again the chance to drink. In the other version of the new task, pecking on the new key light earned a direct infusion of water into the dove's crop (a pouch within the esophagus where doves normally store small quantities of water or food before passing it on to the stomach). The amount of water was the same in both cases, and thus provided about the same degree of physiological need reduction. McFarland found that pigeons who were switched to the natural drinking task simply maintained their performance at high rates. They immediately learned that the new key provided water reward, just as the old key had, and they kept pecking at it. By contrast,

the doves who were switched to the crop infusion of water initially pecked a number of times at the new key, just as they had at the old key, but then gradually ceased to peck at all. Even though they received just as much water into their bodies as before, the water was no longer able to reinforce or maintain their behavior. It appeared no longer to be a reward. The results of McFarland's switched task experiment are important because they show that drive reduction by itself is not only insufficient to train a powerful new level of motivated behavior, but that drive reduction alone is also an insufficient reward to maintain motivation even in a rewarded task that has already been well learned (McFarland, 1969).

B. Drive: Insufficient to Cause Motivation

The Bindra/Bolles/Toates theory posits that motivation as well as reward is incentive based, not drive-based. Just as drive reduction by itself turns out to be insufficient for reward, so physiological drive by itself should not be necessary or sufficient to cause motivation. In an early demonstration that physiological drive was unnecessary for motivation, Solomon and Stellar and their colleagues asked whether appetite for food would be suppressed by intravenous feeding of the average daily intake of calories and nutrients (Turner, Solomon, Stellar, & Wampler, 1975). They provided dogs with intravenous meals, solutions of glucose, and other nutrients delivered directly into the bloodstream through a vein catheter, giving as many calories as the dogs would ordinarily eat in a day. The question was whether this intravenous feeding would satisfy their appetites. The answer was no: if the dogs were given the chance to eat ordinary food, they ate almost as much as they would have ordinarily without the extra calories. As a result of the combined intravenous feeding and voluntary meals, the dogs became quickly overweight. But still they continued to eat.

In a more sophisticated demonstration of this principle, Bedard and Weingarten (1989) precisely parceled out the role of oral incentive cues versus physiological drive reduction in turning off appetite, and found an interactive effect precisely as Toates would have predicted. Bedard and Weingarten used a sham feeding paradigm, in which rats were first implanted with a gastric fistula—a permanent tube into their stomach. If the gastric fistula were open, then what the rat ate or drank would leak out of the tube after it were swallowed—hence the name sham feeding. On the other hand, the open gastric fistula also allowed Bedard and Weingarten to give an intragastric meal directly, by simply passing liquefied food through the tube into the stomach. Bedard and Weingarten took hungry rats and simply asked, what was the crucial event that made them stop being hungry? Was it the physiological effects of food in the stomach and

its absorption into the body? If so, then a gastric infusion of liquid food ought to suffice to satisfy the hungry rats. Or was it the opportunity to experience the oral stimuli of biting, tasting, chewing, and swallowing the food? If so, then sham feeding, with the gastric fistula open, ought to suppress a subsequent meal even though the food from the first meal had leaked out of the stomach. Or was a Toatesian interaction required between physiological and incentive stimuli? The answer was that neither oral stimulation nor drive reduction alone were able to suppress appetite. Bedard and Weingarten found that appetite for a second meal was suppressed only if rats first received the combination of sham feeding and a nutrient infusion (or the chance to simply eat an ordinary meal first, which also combined both oral and physiological aspects of eating). Similarly, it is difficult to satiate thirst by drive reduction alone. Thirsty rats continue to drink even after they are given intragastric infusions of water unless they are also allowed to experience oral stimulation related to the act of licking and drinking (Rowland & Nicolaidis, 1976). According to Toates's view, motivated behavior exists only when both types of external incentive and internal drive stimuli are present. Motivation is terminated only when the individual's requirements for both are satisfied. And reward learning, of the sort needed to produce vigorous behavior directed toward that reward in the future, occurs only when both are delivered interactively to the individual. Though counterintuitive at first glance, Toates's notion of multiplicative interaction between incentive and drive stimuli appears to be crucial in order to properly understand basic incentive motivation and reward.

IV. Do Conditioned Incentive Stimuli Take on the Motivational Properties of Primary Incentive Stimuli?

Let's consider a few other surprising implications of the Bolles/Bindra/Toates incentive process. Toates's theory makes the startling claim that when a cue, or CS, is associatively paired with a primary hedonic reward (tasty food, pleasant drug), the result is not only *learning* of a Pavlovian association (and perhaps a cognitive expectation too) but also a *motivational* transformation. The claim is that, to the individual who has learned the association, the cue actually takes on a number of the incentive motivational properties originally possessed only by that primary hedonic event. The conditioned cue becomes attractive and valuable to the individual, in the same way that the primary reward is valuable, and may even elicit some of the same hedonic impact, or pleasure, elicited by the unconditioned event. The individual may try to interact with the cue in some of the ways normally reserved for interaction with the primary hedonic reward. If the

primary reward were tasty food, for example, then after Pavlovian conditioning, the hungry individual might find the cue alone to be somewhat tantalizing, pleasant, and even perhaps edible.

An edible cue seems to stretch credibility, but cases of edible cues produced by classical conditioning appear to exist. We have encountered one already—autoshaping—which now deserves a closer look.

A. Autoshaping

Autoshaping takes its name from a superficial similarity to operant conditioning or shaping, in which an animal's response such as lever pressing or key pecking is gradually increased by an experimenter who awards reinforcement only after the animal shows an approximation of that response. But in reality autoshaping is a pure incentive process with no response reinforcement at all. It is caused by Pavlovian associations, rather than operant contingencies. It creates an attractive and even edible Bindra–Toates-style conditioned incentive out of a mere Pavlovian cue stimulus.

In autoshaping, originally discovered by Brown and Jenkins (1968) and by Williams and Williams (1969), a pigeon or rat is placed in an operant chamber where there is a signal light above a lever to press (for rats) or a signal light embedded in a key or button to peck (for pigeons). The signal light is occasionally illuminated and followed by presentation of food reward. In other words, the signal is a Pavlovian CS for food. The lever can be pressed (or key pecked for pigeon subjects) as the animal chooses. The crucial feature that distinguishes autoshaping from ordinary shaping is that in autoshaping there is absolutely no contingency between pressing the lever or pecking the key and food reward. Food and its CS signal are presented intermittently regardless of whether or not the animal performs the instrumental response. Under these conditions there is no rational reason for the animal to work. There is no response reinforcement. There is no shaping because food reward is not selectively given when the animal performs the response. The reward is free. Yet, amazingly, under these conditions, as a trial continues, pigeons begin to peck robustly at the key when it is illuminated, and rats begin reliably to press the lever (often biting the light itself as well).

When autoshaping was first discovered, early attempts were made to explain it in traditional reinforcement terms, by suggesting that perhaps reinforcement was occurring fortuitously. If the animal had simply happened to peck the key or push the lever by chance before the food was delivered, reinforcement would occur accidentally, and if they responded again, they would be again accidentally reinforced by another food reward, even though there was no outside contingency between their response

and the reinforcer. In effect, the animals would have created their own response–reinforcer contingency. Skinner had many years earlier advanced a similar explanation of superstitious behavior, that is, particular behavioral patterns that some animals developed when put into an operant chamber, even though the experimenter did not intend to reinforce those behaviors (Skinner, 1948).

A moderate difficulty for the accidental reinforcement explanation is that *any* response should be equally likely to be strengthened, whereas in autoshaping the response is usually pecking or biting of the cue or lever pressing. Clearly something else is needed to explain the directedness of this behavior, something that draws upon the predictive relation between the cue and its particular reward. Staddon and Simmelhag (Staddon & Simmelhag, 1970) offered the first explanation that drew upon this predictive, Pavlovian relationship. They suggested that the cue light became a Pavlovian CS for food or water and elicited a few Pavlovian CRs, similar to the responses that ordinary food or water would elicit as UCs. Then once a few Pavlovian CRs had been predictably elicited by the cue, they suggested, those reward-appropriate responses were particularly and especially reinforced in ordinary fashion by the arrival of a food or water reward.

But even this degree of reinforcement-based explanation was demolished by later experiments that showed autoshaped subjects apparently did not need any kind of response reinforcement, even accidental reinforcement, to develop autoshaping (Schwartz & Williams, 1972). This was shown by adding a *negative* contingency between responses and reward (an omission contingency) to the autoshaping procedure. If the animal responded by pecking the keylight then food was withheld and not delivered, even if the cue had just occurred. If the animal did not respond when the cue was illuminated, then the food reward came as usual. The animal ought to learn to simply sit and wait for signaled food, or at least to refrain from approaching the lever or key, under these conditions. But instead it was found that autoshaping still developed—albeit somewhat more slowly—under an omission contingency. The subjects continued to develop and maintain approach and consummatory-like responses directed toward the food cue—even though it cost them a real food reward every time they did so. This rules out a reinforcement explanation.

B. Consumption of Conditioned Stimuli

Further observations supported yet another aspect of the Bolles/Bindra/Toates interpretation, that is, that CSs take on the natural incentive properties of the reward they represent. It was discovered that autoshaped subjects often attempt to "eat" their reward's CS. Rats, for instance, may gnaw away

at the plastic cover that protects the cue lightbulb, sometimes destroying the cue entirely. Pigeons have a distinctive movement pattern of the beak used in feeding pecks, in which they bring the top and bottom halves of the beak rapidly together as they encounter their normal food, such as a speck of grain. In autoshaping, the same feeding peck is directed towards the keylight itself if it has signaled food. By comparison, pigeons that have been autoshaped while thirsty using a cue-water association show a drinking peck, a different movement pattern in which the lower beak is extended longer to scoop up water. Thus in a literal sense, animals appear sometimes to try to eat or drink their CS cue light, depending on whether their primary hedonic reward was food or water. They treat the CS object as though it were edible food or potable water, and try to consume it in the appropriate fashion (Allan & Zeigler, 1994; Jenkins & Moore, 1973).

Autoshaped eating of a keylight is certainly consistent with the proposition that CSs for a reward take on the properties of that reward, and are treated almost as though they *were* the reward. But an alternative interpretation was also possible, one also based on Pavlovian sensorimotor conditioning. It is conceivable that what has been associated with the cue is not so much an incentive motivational status of food as it is the sensorimotor response pattern that is usually elicited by food. We have already mentioned that the autoshaped keypeck uses a movement pattern very similar to the movement used in real eating or drinking. It has long been known that motor responses, or simple movement patterns, can easily be conditioned by Pavlovian training. Eyeblink movements, for example, elicited by a wind-like puff of air, are also elicited by a tone CS for the airpuff. Could autoshaping too be merely a conditioned motor response? If so, the phenomenon would be of much less motivational interest.

C. Pure Incentive Motivation Elicited by Pavlovian Cues

Fortunately there are ways to show that the sensorimotor conditioning explantation is insufficient, and that a CS indeed takes on motivational properties as the incentive conditioning model suggests. One procedure that has been used to accomplish this uses separate Pavlovian pairings outside of the instrumental training situation, which are then used to influence a different instrumental response that involves a different movement from any Pavlovian CR. This procedure has been given various names, such as the conditioned emotional response or Pavlovian instrumental transfer paradigm. It essentially measures the effect of a Pavlovian *incentive cue* upon motivation reflected through instrumental performance to obtain a reward.

For example, Lovibond (Lovibond, 1983) trained rabbits to perform an instrumental response for sucrose reward and then separately trained them

in a Pavlovian cue–reward association. Instrumentally, the rabbits learned to push a lever to obtain a sucrose squirt when they were hungry. Quite separately, they got free sucrose during training of a Pavlovian cue, a clicking sound. The sound came on for 10 sec, and then the sucrose solution was delivered. No response was necessary during this Pavlovian portion of the training, and the rabbits had no opportunity to press the lever during the cue training. Finally, the lever was returned, and the rabbits were allowed to work again to obtain the sucrose. Now the clicker cue was occasionally presented. In this case, the clicker could elicit no conditioned motor response relative to the lever, but it could still elicit an incentive motivational state. The incentive state could be expressed only through the instrumental lever-pressing response, not through any Pavlovian motor response. And indeed, the rabbits pressed the lever that had earned sucrose more during the period when the clicker cue was presented than during the period when there was no Pavlovian cue. In other words, the cue had clearly potentiated instrumental responding, even though the motor response needed to perform the instrumental task was unrelated to any Pavlovian motor responses. This indicates that a heightened motivation for the reward was truly produced by the cue. (Aversive motivation may work in a similar way: if a Pavlovian cue is paired with shock, that cue can increase the level of instrumental performance of animals that are performing a response to avoid the shock (Rescorla & Lolordo, 1965). In such cases, the cue elicits what has been called a conditioned *emotion,* such as fear of shock or desire for sucrose, that is appropriate to its paired UCS.

This demonstrates that the Pavlovian cue can elicit a motivational state. But does it show that the cue itself is the target of the motivational state? Does the individual merely want the unconditioned reward? Or does it want the CS as well as the UCS, as Toates's model would predict. That is, does the cue's association cause it to become an incentive itself, as well as being able to potentiate incentive motivation for the unconditioned reward?

D. Sensory Preconditioning of Incentive Value

Evidence that the cue itself actually becomes a target for incentive motivation comes from an experimental procedure called *sensory preconditioning*. In sensory preconditioning, a neutral cue is first paired with another neutral cue. Neither stimulus has motivational value at this first stage. They are simply presented together again and again, so that the subject learns an association between them. Then one of the neutral stimuli is given new motivational value. The motivational value may be given by ordinary Pavlovian conditioning, by pairing it with a reward or with a punishment. Alternatively, one of the neutral stimuli may be given motivational value

by performing a physiological manipulation upon the individual that is relevant to that particular stimulus.

For example, in a salt appetite experiment using sensory preconditioning Fudim (1978) combined pairs of mixed flavors for rats: for some rats, banana flavor was combined with salt into a salty banana mix, and almond flavor was combined with sucrose to make a sweet almond mix. For other rats, the pairs were salty almond and sweet banana. Then Fudim induced a physiological state of sodium deficiency in the rats, by giving them an injection of formalin (which causes sodium-containing body fluids to move out of their normal physiological compartments, producing a physiological sodium deficiency). The sodium deficiency would be expected to produce a salt appetite in which salty tastes would be sought out (Schulkin, 1991). While the rats were sodium-deficient, Fudim presented them with a drinking tube that contained pure banana flavor by itself and with another tube that contained pure almond-flavored solution. No salt or sugar was now present in either tube; instead, there was merely a flavor that had previously been paired with salt or sugar. Impressively (from the point of view of the Bindra–Toates model), the sodium-deficient rats drank much more of the banana or almond solution that had been paired with salt than they did of the flavor that had been paired with sucrose. When the rats were not sodium-deficient, on the other hand, they did not show this preference to consume the salt-paired solution. In other words, the drive state caused the rats to seek out and actually ingest the CS itself—precisely as the Bindra/Toates model would suggest.

E. Sensory Preconditioning of Hedonic Value

One last objection could be raised at this point to the Toates–Bindra interpretation. Perhaps the rats in Fudim's experiment did not like the CS flavor they drank when sodium depleted, but rather drank it solely for its association with postingestive effects of salt. The rat might drink the flavor in a way similar to why you might drink an unpleasant tasting medicine: not because you like the medicine's flavor, but because you hope to feel better afterwards. The Toates–Bindra model specifies that conditioned incentives actually become *liked*.

In order to find out whether the Bindra–Toates prediction was literally true, and whether the CS actually gained *hedonic value* itself, Jay Schulkin and I modified Fudim's sensory preconditioning experiment and added an additional test of the flavor's hedonic impact: the taste reactivity measure of hedonic and aversive patterns of behavioral reactions that are elicited by tastes (Berridge & Schulkin, 1989). Similar to Fudim's experiments, our rats first received two flavors that were paired either with salt or with

sucrose. We used bitter-tasting quinine and sour-tasting citric acid as our CS flavors. Bitter and sour were chosen as cues because we wanted to be sure that the CS flavors had no positive hedonic value to begin with (so to be able to tell if positive hedonic value were suddenly acquired). Rats ordinarily respond to bitter or sour tastes with strongly aversive reactions (gapes, etc.). Some of our rats received a salty bitter mixture and a sweet sour mixture. Other rats received a salty sour mixture and a bittersweet mixture for conditioning of the taste–taste association. Later the rats were made to develop a salt appetite by injecting them with furosemide (a drug that disrupts the kidney's hormonal regulation of physiological sodium balance). Once they were sodium deficient, the rats were allowed to drink if they wished either a sour citric acid solution or a bitter quinine solution—but no salt was present in either solution. As Fudim (1978) had found, our rats drank more of the CS flavor that had been paired with salt, and only when they were sodium deficient. Finally, we squirted into the mouths of the rats (through oral cannulae) either the purely sour solution, or the purely bitter solution, and measured their hedonic and aversive affective reactions. Ordinarily the sour or bitter tastes elicited mostly aversive reactions, such as gapes, headshakes, and forelimb flails and this remained true when the rats were tested in a normal physiological state. But when the rats were in a sodium-deficient state, the sour or bitter CS flavor elicited positive hedonic reactions, such as tongue protrusions and paw licking (Figure 3). Only the salt-paired conditioned flavor elicited hedonic reactions, and only when the rats were sodium deficient. The other bitter or sour flavor (which had not been paired with salt) elicited only aversive reactions at all times (Figure 4). In other words, sodium-deficient rats actually ''like'' a sour or bitter taste if the flavor is a CS for saltiness and if they have a physiological salt appetite at the moment. The rats not only wanted to consume these flavors—they also ''liked'' them, as Toates' model suggests—even though they had never ''liked'' concentrated salt itself when they had tasted it before in the taste mixtures (because they had not had a salt appetite then). An integration of associative and physiological information had occurred, passing the alliesthesia increase in hedonic palatability directly to the CS, and giving it new incentive and hedonic value.

V. Splitting Pavlovian Incentives: "Liking" versus "Wanting"

The Bindra–Toates model suggests that Pavlovian incentives become both ''liked'' and ''wanted'' as a consequence of reward learning. Conditioned incentive value is equivalent to conditioned hedonic value according to the original Bindra–Toates model (Toates, 1986). Individuals literally move

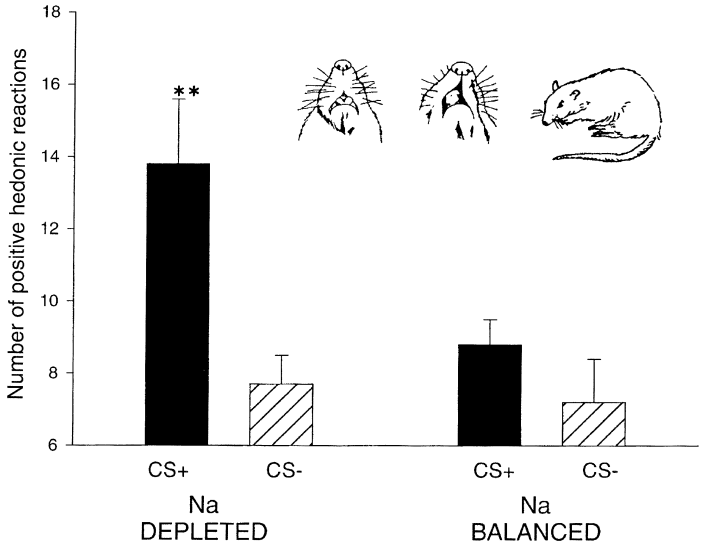


Fig. 4. Alliesthesia of a sour–bitter conditioned stimulus (CS) in salt (NaCl) appetite. Hedonic reactions patterns (lip and paw licking) elicited by a sour (citric acid) or bitter (quinine) taste is modulated by Pavlovian conditioning and physiological deficit state of sodium depletion. The CS+ is the sour/bitter taste that was previously paired with the taste of concentrated NaCl. The CS− is the other sour/bitter taste that was paired with sucrose. Citric acid was the CS+ for some rats, whereas quinine was the CS+ for other rats. The sour/bitter CS+ taste specifically elicited many more hedonic reactions than the CS− when the rats had a salt appetite induced by physiological sodium depletion (NaCl depleted). Neither taste elicited hedonic reactions when the rats were in a normal physiological state of sodium balance. (Figure modified from Berridge & Schulkin, 1989.)

along a gradient of conditioned hedonic stimuli as they get closer to the goal. But recent evidence has led my colleagues and I to suggest there are two modifications that need to be attached to this conclusion. The result is what I have called a modified Bindra–Toates model or the incentive salience model (Figure 3B; Berridge, 1996; Berridge & Robinson, 1998; Berridge & Valenstein, 1991; Robinson & Berridge, 1993).

The first issue concerns how the core processes of incentive value that we've been discussing relate to subjective experience, that is, to conscious pleasure and desire. To many, the term *reward* is often used as a surrogate for conscious pleasure, that is, something that is consciously liked, and indeed one dictionary definition of reward is "a pleasant stimulus" (p. 2584, *Shorter Oxford Dictionary*)(Brown, 1993). But the psychological reality is more complex than the dictionary would suggest. Conscious pleasure may not always be required for reward phenomena. First, a number of studies of

humans indicate that hedonic affective states and the motivational processes triggered by both Pavlovian incentives and unconditioned incentives can sometimes be produced *without* conscious awareness: for example, people may be trained to perform responses for a 'pleasant' drug they are not aware of receiving, or may be made to have an emotional reaction by an event that they do not consciously notice (Fischman & Foltin, 1992; Lamb et al., 1991; Winkielman, Zajonc, & Schwarz, 1997). In other words, unconscious pleasures and desires may exist as unconscious psychological core processes (Berridge, 1999). If so, we need to be able to distinguish conscious pleasures from unconscious core processes of reward. My colleagues and I have adopted the terms ''liking''—in quotation marks—to refer to unconscious core processes underlying conscious liking, since the core process is not really liking in the usual sense. Similarly, we've used the term ''wanting'' in quotation marks to refer to the core process of desire elicited by an incentive, which may occur even unconsciously. It is outside the scope of this chapter to say much more about unconscious core processes of reward or about their relationship to real conscious pleasures and desires, but I refer the interested reader to another chapter that focuses on unconscious affective core processes (Berridge, 1999).

Second and more relevant to the Bindra–Toates model of Pavlovian incentive learning is a *split* that may occur between the incentive processes of ''liking'' and ''wanting.'' Typically these processes go together, and both obey Bindra–Toates rules. But in a number of affective neuroscience experiments my colleagues and I have found that in some instances ''liking'' and ''wanting'' separate dramatically, and we have therefore proposed that the Bindra–Toates model of incentive motivation be modified to allow separate control of these two incentive processes (Berridge, 1996; Berridge & Valenstein, 1991; Berridge, Venier, & Robinson, 1989; Robinson & Berridge, 1993). Separation of ''wanting'' from ''liking'' happens especially in cases involving certain neural or psychopharmacological manipulations that can occur in animal experiments and even in real-life human predicaments, such as drug addiction. In such cases, incentive motivation can reflect Bindra–Toates rules applied to ''wanting'' alone (Figure 3B).

A. ''Liking'' without ''Wanting''

A number of studies from the past 10 years, mostly involving brain manipulations, have shown that it is possible to have ''liking'' without ''wanting'' for an incentive, or vice versa, ''wanting'' without ''liking.'' For example, ''liking'' without ''wanting'' appears to be the result of extensive damage to brain dopamine systems. Such damage leaves individuals apparently without motivation for any incentive, conditioned or unconditioned (Fibi-

ger & Phillips, 1986; Marshall, Richardson, & Teitelbaum, 1974; Ungerstedt, 1971), but ''liking'' for the same incentives remains normal, at least for the hedonic impact of food rewards. For example, in studies of taste-elicited affective reactions, Robinson and I have found that positive hedonic reaction patterns to sweet tastes remain essentially normal in animals that have lost nearly all of their mesolimbic dopamine neurons that project from midbrain to forebrain structures (Berridge & Robinson, 1998; Berridge et al., 1989).

B. ''Wanting'' without ''Liking'': Pure Pavlovian Incentive Salience

Conversely, ''wanting'' without ''liking'' can be produced by several manipulations in the laboratory and probably in real human life. In the laboratory, electrical stimulation of the lateral hypothalamus in rats triggers a number of motivated behaviors, such as eating. Increased appetite is normally accompanied by increased hedonic appreciation of food. This ''alliesthesia'' or hedonic enhancement is also caused by many other manipulations that cause eating, such as drugs and conditioned appetites. But Elliot Valenstein and I found eating caused by electrical stimulation of the lateral hypothalamus was not accompanied by enhanced hedonic reactions to the taste of food. If anything, rats responded more aversively to a sweet taste, as though it became bitter, when their hypothalamus was stimulated.

Similarly, Cindy Wyvell, working in our laboratory, has recently found that a related ''wanting'' without ''liking'' can be triggered by microinjections of amphetamine that activate dopamine neurons in the nucleus accumbens (Figure 5). The presentation of a Pavlovian CS for food causes rats to work even harder than normal in order to obtain food. Microinjection of amphetamine into the accumbens specifically potentiates the conditioned incentive motivation properties of the food cue and increases the ability of the cue to trigger ''wanting'' for the food reward. But the enhanced ''wanting'' for food is not accompanied by increased ''liking,'' as evidenced by the microinjection's failure to increase positive hedonic patterns of behavior elicited from the rats by a sweet taste (Wyvell & Berridge, 2000). Accumbens amphetamine magnifies the conditioned incentive ability of a food cue to trigger ''wanting'' for food, but has no effect on ''liking'' for food.

A related ''wanting'' without ''liking'' phenomenon for a food UCS itself, very similar to the effects of hypothalamic stimulation, has been found recently by Sheila Reynolds in our laboratory to be produced by accumbens microinjections of muscimol, a drug that stimulates γ-aminobutyric acid (GABA) receptors. GABA neurons in the accumbens receive dopamine inputs. These neurons may be the next link in the neural chain of ''wanting,''

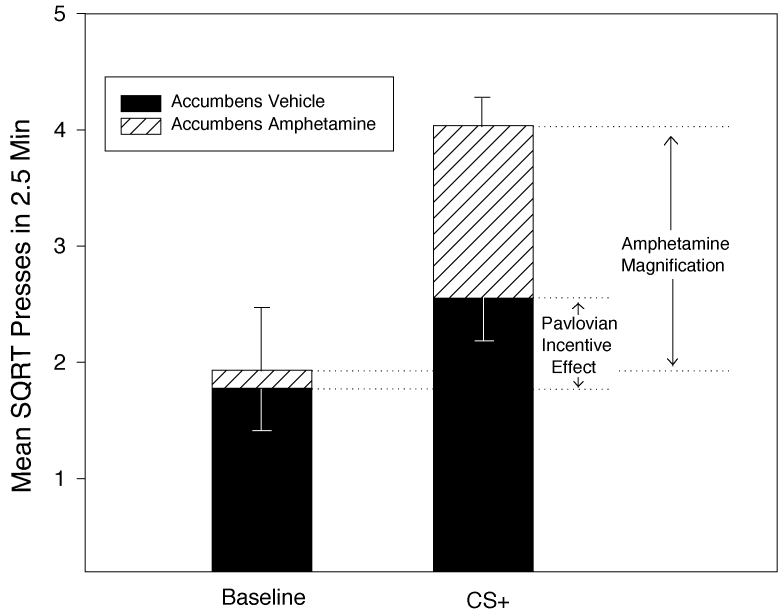


Fig. 5. Magnification of a Pavlovian food cue's incentive properties by a microinjection of amphetamine into the nucleus accumbens (20-μg dose). Bar pressing by rats on a bar that previously earned them sucrose pellets (measured under extinction conditions when no food pellets were actually delivered). Presentation of a Pavlovian conditioned stimulus (CS) for food reward increases bar pressing for the reward. Facilitation of mesolimbic dopamine neurotransmission markedly magnifies the incentive salience attributed to the Pavlovian reward cue, thus increasing its ability to trigger "wanting" for the food reward (without much affecting baseline rate of pressing in the absence of the food cue). (From Wyvell & Berridge, 2000.)

as modulation of these GABA neurons has been suggested to be a shared output mechanism for all mesolimbic reward (Carlezon & Wise, 1996). Muscimol microinjections in the accumbens are known to cause rats to eat intensely, much as LH stimulation does (Stratford & Kelley, 1997). Yet even though this drug microinjection makes the rats avidly seek and consume food, it does not at all increase their hedonic affective reactions elicited by a sweet taste (Reynolds & Berridge, 2000). The GABA agonist in the accumbens makes rats "want" food—and so they eat avidly—but does not make them "like" any more the food they "want."

Finally, human drug addicts apparently may often crave drugs even when they don't derive much pleasure from them. For example, some drugs, such as nicotine, generally fail to produce substantial pleasure at all in most people, but can still be quite addictive. There is reason to believe that

the same dopamine-accumbens neural system may become sensitized or hyperresponsive to drugs and CSs in the brains of drug addicts (Robinson & Berridge, 1993). If so, Terry Robinson and I have suggested that this may cause heightened incentive salience to be attributed to drug cues, causing addicts to ''want'' to take drugs even if they don't particularly ''like'' the drugs.

Incentive salience attribution, operating by Bindra–Toates rules, may thus produce ''irrational'' pursuit of goals and sometimes cause pathological behavior in human daily life as well as in animal laboratories. Incentive salience in short is the property of a perceived reward stimulus that makes it attractive, attention grabbing, ''wanted,'' and a target for goal-directed strategies (Berridge, 1996; Berridge & Robinson, 1998; Robinson & Berridge, 1993). When attributed to a specific stimulus, incentive salience can make an autoshaping cue light appear ''food-like'' to the autoshaped pigeon or rat that perceives it. When attributed to the smell emanating from a bakery, incentive salience can rivet a person's attention and trigger sudden thoughts of lunch. When attributed in an excessive or sensitized fashion to the sight of drug paraphernalia or drug-associated settings, incentive salience can trigger sudden and compulsive ''wanting'' in an abstinent heroin addict to take the drug again. Or if attributed to the sight of small white chunks of powder, incentive salience may lead crack cocaine addicts to rummage desperately about on the floor ''chasing ghosts''—even if they know cognitively that the white chunks they find there are more likely to be table sugar than cocaine—a not-quite-rational act more than a little reminiscent of autoshaping in animals.

Addiction is a pathological case, but incentive salience attribution operates by Bindra–Toates rules, and is crucial to normal reward learning. Reward in the full sense cannot happen without it. ''Liking'' by itself is not true reward any more than is ''wanting'' by itself: nothing need be reward*ed* by hedonic activation alone. Pleasure is not itself goal directed or necessarily associated with objects or events. Pleasure by itself is simply a *triggered affective state*—there need be no object of desire. It is the process of incentive salience attribution that makes a specific associated stimulus or action the object of desire. ''Liking'' and ''wanting'' are needed together for full reward (Figure 4).

VI. From Pavlovian Incentives to Incentive Expectations: Dickinson–Balleine

My discussion has stressed the evidence for the modified Bindra–Toates model of incentive motivation and explored at some length its implications.

By this model, incentive motivation is guided by Pavlovian cues for rewards. The cues become attributed with incentive salience (and sometimes with hedonic properties as well), and so the Pavlovian reward cues become beacons, triggers for motivation, enablers for the arousing effects of physiological drive states, and almost partial substitutes for the rewards they represent.

I have emphasized the evidence for this Pavlovian incentive process in part because it is counterintuitive that the incentive salience of Pavlovian reward cues, rather than drive reduction learning or cognitive expectation, is a chief mechanism for motivation and reward learning. The conclusion is unlikely to be accepted without emphasis and evidence. But evidence such as that reviewed above demonstrates that incentive salience attribution to Pavlovian reward cues is a very powerful mechanism of reward learning in animals, and may be also be influential aspects of human reward and motivation (e.g., addiction).

Still, incentive salience mechanisms operating by Bindra–Toates rules do not work alone. Cognitive expectation of a reward for goal-oriented action is surely another important mechanism for human incentive motivation, and quite possibly for animals too, and is far more accessible to the inner eye of human introspection. Cognitive expectation of a future reward is the form of incentive learning that comes naturally to mind for one who is asked, What does it mean to work for an incentive? The commonsense or *rational* meaning is that the incentive is known in advance, that the incentive and its hedonic value is held explicitly in mind as a declarative cognitive representation by the seeker during the work, and that the action is coordinated specifically in order to attain the hoped for goal. Incentive motivation in this highly cognitive sense clearly happens in us many times each day. Cognitive incentive learning may happen in animals also, but has been less tractable to verification or experimental analysis in animals for the obvious reason that we have no access to animal introspection and must rely on nonverbal tests of cognition. Yet the study of cognitive forms of incentive motivation can still be approached indirectly in animals, and thus its basic mechanisms can be probed with rigorous experimental tools. Foremost in attempts to examine the psychological properties and brain substrates of cognitive incentive learning in animals has been the efforts of Anthony Dickinson and his colleagues, especially Bernard Balleine.

A. Cognitive Expectation of Incentive Value Based on Past Hedonic Experience

Building on a scheme sketched originally by Tolman (Tolman, 1949), Dickinson (1989) defined cognitive incentive learning by several features that

can be demonstrated in goal-directed behavior. He and his colleagues have argued that if one learns cognitively how to gain a particular incentive, then this must entail understanding of the causal relationship between working (as a cause) and gain of the incentive reward (as a consequence). In other words, in the cognitive sense of incentive motivation, one knows that one's effort can produce the incentive. Dickinson uses the term *''act–outcome representation''* to denote cognitive understanding of the causal relationship. Once one has learned to represent the act–outcome relationship between a particular action and the incentive it produces, then it may be the case that when one engages specifically in that incentive-related action, one does so with the *expectation* of earning the reward. Whether this description is true or not can be tested experimentally by probing whether the individual takes into account new information regarding the value of the reward. An experimenter can ascertain under what conditions expected reward value equals the reward's true value.

A cognitive expectation of reward, according to Dickinson and Balleine (1995), is what gives the representation of a stimulus *incentive value* in the everyday and fullest sense of the term. The representation of a stimulus has these properties if ''instrumental behavior is mediated not only by a representation of the action-outcome relation, but also by a representation of the incentive value of the outcome, or what in common parlance would be referred to as the desire for the outcome'' (Dickinson & Balleine, 1995, p. 163). This is a cognitive sense of incentive value, not a Pavlovian or Bindra–Toates sense, and is more complex than incentive salience ''wanting.''

Dickinson's work and related studies have shown that animals seem to have the rational sense of incentive learning too. Even rats can be shown to expect particular outcomes, and to govern their instrumental performance based on their expectation. One remarkable thing is that the explicit cognitive expectation sense of incentive learning (responding with cognitive expectation to gain the reward) and the Pavlovian incentive salience process (directly ''wanting'' and pursuing the conditioned cues of reward) both seem to exist simultaneously. They usually guide behavior together in the same direction: towards the goal. But under certain conditions they guide behavior in different directions, and in those situations the two processes can be pitted against each other to better study them. Sometimes one wins, sometimes the other. Surprisingly, Dickinson and colleagues have shown that the rational sense of incentive learning is in one way *stupider* than the Pavlovian form of incentive learning. Or rather, to put a more positive gloss on it, the cognitive mechanism is rational: it expects an incentive always to be as good as it has been in the past. Sometimes this rational expectation is wrong. It is wrong chiefly when the hedonic value of the

incentive is suddenly changed. Under those conditions, the Pavlovian mechanism may have an advantage.

B. Pulling Apart Cognitive Expectations from Pavlovian Incentive Salience

In a typical experiment, Dickinson and Balleine and colleagues have assessed whether an animal has a cognitive representation of a reward by asking whether it can act appropriately when it is suddenly placed in a new motivational state. For example, if a rat had learned while hungry to perform two instrumental responses, one for sucrose solution and another for food pellets it might be tested in a state of thirst or in a state of caloric satiety, or after one food had been associatively paired (in a different setting) with LiCl illness to induce a taste aversion. Any of these manipulations would change the relative hedonic values of the foods. Two features of the new state test are particularly important in these experiments in order to reveal the role of the cognitive act–outcome expectation. First, the rat is tested in extinction, so that operant responses no longer earn the actual food rewards. The rat must therefore choose to work based solely on its *representation* or expectation of the rewards, because the food reward itself does not arrive, and the rat never gets to *experience* its altered hedonic value. In many experiments, the rat's expectation of a food's hedonic value appears to be based directly on the explicit hedonic experiences it has had of the food in the past. Second, Dickinson and Balleine take pains to arrange the experimental situation so that potential Pavlovian CSs predict both foods equally well. This obviates control of behavior by Pavlovian-related motivational processes, which otherwise could guide behavior to one goal or the other (and which would be more related to the sense of incentive salience and to the Bindra–Toates sense of incentive learning).

Under these conditions, Dickinson and Balleine have shown that rats quite often fail to appropriately modify their instrumental behavior for a newly revalued food, when the rats are forced to act on the basis of their expectations. Instead they continue mistakenly to expect the food to have the same hedonic value it had before. If a food had high hedonic value when it was experienced during training (e.g., while hungry), then the rat often continues to work hard in the test—even if the food would now have reduced hedonic value (e.g., either because now the rat is sated or because food has subsequently received one conditioned aversion pairing with LiCl) (Balleine & Dickinson, 1991; Dickinson & Balleine, 1994). If a rat had not worked vigorously before, because the food did not have great value during training, it still does not work harder during the test, even though the food would now have enhanced hedonic value (e.g., because physiological

deprivation is now higher, or because a benzodiazepine drug has been given that increases food palatability) (Balleine, 1992; Balleine, Ball, & Dickinson, 1994; Dickinson & Balleine, 1994). The rat's expectation of reward apparently is that the reward will be the same as before—even though now it won't. In these situations a rat can be said to have a cognitive expectation of reward in one sense, but the expectation is an incorrect one. The rat has not yet fully learned about the new hedonic value of the food (its cognitive representation is not updated), and its behavior is not guided by an accurate cognitive representation of the food at that moment.

Dickinson and colleagues have found a crucial way in which a rat can be provided with information sufficient to correct its mistake, allowing it to change its expectation-driven instrumental behavior appropriately to the new incentive value of the food or water reward. What the rat needs is to actually *experience the new hedonic value* of the reward *while it is in the changed state* of thirst, hunger, drug-state, or after a single taste-LiCl pairing (Balleine & Dickinson, 1998; Dickinson & Balleine, 1994). This would ordinarily happen naturally if the test were not conducted under extinction conditions. Then the rat would simply sample both rewards, quickly experience which one was better in the new state, and modify its behavior accordingly. But it is also sufficient, Dickinson and Balleine show, simply to allow a rat to experience one taste of the new hedonic value of the food anyplace else, such as in its home cage (as long as the rat is in the same physiological state as to be tested later). If the rat has experienced the new hedonic value prior to the instrumental extinction test, it can later employ the now known value of the food in the cognitive expectation it has of the food's value, and of the food's causal relationship to the two different actions. It then modifies its instrumental response in the extinction test appropriate to the revalued food. Other instrumental responses, used to obtain another reward that was not revalued, are not altered. For Dickinson and colleagues, this selective and intelligent change in behavior demonstrates true incentive learning in the common sense or rational meaning of the phrase. It demonstrates that a rat knows in a cognitive sense what it is working for and how to get it.

C. Semipermeable Boundary between Cognitive Expectation and Pavlovian Incentive Salience

The distinction described above between cognitive expectation of incentive value and Pavlovian incentive salience of CS has emerged quite clearly in many experiments conducted by Dickinson and his colleagues. Cognitive expectation appears to be based squarely on past experience of the incentive value. Incentive salience attributed to Pavlovian CS, on the other hand,

follows Bindra–Toates rules and can integrate new physiological states with associative information even in advance of explicit experience of a new hedonic value.

However, it should probably be acknowledged that the division between the two forms of control is not absolute. In a number of other experiments, leakage between the processes seems to have occurred, and integration in advance has been observed even in behavior apparently guided by the cognitive mechanism of act–outcome representations (Rescorla, 1994a, 1994b; Shipley & Colwill, 1996). For example, in contrast to the findings of Dickinson and his colleagues (Balleine & Dickinson, 1991, 1992; Lopez, Balleine, & Dickinson, 1992), Rescorla (1992, 1994b) reported that rats did suppress their instrumental performance for a food (under extinction conditions) directly after they had had a single pairing of a taste with LiCl illness (i.e., *before* they had ever had a chance to experience the newly aversive palatability of that food, which would have required a second food presentation). Similarly, in a hunger-to-thirst transfer, Shipley and Colwill (1996) found that when rats that had learned while hungry to work for sucrose solution and for food pellets were later tested in extinction while thirsty, they would directly work more for the solution—before experiencing the new hedonic values of pellets and solution in the thirsty state. That is again different from the hunger-to-thirst results of Dickinson and colleagues (Dickinson & Dawson, 1987). Yet even Dickinson and colleagues themselves find that expectation-guided behavior in a few conditions is able to integrate in advance, without need of the usual new hedonic experience, for example after shifts from thirst to hunger (Dickinson & Balleine, 1990), though not after shifts from hunger to thirst (Dickinson & Dawson, 1987).

Thus there appears to be at least some ability even in rats for the cognitive mechanism to skip over the usual rationalist procedural assumption that the future will reflect the past, and instead to integrate associative and physiological information in advance of experiencing the new hedonic value, as the Bindra–Toates conditioned incentive salience mechanism automatically can do. This does not mean that the distinction is not real between cognitive expectation of reward and attribution of incentive salience to Pavlovian CS. As Shipley and Colwill (1996) noted regarding the difference between their result and Dickinson's, ''it is not impossible to reconcile our finding of a direct effect of outcome revaluation on instrumental performance with the general thesis proposed by . . . Dickinson and his colleagues'' (p. 65). There are several reasons why the cognitive mechanism might sometimes be able to make special use of information ''in advance'' yet still be distinct from other psychological incentive mechanisms, as Dickinson suggests.

First, as Shipley and Colwill (1996) point out, in some cases the cognitive expectation might not have to be readjusted in advance after all, but rather might make use of past experiences that were not intended by the experimenter. For example, their rats may have previously been thirsty and have experienced the increased hedonic value of the taste of water when drunk in a thirsty state. If so, they might have generalized that previously gained information about hedonic experience to the sucrose solution when they were thirsty.

Second, if a manipulation of incentive value produces an explicit hedonic experience by itself, then that hedonic experience should revalue the cognitive expectation of future hedonic experience. For example, when Rescorla (1992) found that a single pairing of a food with an injection of hypertonic LiCl reduced subsequent performance for that food (even before the rats were able to taste the now-aversive food), Balleine and Dickinson (1992) argued that the hypertonic injection had produced pain, a direct unpleasant experience (hypertonic solutions are more concentrated than body fluids and irritate pain receptors; isotonic solutions do not produce this irritation pain and so are usually used for injection instead—as Rescorla [1994b] did in a later replication). Pairing of the food with pain would have given the new (negative) hedonic experience needed by the cognitive mechanism to revalue its act-outcome expectation. It is probably worth mentioning that ordinary taste aversion learning produced by isotonic LiCl works because it makes the next experience of the food taste bad, according to Dickinson and colleagues, not because the rat carries an explicit association of the unpleasant nausea state. Presumably the nausea is too delayed to enter into an explicit cognitive representation with the idea of the food (even though it enters into a Pavlovian association). The disgusting experience of the unpleasant taste, when next encountered, is the immediate and explicit (negative) hedonic event that is usually needed to reduce the cognitively labeled incentive value of the taste.

Third, if Pavlovian cues are present in a way that can selectively guide a response, then Bindra–Toates incentive salience should be attributed to those reward cues, to control behavior in an incentive-appropriate fashion even in advance of experiencing the new consequences. As we saw earlier, the Bindra–Toates incentive conditioning mechanism can integrate physiological state to appropriately change ''wanting'' and ''liking'' for reward cues as soon as the physiological shift occurs, without need of experiencing the new hedonic value of the S* reward. That is enough in many cases to guide goal-directed behavior. So for example, Krieckhaus and Wolf (1968) found that rats that had learned that a concentrated salty solution was to be found from a certain spout when later made sodium deficient went directly to obtain the solution from that spout. Because attribution of

incentive salience to Pavlovian cues, such as a salt spout, can always be expected to follow integrative Bindra–Toates rules, any situation that has such conditioned incentive cues available will be able to guide behavior, and perhaps overrule or influence the cognitively expected hedonic value.

In general, the cognitive incentive expectation and Bindra–Toates incentive salience attribution mechanisms usually act to guide behavior in the same direction, not different directions. It is only in special cases, as when hedonic value is suddenly shifted and not yet re-experienced, that divergence between the mechanisms can be expected to happen at all. The two processes operate ordinarily in close association, and information leakage or communication between them might sometimes occur. The anticipatory integration of information that ordinarily drives the Bindra–Toates condition incentive salience attribution may under some conditions—not yet identified—be able to permeate the boundary and to affect the cognitive expectation even before the incentive is encountered again. We must accept for the moment that we do not yet understand what factors gate the access of the cognitive expectation to anticipatory integration. Presumably, cognitive access to Bindra–Toates integration might occur more often in humans than in rats, but there is surprisingly little evidence on the matter. When can cognitive expectation integrate in advance? When does it remain tied to the rationalist prediction that future value will equal past value? The answers to these questions remain for future research. For now, the best conclusion may be that there is a fundamental difference between these two mechanisms of reward learning, but that the divide between them can occasionally be bridged.

VII. Brain Mechanisms of Reward Learning

It is not my purpose to provide here a review of brain mechanisms of reward learning (for excellent reviews see Everitt et al., 1999, in press; Robbins & Everitt, 1996; Rolls, 1999; Schultz, 1998; Shizgal, 1999). Instead I only wish to touch on a few studies of brain substrates that are especially relevant to the distinction between cognitive incentive expectation to and the Bindra–Toates attribution of incentive salience to CS. Recent evidence indicates that these two processes differ not only psychologically but also in terms of their brain mechanisms, giving further justification for considering them as distinct. Incentive salience 'wants' for Pavlovian reward cues appear to be mediated primarily by the mesolimbic dopamine system and its forebrain targets (Balleine & Killcross, 1994; Dickinson, Smith, & Mirenowicz, 2000; Wyvell & Berridge, 2000). Cognitive act–outcome expectancy appears to be more directly mediated by specific cortical regions, such as the insular cortex and the prefrontal cortex (Balleine & Dickinson, 1998).

A. Mesolimbic Dopamine Systems Mediate Bindra–Toates Incentive Salience Attribution

Regarding reward learning, the brain system that has received the most attention in the past decade is the mesolimbic dopamine system, or dense projection of dopamine neurons from the midbrain's tegmentum to the forebrain's nucleus accumbens (and prefrontal cortex and amygdala to a lesser degree) (Beninger & Miller, 1998; Braver & Cohen, 1998; Montague, Dayan, & Sejnowski, 1996; Phillips et al., 1992; Schultz, 1998; Schultz, Dayan, & Montague, 1997; Vaccarino, Schiff, & Glickman, 1989). Mesolimbic dopamine neurons are activated by a variety of hedonic rewards, including palatable foods, sexual copulation, and drugs such as cocaine, amphetamine, or heroin, in a way that seems related to the incentive value of the reward (Di Chiara, 1998; Koob, 1996; Phillips et al., 1992; Wise, 1985, 1998). More relevant to reward *learning,* mesolimbic dopamine neurons and their accumbens targets are also activated by Pavlovian CSs for rewards (Di Ciano, Blaha, & Phillips, 1998; Fiorino, Coury, & Phillips, 1997; Schultz, Tremblay, & Hollerman, 1998). In many instances where predictive Pavlovian cues exist, the dopamine and accumbens neurons are activated before the hedonic reward occurs.

The predictive or anticipatory nature of mesolimbic dopamine responses has led to suggestions that this is a specialized brain substrate for reward learning (Montague et al., 1996; Schultz, 1998). These models have typically proposed that dopamine neurons play two roles in reward learning. First, the models suggest that dopamine neurons mediate the predictive *expectation* of a reward triggered by a CS. Second, they suggest that dopamine neurons play a role in *''teaching'' new predictions* by signaling the occurrence of outcomes that deviate from the expected. Generally these models have not attempted to choose between the Pavlovian incentive type versus cognitive incentive type of predictions, or to at all specify the precise psychological nature of the predictive expectation they invoke. However, the results of several studies indicate that the specific psychological nature of the expectation mediated by mesolimbic dopamine systems best corresponds to the Bindra–Toates sense of incentive salience attribution to CS—and not the cognitive sense of act–outcome expectation (Balleine & Killcross, 1994; Dickinson et al., 1999; Wyvell & Berridge, 2000).

B. Mesolimbic Dopamine Substrates of Pavlovian Incentive Salience

In one study, Dickinson et al. (2000) trained rats in the Dickinsonian paradigm that allows separation of cognitive incentive learning from Pavlovian incentive learning. They then tested the effects of drugs that suppress the neurotransmission of mesolimbic dopamine systems. They gave rats a

dopamine antagonist drug that blocks neural receptors for dopamine (either pimozide or alpha-flupenthixol). Dopamine suppression blocked both the acquisition and expression of Pavlovian incentive learning. If the rats received the dopamine antagonist just before their Pavlovian training trials, in which a Pavlovian CS was paired in a predictive fashion with food reward, they later behaved as though they never had received the Pavlovian training. When the CS was later presented to them as they bar pressed for food under extinction conditions in an undrugged state, they pressed no more than before the CS arrived. By contrast, rats that had been given vehicle during the Pavlovian training trials pressed harder when presented with the Pavlovian CS, as expected. In other words, the dopamine antagonist appeared to have blocked the acquisition by the Pavlovian CS of the conditioned incentive salience that would ordinarily have increased "wanting" for food.

In a follow-up experiment, Dickinson and colleagues (2000) found that the dopamine antagonist similarly prevented the *expression* of conditioned incentive salience attribution to a CS when the drug was given during the final test rather than during Pavlovian training. In this condition, all the rats received the CS–food Pavlovian training in an undrugged state. Later the hungry rats were tested for bar pressing under extinction conditions, while the CS was occasionally presented to them, after half the rats had received a dopamine antagonist drug and the others had not. The rats that were tested in a suppressed state of dopamine transmission failed to show any enhancement for bar pressing when the CS was presented, whereas the undrugged rats did press more during the CS. In other words, dopamine suppression blocked the expression of Pavlovian incentive value that had previously been learned while undrugged. The dopamine antagonists appeared to block the attribution of incentive salience to the CS that would ordinarily have caused cue-elicited "wanting" for the food.

A recent study by Cindy Wyvell in our laboratory, mentioned earlier, further found that conditioned incentive salience can be *enhanced* if mesoaccumbens dopamine systems are stimulated by amphetamine microinjection rather than suppressed (Wyvell & Berridge, 2000). Wyvell found that activation of dopamine neurotransmission in the accumbens, caused by microinjections of amphetamine directly into that brain structure, magnified the incentive salience attributed to a Pavlovian CS (Figure 5). In this study, as in the Dickinson experiments, rats learned to press a bar for food and separately learned an association between a predictive Pavlovian CS and food reward. A microinjection of amphetamine was made into the nucleus accumbens of the rats just before they were tested while bar pressing in extinction. Amphetamine microinjection dramatically magnified the increase in bar pressing caused by presentations of the Pavlovian food CS,

without much changing baseline bar pressing when the CS was not present. In other words, the activation of accumbens dopamine system apparently magnified the ability of the Pavlovian CS to trigger ''wanting'' for a food reward in a specific conditioned incentive fashion. It seemed to magnify the attribution of incentive salience specifically to the Pavlovian food cue (as the modified Bindra–Toates model of incentive salience would suggest).

The mediation of Pavlovian incentive learning by mesolimbic dopamine systems appears to be restricted to Pavlovian based ''wanting'' rather than to ''liking.'' Pavlovian conditioning can still cause shifts in the hedonic reaction pattern elicited by a CS even after the mesolimbic dopamine system is removed. For example, Berridge and Robinson found that changes in ''liking'' caused by a conditioned taste aversion could still be learned by rats that had lost their mesolimbic dopamine system due to lesions caused by a neurotoxin, 6-hydroxydopamine, which destroyed 98–99% of their mesolimbic dopamine neurons (Berridge & Robinson, 1998). We found that rats that had lost their dopamine systems still were perfectly capable of this learned suppression of incentive ''liking'' caused by Pavlovian conditioning. Rats with dopamine lesions switched their reaction to a sweet polycose-saccharin taste from a positive hedonic pattern (licking of their lips and paws) to an aversive pattern (gapes, arm flails, headshakes, etc.) just as normal rats did. In other words, it is Pavlovian incentive learning, not Pavlovian hedonic learning, that is mediated by mesoaccumbens dopamine systems. Dopamine in the accumbens is needed to ''want'' a conditioned incentive, but not necessarily to ''like'' it or to learn new ''likes'' and ''dislikes.''

C. Brain Substrates of Cognitive Incentive Learning

Less is known about the neural substrate for the cognitive expectation form of incentive learning than about conditioned incentive forms of reward learning. However, recent studies by Balleine and Dickinson and colleagues suggest that cognitive act–outcome representations of incentive value appear *not* to depend on the mesolimbic dopamine system. Instead, the prefrontal cortex and the insular cortex appear to be more important.

D. Mesolimbic Dopamine Does Not Mediate Cognitive Act–Outcome Expectation

At least two studies indicate that *cognitive representations* of act–outcome relations are *not* mediated by dopamine projections to the nucleus accumbens or by the nucleus accumbens itself. In one study, Balleine and Killcross (1994) found that accumbens lesions did not impair a rat's ability to adjust its level of instrumental response based on a change in reward value.

They concluded that accumbens lesions "do not influence sensitivity to the instrumental contingency" (p. 191) and that the role of the nucleus accumbens is "dissociated from the control of performance mediated by the act-outcome relation" (p. 181). The lesions did impair classically conditioned approach responses directed toward the food cup, consistent with an interpretation that there was reduced incentive salience attributed to reward-related stimuli such as the food cup, and Balleine and Killcross accordingly interpreted their accumbens lesion deficit as reduced "conditioned affective arousal produced by classical conditioning" (p. 191).

Second, in the dopamine antagonist study described earlier, Dickinson and colleagues (2000) tested the effects suppressing dopamine projection systems on cognitive incentive value versus Pavlovian incentive value. They gave rats dopamine-suppressing drugs that suppressed Pavlovian incentive salience. In this experiment they gave the drug under two conditions that would be expected to determine the rats' experienced-based expectation of the food's hedonic value. First, they let some rats eat the food while under the influence of the drug (but while outside of the instrumental test chamber), which allowed them to experience any hedonic consequences of the drug on food reward. Dopamine blockade failed to suppress the cognitive incentive value, or act–outcome representation, of the food for the rats. When the rats were allowed to eat food under the influence of the dopamine-blocking drug and thus gain an explicit representation of the remembered hedonic impact that food had in the drugged state, that memory did not reduce their bar pressing for the food later when they were tested under extinction conditions. In these conditions, food was no longer delivered, and so the rats had to operate entirely on the basis of their memory of the past food. Apparently, the incentive expectation based on that memory had not been diminished by the dopamine-blocking drug.

Dickinson and colleagues gave other rats in the same experiment the dopamine antagonist while they were being tested on the instrumental task under extinction conditions—that is, while they had to rely on their expectation of incentive value to guide their behavior. Again the dopamine antagonists failed to suppress the cognitive incentive value. Even when the rats were tested directly under the drug's influence (Dickinson et al., 2000), they still pressed at a normal rate under extinction conditions. Even though the dopamine antagonist would have suppressed bar pressing if the food were actually earned by a bar press (when it could disrupt the reboosting of incentive salience to its conditioned representation), it did not suppress the purely cognitive *expectation* of incentive value.

In other words, dopamine blockade suppresses the incentive motivation of rats to work for real food, if they experience the food as they are working, and also suppresses the Pavlovian incentive ability of cues to stimulate

motivation. But dopamine blockade apparently does not suppress the cognitive incentive expectation of the same food reward, which is based upon explicit *memory* of the food's hedonic value and on a cognitive representation of the act–outcome causal relation. As Dickinson and colleagues put it, "the absence of the effect of DA [dopamine] antagonists on incentive learning suggests that the incentive values of food rewards that control instrumental responding through a representation of the action-outcome contingency are mediated by a system that is not strongly modulated by a DA input" (Dickinson et al., 2000). That is, cognitive expectations of incentive value must be mediated by other brain systems.

E. Prefrontal Cortex and Insular Cortex Mediate Cognitive Expectation of Incentive Value

The brain systems that mediate cognitive incentive value are just beginning to be understood. So far, Balleine and Dickinson's studies have implicated structures such as prefrontal cortex and insular cortex in cognitive incentive expectation (and neurochemical substrates such as benzodiazepine and cholecystokinin systems that may mediate the experienced hedonic value of a food, which forms the explicit memory basis of expected hedonic value) (Balleine et al., 1994; Balleine & Dickinson, 1998).

The human and primate prefrontal cortex contains a dorsolateral region that has been especially implicated in act–outcome planning and related executive functions (Damasio, 1996; Niedermeyer, 1998; Robbins, 1996; Smith & Jonides, 1999), and in rats the corresponding region of prefrontal cortex is called the prelimbic area. Balleine and Dickinson made prelimbic lesions in the prefrontal cortex of rats that learned to perform two responses (press a bar, pull a chain), each of which earned its own specific food reward (Balleine & Dickinson, 1998). Then the rats were allowed to eat their fill of one of the foods elsewhere, before being tested on the two responses under extinction conditions (remember again, when neither response produced real food, so the rats were forced to operate solely on the basis of their expected incentive values). Eating a food to satiety is well known in humans to cause a temporary decrease in appetite and palatability ratings for that particular food, a phenomenon called *sensory-specific satiety* (Hetherington, 1996; Rolls, 1986). Rats also show a sensory-specific satiety decrease in hedonic behavioral affective reactions to a sweet taste just after they have consumed a large meal of it (Berridge, 1991). Thus, Dickinson and Balleine could expect that their rats would experience reduced hedonic value of the particular food they were sated upon by the end of their large meal of it. When tested immediately afterwards, normal rats did not work much on the response associated with the food they had just become sated

on, but they still continued to work robustly on the response associated with the other food. Rats that had prelimbic prefrontal cortex lesions, by contrast, failed to work hard at either response (Balleine & Dickinson, 1998). Damage to the prefrontal cortex appeared to impair the ability of rats to generate selective goal-directed behavior, when faced with multiple tasks carrying multiple outcomes, based on their expectations of relative hedonic value. It is interesting to note that human patients with damage to the prefrontal cortex have been to reported to lack an anticipatory galvanic skin response that normal people show when faced with a risky choice in certain situations (Bechara, Damasio, Tranel, & Damasio, 1997). This could be construed as a failure by these humans to generate an appropriate predictive representation of the causal relation between an act and its outcome's affective value. Conceivably there could be a link between prefrontal cortex mediation of the anticipatory prediction of emotional outcomes in humans, reflected in human anticipatory autonomic responses, and the type of outcome value expectation posited by Dickinson and his colleagues for rats (though it should be acknowledged that the brain lesions, tasks, and response measures, as well as the species involved, are all quite different in the two experimental paradigms).

The insular cortex, by contrast, is a cortical region involved in gustatory processing and has been shown to be important in mediating learned taste preference and avoidance (Dunn & Everitt, 1988). When Balleine and Dickinson (1998) made lesions in this area and tested rats in a similar experiment, rats with insular cortex lesions still worked under extinction conditions, apparently guided by an intact expectation of reward. However, after eating one food to satiety, they continued to work robustly on both responses under extinction conditions, and did not reduce responding on the task associated with the pre-fed food (Balleine & Dickinson, 1998). This was not due to an inability to experience sensory-specific satiety, as Balleine and Dickinson found that rats with insular cortex lesions showed a normal selective decline in responding after prefeeding if the two responses actually still earned the two foods (i.e., not extinction). The insular cortex lesions only prevented the *satiety devaluation* when tested under conditions that forced the rats to rely solely on *expectations* regarding the two foods. Balleine and Dickinson concluded that "what is impaired, however, is the ability to store and retain information about the changed value" (Balleine & Dickinson, 1998, p. 415).

Although much remains to be done to clarify the brain systems that mediate cognitive forms of incentive learning, these results suggest that prefrontal and insular cortical regions may play special roles. Balleine and Dickinson suggest that, in particular, the prelimbic prefrontal area may mediate learning about the causal contingency between act and outcome,

whereas the insular area may be needed to mediate cognitive expectation of the hedonic value of a particular food reward that is based on previous hedonic experience.

VIII. Conclusion

Many processes have been considered over the years as potential psychological mechanisms for reward learning. First of all, *reinforcement* concepts have been advanced in several forms. These include most notably stamping in accounts of a S-R habit via a mechanism triggered either by a hedonic satisfier or by drive reduction. Reinforcement concepts still have detectable influence on research and theory today, in fields that range from behavioral neuroscience to human motivational psychology. Next, *conditioned incentive* concepts of reward learning and motivation have become viewed as increasingly important, from the S-S* associations posited by Bolles (1972), so important in guiding goal-directed behavior, to the acquisition by CS of flexible and state-dependent incentive properties posited by Bindra and Toates, and finally to the splitting of ''wanting'' from ''liking'' and the attribution of conditioned ''wanting'' to cues posited by the incentive salience account. Finally, we have seen *cognitive expectation* mechanisms of hedonic-laden incentive outcomes. These concepts posit motivated behavior to be guided via representations of a causal relation between an act and its expected hedonic consequences, which has been constructed on the basis of explicit past experiences.

This diversity of potential mechanisms reflects more than changes in the theoretical orientation of psychologists over the decades and more than the existence of multiple schools of thought. It reflects a real diversity in the underlying psychological phenomena. Regarding S-R reinforcement, many experimental demonstrations, beginning with Watson's ''kerplunk'' experiment, leave no room to doubt that S-R habits indeed exist. Similarly, many demonstrations show that Pavlovian CS for rewards take on powerful incentive properties. Pavlovian reward cues and their representations elicit desire and are often needed for motivation. Conditioned cues themselves become the target of ''wanting,'' and become triggers for conditioned ''liking.'' They are sought after and even sometimes consumed. These Bindra–Toates incentive properties are by no means restricted to animals: from the ability of a taste of food to prime appetite for that food in normal humans to the intense, irrational, and overpowering cravings of drug addiction, there are numerous phenomena in daily life to testify to the power of incentive conditioning processes in the control of human motivation and behavior. Finally, because cognitive expectations are of all these processes the most

accessible to human introspection, as well as because of the focus of contemporary psychology on cognition, we need little to convince ourselves that cognitive incentive expectations form the basis of much of our daily behavior and of nearly all the motivated plans of which we are aware.

The point is that all these processes of reward learning and motivation exist simultaneously as psychological processes. They are all within each of us, operating in parallel and usually in cooperation. They diverge from one another under limited circumstances, and when that happens they may compete for control of behavior. Sometimes one process wins, sometimes another. We should not aim to dissolve all reward learning into one type of explanation, whether it be cognitive or associative, or reinforcement or incentive in nature. Future progress in the psychology of reward learning and of appetitive motivation will instead be best gained by a clearer understanding of the nature and features of each of these psychological processes, their relation to brain mechanisms, and the rules that govern interactions among them.

Acknowledgments

I am very grateful to Prof. Anthony Dickinson, Prof. Frederick Toates, and Dr. Cindy Wyvell for helpful comments on an earlier version of this manuscript. They helped me to correct many errors. Mistakes that remain are solely my own.

References

Allan, R. W., & Zeigler, H. P. (1994). Autoshaping the pigeon's gape response: Acquisition and topography as a function of reinforcer type and magnitude. *Journal of Experimental Analysis Behavior, 62*(2), 201–223.

Balleine, B. (1992). Instrumental performance following a shift in primary motivation depends on incentive learning. *Journal of Experimental Psychology: Animal Behavior Processes, 18*(3), 236–250.

Balleine, B., Ball, J., & Dickinson, A. (1994). Benzodiazepine-induced outcome revaluation and the motivational control of instrumental action in rats. *Behavioral Neuroscience, 108*(3), 573–589.

Balleine, B., & Dickinson, A. (1991). Instrumental performance following reinforcer devaluation depends upon incentive learning. *Quarterly Journal of Experimental Psychology Section B—Comparative and Physiological Psychology, 43*(3), 279–296.

Balleine, B., & Dickinson, A. (1992). Signaling and incentive processes in instrumental reinforcer devaluation. *Quarterly Journal of Experimental Psychology Section B—Comparative and Physiological Psychology, 45*(4), 285–301.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology, 37*(4-5), 407–419.

Balleine, B., & Killcross, S. (1994). Effects of ibotenic acid lesions of the nucleus accumbens on instrumental action. *Behavioural Brain Research, 65*(2), 181–193.

Beauchamp, G. K., Bertino, M., Burke, D., & Engelman, K. (1990). Experimental sodium depletion and salt taste in normal human volunteers. *American Journal of Clinical Nutrition, 51*(5), 881–889.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*(5304), 1293–1295.

Bedard, M., & Weingarten, H. P. (1989). Postabsorptive glucose decreases excitatory effects of taste on ingestion. *American Journal of Physiology, 256* (5 Pt 2), R1142–1147.

Beninger, R. J., & Miller, R. (1998). Dopamine D-1 like receptors and reward-related incentive learning. *Neuroscience & Biobehavioral Reviews, 22*(2), 335–345.

Berridge, K., Grill, H. J., & Norgren, R. (1981). Relation of consummatory responses and preabsorptive insulin release to palatability and learned taste aversions. *Journal of Comparative and Physiological Psychology, 95*(3), 363–82.

Berridge, K. C. (1991). Modulation of taste affect by hunger, caloric satiety, and sensory-specific satiety in the rat. *Appetite, 16*(2), 103–20.

Berridge, K. C. (1996). Food reward: brain substrates of wanting and liking. *Neuroscience and Biobehavioral Review, 20*(1), 1–25.

Berridge, K. C. (1999). Pleasure, pain, desire, and dread: Hidden core processes of emotion. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 525–557). New York: Russell Sage Foundation.

Berridge, K. C. (2000). Taste reactivity: Measuring hedonic impact in human infants and animals. *Neuroscience Biobehavioral Reviews, 24,* 173–198.

Berridge, K. C., Flynn, F. W., Schulkin, J., & Grill, H. J. (1984). Sodium depletion enhances salt palatability in rats. *Behavioral Neuroscience, 98*(4), 652–660.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research—Brain Research Reviews, 28*(3), 309–369.

Berridge, K. C., & Schulkin, J. (1989). Palatability shift of a salt-associated incentive during sodium depletion. *Quarterly Journal of Experimental Psychology* [*b*], *41*(2), 121–138.

Berridge, K. C., & Valenstein, E. S. (1991). What psychological process mediates feeding evoked by electrical stimulation of the lateral hypothalamus? *Behavioral Neuroscience, 105*(1), 3–14.

Berridge, K. C., Venier, I. L., & Robinson, T. E. (1989). Taste reactivity analysis of 6-hydroxydopamine-induced aphagia: Implications for arousal and anhedonia hypotheses of dopamine function. *Behavioral Neuroscience, 103*(1), 36–45.

Bindra, D. (1974). A motivational view of learning, performance, and behavior modification. *Psychological Review, 81*(3), 199–213.

Bindra, D. (1978). How adaptive behavior is produced: a perceptual-motivation alternative to response reinforcement. *Behavioral and Brain Sciences, 1,* 41–91.

Blackburn, J. R., Pfaus, J. G., & Phillips, A. G. (1992). Dopamine functions in appetitive and defensive behaviours. *Progress in Neurobiology, 39*(3), 247–279.

Bolles, R. C. (1972). Reinforcement, expectancy, and learning. *Psychological Review, 79,* 394–409.

Braver, T. S., & Cohen, J. D. (1998). An integrated computational model of dopamine function in reinforcement learning and working memory. *Journal of Cognitive Neuroscience,* 82–82.

Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist, 16,* 681–684.

Brown, L. (1993). *The New Shorter Oxford English Dictionary on Historical Principles*. Oxford & New York: Clarendon Press: Oxford University Press.

Brown, P. L., & Jenkins, H. M. (1968). Auto-shaping of the pigeon's key-peck. *Journal of the Experimental Analysis of Behavior, 11*(1), 1–8.

Cabanac, M. (1979). Sensory pleasure. *Quarterly Review of Biology, 54*(1), 1–29.
Cabanac, M. (1992). Pleasure: The common currency. *Journal of Theoretical Biology, 155,* 173–200.
Cabanac, M., & Lafrance, L. (1990). Postingestive alliesthesia: the rat tells the same story. *Physiology and Behavior, 47*(3), 539–43.
Carlezon, W. A., Jr., & Wise, R. A. (1996). Rewarding actions of phencyclidine and related drugs in nucleus accumbens shell and frontal cortex. *Journal of Neuroscience, 16*(9), 3112–22.
Carr, H., & Watson, J. B. (1908). Orientation of the white rat. *Journal of Comparative Neurology and Psychology, 18,* 27–44.
Cornell, C. E., Rodin, J., & Weingarten, H. (1989). Stimulus-induced eating when satiated. *Physiology & Behavior, 45*(4), 695–704.
Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London-Series B: Biological Sciences, 351*(1346), 1413–20.
Delamater, A. R., LoLordo, V. M., & Berridge, K. C. (1986). Control of fluid palatability by exteroceptive Pavlovian signals. *J Exp Psychol* [*Anim Behav*], *12*(2), 143–52.
Depue, R. A., & Collins, P. F. (1999). Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion. *Behavioral and Brain Sciences, 22*(3), 491–569.
Di Chiara, G. (1998). A motivational learning hypothesis of the role of mesolimbic dopamine in compulsive drug use. *Journal of Psychopharmacology, 12*(1), 54–67.
Di Ciano, P., Blaha, C. D., & Phillips, A. G. (1998). The relation between dopamine oxidation currents in the nucleus accumbens and conditioned increases in motor activity in rats following repeated administration of d-amphetamine or cocaine. *European Journal of Neuroscience, 10*(3), 1113–20.
Dickinson, A. (1989). Expectancy theory in animal conditioning. In R. R. M. Stephen & B. Klein (Eds.), *Contemporary learning theories: Pavlovian conditioning and the status of traditional learning theory* (pp. 279–308). Hillsdale, NJ. Lawrence Erlbaum Associates, Inc.
Dickinson, A., & Balleine, B. (1990). Motivational control of instrumental performance following a shift from thirst to hunger. *Quarterly Journal of Experimental Psychology* [*B*], *42*(4), 413–431.
Dickinson, A., & Balleine, B. (1994) Motivational control of goal-directed action. *Animal Learning & Behavior, 22,* 1–18.
Dickinson, A., & Balleine, B. (1995). Motivational control of instrumental action. *Current Directions in Psychological Science, 4,* 162–167.
Dickinson, A., & Dawson, G. R. (1987). Pavlovian processes in the motivational control of instrumental performance. *Quarterly Journal of Experimental Psychology, B* (*Comparative & Physiological Psychology, 39,* 201–213.
Dickinson, A., Smith, J., & Mirenowicz, J. (2000). Dissociation of Pavlovian and instrumental incentive learning under dopamine antagonists. *Behavioral Neuroscience, 40,* 468–483.
Dunn, L. T., & Everitt, B. J. (1988). Double dissociations of the effects of amygdala and insular cortex lesions on conditioned taste aversion, passive avoidance, and neophobia in the rat using the excitotoxin ibotenic acid. *Behavioral Neuroscience, 102*(1), 3–23.
Everitt, B. J., Cardinal, R. N., Hall, J., Parkinson, J. A., & Robbins, T. R. (in press). Differential involvement of amygdala subsystems in repetitive conditioning and drug addiction. In J. P. Aggleton (Ed.), *The amygdala: a functional analysis*. Oxford: Oxford University Press.
Everitt, B. J., Parkinson, J. A., Olmstead, M. C., Arroyo, M., Robledo, P., & Robbins, T. W. (1999). Associative processes in addiction and reward. The role of amygdala-ventral striatal subsystems. *Annals of the New York Academy of Sciences, 877,* 412–38.

Falk, J. L. (1971). The nature and determinants of adjunctive behavior. *Physiology & Behavior, 6*(5), 577–88.

Fibiger, H. C., & Phillips, A. G. (1986). Reward, motivation, cognition: Psychobiology of mesotelencephalic systems. *Handbook of Physiology—The Nervous System, 4,* 647–675.

Fiorino, D. F., Coury, A., & Phillips, A. G. (1997). Dynamic changes in nucleus accumbens dopamine efflux during the Coolidge effect in male rats. *Journal of Neuroscience, 17*(12), 4849–55.

Fischman, M. W., & Foltin, R. W. (1992). Self-administration of cocaine by humans: A laboratory perspective. In G. R. Bock & J. Whelan (Eds.), *Cocaine: Scientific and social dimensions* (Vol. 166, pp. 165–180). Chichester, UK: Wiley.

Flaherty, C. F. (1996). *Incentive relativity*. New York: Cambridge University Press.

Fudim, O. K. (1978). Sensory preconditioning of flavors with a formalin-produced sodium need. *Journal of Experimental Psychology: Animal Behavior Processes, 4*(3), 276–85.

Galef, B. G., Jr. (1998). Edward Thorndike: Revolutionary psychologist, ambiguous biologist. *American Psychologist: Special Issue: Celebrating E. L. Thorndike, 53*(10), 1128–1134.

Gallistel, C. R. (1978). Irrelevance of past pleasure. *Behavioral and Brain Sciences, 1,* 59–60.

Garcia, J., McGowan, B. K., Ervin, F. R., & Koelling, R. A. (1968). Cues: their relative effectiveness as a function of the reinforcer. *Science, 160*(829), 794–5.

Grill, H. J., & Norgren, R. (1978). Chronically decerebrate rats demonstrate satiation but not bait shyness. *Science, 201*(4352), 267–9.

Hetherington, M. M. (1996). Sensory-specific satiety and its importance in meal termination. *Neuroscience & Biobehavioral Reviews, 20*(1), 113–7.

Holland, P. C. (1990). Event representation in Pavlovian conditioning: image and action. *Cognition, 37*(1-2), 105–31.

Holland, P. C. (1993). Cognitive aspects of classicial conditioning. *Current Opinion in Neurobiology, 3*(2), 230–6.

Hull, C. L. (1943). *Principles of behavior, an introduction to behavior theory*. New York: D. Appleton-Century Co.

Jenkins, H. M., & Moore, B. R. (1973). The form of the auto-shaped response with food or water reinforcers. *Journal of the Experimental Analysis of Behavior, 20*(2), 163–81.

Kelley, A. E. (1999). Neural integrative activities of nucleus accumbens subregions in relation to learning and motivation. *Psychobiology, 27*(2), 198–213.

Konorski, J. (1967). *Integrative activity of the brain: an interdisciplinary approach*. Chicago: University of Chicago Press.

Koob, G. F. (1996). Hedonic valence, dopamine and motivation. *Molecular Psychiatry, 1*(3), 4.

Krieckhaus, E. E., & Wolf, G. (1968). Acquisition of sodium by rats: interaction of innate mechanisms and latent learning. *Journal of Comparative & Physiological Psychology, 65*(2), 197–201.

Lamb, R. J., Preston, K. L., Schindler, C. W., Meisch, R. A., Davis, F., Katz, J. L., Henningfield, J. E., & Goldberg, S. R. (1991). The reinforcing and subjective effects of morphine in post-addicts: a dose-response study. *J. Pharmacol Exp Ther, 259*(3), 1165–73.

Lopez, M., Balleine, B., & Dickinson, A. (1992). Incentive learning following reinforcer devaluation is not conditional upon the motivational state during re-exposure. *Quarterly Journal of Experimental Psychology* [*B*], *45*(4), 265–84.

Lovibond, P. F. (1983). Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes, 9*(3), 225–47.

Marshall, J. F., Richardson, J. S., & Teitelbaum, P. (1974). Nigrostriatal bundle damage and the lateral hypothalamic syndrome. *Journal of Comparative and Physiological Psychology, 87,* 808–830.

McClelland, D. C. (1987). *Human motivation*. Cambridge England; New York: Cambridge University Press.

McFarland, D. J. (1969). Separation of satiating and rewarding consequences of drinking. *Physiology and Behavior, 4,* 987–989.

McFarland, K., & Ettenberg, A. (1998). Naloxone blocks reinforcement but not motivation in an operant runway model of heroin-seeking behavior. *Experimental and Clinical Psychopharmacology, 6*(4), 353–359.

Miller, N. E. (1973). How the project started. In E. S. Valenstein (Ed.), *Brain Stimulation and Motivation: research and commentary* (pp. 53–68). Glenview, Ill.: Scott, Foresman and company.

Miller, N. E., & Kessen, M. L. (1952). Reward effects of food via stomach fistula compared with those of food via mouth. *Journal of Comparative and Physiological Psychology, 45,* 555–564.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience, 16*(5), 1936–47.

Moss, F. A., & Thorndike, E. L. (1934). *Comparative psychology*. New York: Prentice-Hall.

Niedermeyer, E. (1998). Frontal lobe functions and dysfunctions. *Clinical Electroencephalography, 29*(2), 79–90.

Olds, J. (1973). The discovery of reward systems in the brain. In E. S. Valenstein (Ed.), *Brain Stimulation and Motivation: research and commentary* (pp. 81–99). Glenview, Ill.: Scott, Foresman and company.

Olds, J., & Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology, 47,* 419–427.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. Oxford, UK: Oxford University Press.

Phillips, A. G., Blaha, C. D., Pfaus, J. G., & Blackburn, J. R. (1992). Neurobiological correlates of positive emotional states: Dopamine, anticipation, and reward. In K. T. Strongman (Ed.), *International Review of Studies on Emotion* (Vol. 2, pp. 31–50). New York: John Wiley & Sons.

Rescorla, R. A. (1992). Depression of an instrumental response by a single devaluation of its outcome. *Q J Exp Psychol* [*B*], *44*(2), 123–36.

Rescorla, R. A. (1994a). Control of instrumental performance by Pavlovian and instrumental stimuli. *J Exp Psychol Anim Behav Process, 20*(1), 44–50.

Rescorla, R. A. (1994b). A note on depression of instrumental responding after one trial of outcome devaluation. *Quarterly Journal of Experimental Psychology, 47B*(1), 27–37.

Rescorla, R. A., & Lolordo, V. M. (1965). Inhibition of avoidance behavior. *Journal of Comparative & Physiological Psychology, 59,* 406–412.

Rescorla, R. A., & Solomon, R. L. (1967). Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological Review, 74*(3), 151–82.

Reynolds, S., & Berridge, K. C. (2000). Anatomical segregation of behaviors elicited by GABA agonist microinjections in the nucleus accumbens shell: eating versus defensive treading. *Manuscript in preparation*.

Robbins, T. W. (1996). Dissociating executive functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London—Series B: Biological Sciences, 351*(1346), 1463–1470; discussion 1470–1471.

Robbins, T. W., & Everitt, B. J. (1996). Neurobehavioural mechanisms of reward and motivation. *Current Opinion in Neurobiology, 6*(2), 228–36.

Robinson, T. E., & Berridge, K. C. (1993). The neural basis of drug craving: an incentive-sensitization theory of addiction. *Brain Research Reviews, 18*(3), 247–291.
Rolls, B. J. (1986). Sensory-specific satiety. *Nutrition Reviews, 44*(3), 93–101.
Rolls, E. T. (1999). *The brain and emotion*. Oxford: Oxford University Press.
Rowland, N., & Nicolaidis, S. (1976). Metering of fluid intake and determinants of ad libitum drinking in rats. *American Journal of Physiology, 231*(1), 1–8.
Rozin, P. (1999). Preadaptation and the puzzles and properties of pleasure. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 109–133). New York: Russell Sage Foundation.
Schulkin, J. (1991). *Sodium hunger: The search for a salty taste*. New York: Cambridge University Press.
Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology, 80*(1), 1–27.
Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593–1599.
Schultz, W., Tremblay, L., & Hollerman, J. R. (1998). Reward prediction in primate basal ganglia and frontal cortex. *Neuropharmacology, 37*(4-5), 421–429.
Schwartz, B., & Williams, D. R. (1972). The role of the response-reinforcer contingency in negative automaintenance. *Journal of the Experimental Analysis of Behavior, 17,* 351–357.
Shipley, B. E., & Colwill, R. M. (1996). Direct effects on instrumental performance of outcome revaluation by drive shifts. *Animal Learning & Behavior, 24*(1), 57–67.
Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York: Appleton-Century Company.
Skinner, B. F. (1948). 'Superstition' in the pigeon. *Journal of Experimental Psychology, 38,* 168–172.
Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review, 57,* 193–216.
Smith, E. E., & Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science, 283*(5408), 1657–1661.
Staddon, J. E., & Simmelhag, V. L. (1970). The "superstition" experiment: A reexamination of its A reexamination of its implications for the principles of adaptive behavior. *Psychological Review, 78,* 3–43.
Staddon, J. E. R. (1999). Theoretical behaviorism. In W. T. O'Donohue & R. F. Kitchener (Eds.), *Handbook of behaviorism* (pp. 217–241). San Diego, CA: Academic Press, Inc.
Steiner, J. E. (1979). Human facial expressions in response to taste and smell stimulation. *Advances in Child Development and Behavior, 13,* 257–295.
Steiner, J. E., Glaser, D., Hawilo, M. E., & Berridge, K. C. (1999). *Comparative expression of hedonic impact: Affective reactions to taste by human infants and other primates.* Manuscript submitted for publication.
Stewart, J., de Wit, H., & Eikelboom, R. (1984). Role of unconditioned and conditioned drug effects in the self-administration of opiates and stimulants. *Psychological Review, 91*(2), 251–268.
Stratford, T. R., & Kelley, A. E. (1997). GABA in the nucleus accumbens shell participates in the central regulation of feeding behavior. *Journal of Neuroscience, 17*(11), 4434–4440.
Thorndike, E. L. (1898). *Animal intelligence: An experimental study of the associative processes in animals*. New York: The Macmillan Company.
Thorndike, E. L. (1998). Animal intelligence: An experimental study of the associate processes in animals (originally published 1898). *American Psychologist, 53*(10), 1125–1127.
Toates, F. (1986). *Motivational systems*. Cambridge, UK: Cambridge University Press.
Toates, F. (1998). The interaction of cognitive and stimulus-response processes in the control of behaviour. *Neuroscience and Biobehavioral Reviews, 22*(1), 59–83.

Toates, F. M. (1994). Comparing motivational systems—an incentive motivation perspective. In C. R. Legg & D. A. Booth (Eds.), *Appetite: Neural and Behavioural Bases* (pp. 305–327). New York: Oxford University Press.

Tolman, E. C. (1949). *Purposive behavior in animals and men.* Berkeley, CA: Appleton-Century-Craft.

Tomie, A. (1996). Locating reward cue at response manipulandum (CAM) induces symptoms of drug abuse. *Neuroscience and Biobehavioral Reviews, 20*(3), 31.

Turner, L. H., Solomon, R. L., Stellar, E., & Wampler, S. N. (1975). Humoral factors controlling food intake in dogs. *Acta Neurobiologiae Experimentalis, 35*(5-6), 491–498.

Ungerstedt, U. (1971). Adipsia and aphagia after 6-hydroxydopamine induced degeneration of the nigro-striatal dopamine system. *Acta physiologica Scandinavica. Supplementum., 367,* 95–122.

Uttal, W. R. (2000). *The war between mentalism and behaviorism: On the accessibility of mental processes*. Mahwah, NJ, US: Lawrence Erlbaum Associates, Inc., Publishers.

Vaccarino, F. J., Schiff, B. B., & Glickman, S. E. (1989). Biological view of reinforcement. In S. B. Klein & R. R. Mowrer (Eds.), *Contemporary learning theories: Instrumental conditioning theory and the impact of biological constraints on learning* (pp. 111–142). Hillsdale, NJ: Lawrence Erlbaum Associates.

Valenstein, E. S., (1976). The interpretation of behavior evoked by brain stimulation. In A. Wauquier & E. T. Rolls (Eds.), *Brain-stimulation reward* (pp. 557–575). New York: Elsevier.

Valenstein, E. S. Cox, V. C., & Kakolewski, J. W. (1970). Reexamination of the role of the hypothalamus in motivation. *Psychological Review, 77*(1), 16–31.

Watson, J. B. (1913). Psychology as the behaviourist views it. *Psychological Review, 20*(2), 158–177.

Weingarten, H. P. (1983). Conditioned cues elicit feeding in sated rats: a role for learning in meal initiation. *Science, 220*(4595), 431–433.

Weingarten, H. P., & Martin, G. M. (1989). Mechanisms of conditioned meal initiation. *Physiology & Behavior, 45*(4), 735–40.

White, N. M. (1989). Reward or reinforcement: what's the difference? *Neuroscience and Biobehavioral Review, 13,* 181–186.

Williams, D. R., & Williams, H. (1969). Auto-maintenance in the pigeon: Sustained pecking despite contingent non-reinforcement. *Journal of the Experimental Analysis of Behavior, 12,* 511–520.

Winkielman, P., Zajonc, R. B., & Schwarz, N. (1997). Subliminal affective priming resists attributional interventions. *Cognition and Emotion, 11,* 433–465.

Wise, R. A. (1985). The anhedonia hypothesis: Mark III. *Behavioral and Brain Sciences, 8,* 178–186.

Wise, R. A. (1998). Drug-activation of brain reward pathways. *Drug & Alcohol Dependence, 51*(1-2), 13–22.

Wyvell, C. L., & Berridge, K. C. (2000). Intra-accumbens amphetamine increases the pure incentive salience of a Pavlovian cue for food reward: Enhancement of 'wanting' without either 'liking' or reinforcement. *Manuscript submitted for publication*.

Young, P. T. (1966). Hedonic organization and regulation of behavior. *Psychological Review, 73*(1), 59–86.

SPATIAL DIAGRAMS: Key Instruments in the Toolbox for Thought

Laura R. Novick

I. The Toolbox for Thought

When I have a home maintenance problem to solve, I open the closet and look through my red, metal toolbox for the tool that best fits the constraints of the situation (Figure 1a). To hang a picture, I am likely to choose the hammer. To mount a smoke detector on the wall, I may choose a screwdriver. To tighten the bolt on my bicycle handlebars, a wrench probably would be most useful. Similarly, when I have a cognitive problem to solve, I examine my toolbox for thought for the reasoning tool that best fits the constraints of the situation (Figure 1b). In some cases, the appropriate tool is a procedure for transforming the given information into the desired goal state. Asked to find the average height of the players on my child's basketball team, I am likely to retrieve the formula for computing the arithmetic mean. Given a river crossing problem to solve, I may engage in means-ends analysis. An extensive literature in cognitive psychology has documented the wealth of problem-solving procedures available in the toolbox for thought.

But problem-solving procedures are not the only items in one's cognitive toolbox. The toolbox also contains various methods for representing the

THE PSYCHOLOGY OF LEARNING AND MOTIVATION, VOL. 40

Copyright © 2001 by Academic Press.
All rights of reproduction in any form reserved.
0079-7421/00 $35.00

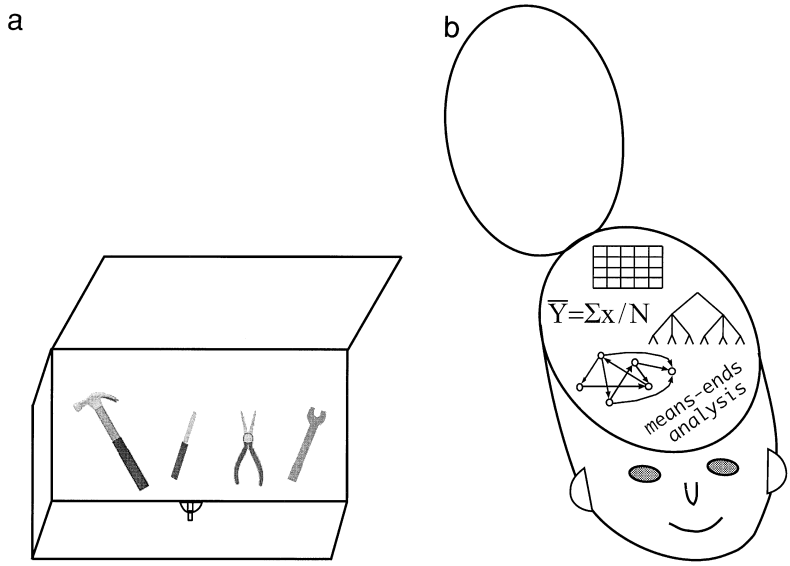


Fig. 1. (a) Home maintenance tools and (b) tools for thought.

information in a problem. Pictorial formats for problem representation are key instruments in the toolbox for thought. As McKim (1980) noted, "you will find graphic-language expressions on the blackboards of almost every department of a university, from aeronautics to zoology" (p. 142). Hegarty, Carpenter, and Just (1991) have classified such expressions into three broad categories. Iconic diagrams, such as photographs and line drawings, depict concrete objects in a very transparent manner, in that the spatial relations in the diagram are isomorphic to those in the referent object. In contrast, schematic diagrams or conceptual graphs (Butler, 1993), such as circuit diagrams, Venn diagrams, and hierarchical trees, depict abstract concepts. They typically rely on learned conventions to depict both the components of the concept and their organization (Dufour-Janvier, Bednarz, & Belanger, 1987; Hegarty et al., 1991). The final category in Hegarty et al.'s taxonomy is charts and graphs. The referent depicted in these diagrams is a set of related (typically quantitative) data. Common examples are pie charts and bar and line graphs.

Among the representational tools available, the superiority of visual tools is part of our folk wisdom. According to a familiar Chinese proverb, a picture is worth ten thousand words. A similar Japanese proverb states

that "a hundred hearings do not equal one look" (Iwasaki, 1995, p. 657). Research supports the folk wisdom, as many studies have documented the benefits of pictures and diagrams for learning, reasoning, and problem solving (e.g., Bartram, 1980; Bauer & Johnson-Laird, 1993; Day, 1988; Guri-Rozenblit, 1988; Hegarty & Just, 1993; Levin, 1989; Mayer & Gallini, 1990; Novick & Hmelo, 1994; Novick & Morse, in press; Polich & Schwartz, 1974; Scanlon, 1989; S. Schwartz, 1971; Sweller, Chandler, Tierney, & Cooper, 1990; Vessey & Weber, 1986). Among the various types of pictorial representations, a number of researchers and theoreticians have particularly touted the virtues of abstract diagrams as tools for thinking (e.g., Dufour-Janvier et al., 1987; Kindfield, 1993/1994; Larkin & Simon, 1987; Polya, 1957). A primary reason for this is that they provide a coherent representation of the objects and relations presented verbally in a problem—that is, a representation whose parts are interconnected by a high-level relational property or global theme (e.g., Greeno, 1977).

II. The Importance of Selecting the Appropriate Representational Tool

One may distinguish three aspects of skilled use of representational tools for reasoning and problem solving. First, solvers must be able to identify the most appropriate (i.e., most efficient) type of representation for the problem at hand. That is, they need to determine, for example, whether the structure of the problem better fits that of a hierarchy or a Venn diagram. Second, they must be able to construct the selected representation so that the important concepts and relations in the problem are accurately mapped onto the components of the representation. Finally, to generate an answer to the problem, solvers must be able to apply appropriate inference rules to the information in the representation (also see Greeno, 1983; Larkin & Simon, 1987).

I would like to argue for the fundamental importance of investigating the first step, namely people's ability to select the appropriate representation for a given situation. The selected representation serves as the foundation for the remainder of the reasoning process. The representation one adopts for a problem indicates one's understanding of the problem, and understanding is arguably the most important component of problem solving (e.g., Duncker, 1945; Greeno, 1977). As Greeno (1977) noted, "the difference between understanding and not understanding is in the nature of the representation . . . good understanding involves achievement of a coherent representation" (p. 44). Although in some cases solvers are able to compute the correct answer without really understanding what the problem is about

(e.g., Paige & Simon, 1966; Wertheimer, 1959/1982), we would not ascribe great problem-solving skill to them. In the more typical case, the type of representation used for a problem affects the ease or likelihood of a solution (e.g., Bassok & Olseth, 1995; Day, 1988; Hayes & Simon, 1977; McGuinness, 1986; S. Schwartz, 1971): Superior performance is obtained only when the display format and the structure of the environment are consistent (Sanfey & Hastie, 1998). Effective problem solving, therefore, depends on being able to select a representational tool whose structure matches that of the information to be represented.

III. Spatial Diagram Tools

I noted earlier that abstract diagrams are particularly important tools for thinking. Among the many different kinds of abstract diagrams, I am particularly interested in the three interrelated spatial diagrams shown in Figure 1b: hierarchies, matrices, and networks (path diagrams). The general superiority of diagrams over sentential representations in many learning and problem-solving situations applies to the three spatial diagrams as well (e.g., Bartram, 1980; Day, 1988; Guri-Rozenblit, 1988; Novick & Hmelo, 1994; Polich & Schwartz, 1974; Scanlon, 1989; S. Schwartz, 1971; Vessey & Weber, 1986).

Because of their broad applicability, these diagrams highlight structural similarities across situations that are superficially very different (Novick, Hurley, & Francis, 1999). For example, a hierarchy can be used to represent ''familial'' relations among members of the animal kingdom or the search space at a given point in a chess game; a matrix can be used to represent the grade sheet for a class or possible item pairings in a deductive-reasoning problem; and a network can be used to represent the food web for an ecosystem or the alliances among nations. Thus, by successfully constructing these diagrams, solvers would be led to see deep (i.e., structural) similarities among diverse situations that otherwise might not be salient. Structural understanding is a key factor underlying expertise (e.g., Chi, Feltovich, & Glaser, 1981; Schoenfeld & Herrmann, 1982).

IV. Overview of the Chapter

Because selection of an appropriate tool provides the foundation for subsequent reasoning, most of the remainder of this chapter will focus on people's ability to select the spatial diagram that best captures the structure of a situation. Most of my research on this topic adopts a competence perspec-

tive. That is, I seek to determine what people know about these tools and when to use them, under conditions that are relatively favorable for them to exhibit their knowledge. Except in the case of experts, people's performance typically underestimates their competence, so it is difficult to reason backwards from performance to competence. In the other direction, competence places an upper bound on performance. Improving performance in a particular situation may require enhancing competence, or it may simply require helping people to access the appropriate knowledge they already possess.

Broadly speaking, my research has considered two means of determining an appropriate spatial diagram to use for a particular situation. One possibility is that people's decisions in this matter are primarily exemplar-driven, relying on the similarity of the current situation to those encountered previously. In this case, solvers recall examples of specific situations encountered in the past in which a particular type of diagram has proven useful. College students clearly have such information at their disposal. Novick et al. (1999) asked students to generate as many examples as they could of situations for which each type of spatial diagram could be used to organize information. The subjects were given 4–5 min to generate examples for each diagram. Averaged across two studies, subjects generated approximately 3.5–4 (appropriate) examples for each type of diagram. Their examples belonged to roughly 2–3 identifiable subtypes for each diagram. Across all the subjects, the authors identified four different subtypes of examples for the network, five for the matrix, and seven for the hierarchy. For example, the hierarchy subtypes include descent (e.g., family tree), power (e.g., managerial levels in a corporation), and elimination (e.g., basketball tournament pairings).

This exemplar-driven, knowledge transfer strategy is often employed by problem solvers when they are searching for an appropriate procedural tool to use. For example, in solving the problems on a math homework assignment, students often try to match each homework problem to an example problem given in the textbook or in class so that they can adapt the procedure used to solve the example problem to work for the new problem. In these cases, the use of the formula or procedure appropriate for the homework problem is guided by the correspondences computed between that problem and the example problem (e.g., Novick, 1995; Novick & Holyoak, 1991; Reed, 1987; Ross, 1987, 1989).

If people's knowledge about the spatial diagram tools is primarily exemplar-based, then selection of an appropriate type of spatial diagram to use in a new situation would rely on the mechanism of representational transfer. I will describe the results of several studies examining the extent to which students use exemplar-based reasoning (i.e., representational transfer) to guide their selection of an appropriate spatial diagram tool to

use (Novick, 1990; Novick & Hmelo, 1994; Novick et al., 1999). These studies indicate that although students can use such reasoning, it does not seem to be their preferred mode of operation, nor is it the most effective strategy.

A second possibility is that reasoners retrieve an abstract schema for one or more diagrammatic tools and choose the type of tool whose structural characteristics best match the situation at hand. This is similar to retrieving a mathematical formula when one is considering the best procedural tool to use. There is precedence for this proposal, as Koedinger and Anderson (1990) have found evidence for the use of diagram schemas that are specific to solving geometry proofs. A matrix schema, for example, might indicate that matrices are particularly useful when (a) there are items in two distinct sets, (b) the items within a set cannot be combined, (c) all possible combinations of the items across sets must be considered, and (d) the links between items are nondirectional. The results of several studies suggest that students have rudimentary abstract schemas for the three spatial diagrams and that they use these schemas to guide their selections of the appropriate diagrammatic tool to use (Novick et al., 1999; Novick & Hurley, 2000).

If representational decisions are based in part on abstract, schematic knowledge, then to really understand students' knowledge of the three spatial diagrams requires some understanding of the content of their schemas. Although it might be possible to address this issue completely inductively, it is likely to be more fruitful to start with a set of hypotheses about what information ought to be contained in the diagram schemas. If these schemas are to be used to select appropriate representations, they need to contain information about the applicability conditions for their respective diagrams. Just as people need to know the structural properties of situations that specify the use of a hammer versus a wrench, they need to know the structural properties that indicate the appropriateness of using a matrix versus a hierarchy. Determining what the applicability conditions for the three spatial diagrams are thus requires a structural analysis of the tools themselves. Novick and Hurley (2000) have provided such an analysis, which I will summarize here. They also presented data validating their structural analysis and describing students' knowledge about the applicability conditions for the three spatial diagrams. Their data came from students who were relatively knowledgeable about the three spatial diagrams due to their college coursework. In this chapter, I will also present, for the first time, data concerning more typical students' knowledge about the applicability conditions for the three spatial diagrams.

Ultimately, it will be important to link people's knowledge about the three spatial diagrams to their ability to construct and use these representations in the service of reasoning and problem solving. Because the research

on people's knowledge is very recent, these links have not yet been made. Thus, I will only briefly consider the process of constructing spatial diagrams, near the end of the chapter.

V. The Use of Exemplar Knowledge in Choosing Diagrammatic Tools

A. Evidence For the Use of Exemplar Knowledge

A common paradigm for investigating people's use of exemplar knowledge for reasoning and problem solving is the transfer paradigm (e.g., Gick & Holyoak, 1980; Novick, 1988; Novick & Holyoak, 1991; Reed, 1987; Ross, 1987, 1989). In this paradigm, subjects are first introduced to several example problems. Then they are given one or more additional problems to solve, and one assesses whether they are able to appropriately transfer what they learned from the example problems to the new problems. In the typical study, subjects are not told anything about the possible relation between the example and test problems, so their test problem performance indicates the extent to which they spontaneously realize the relation between the two sets of problems, correctly identify the appropriate example problem for the test problem, and accurately transfer the learned information to the new situation. The processes of selecting and using the appropriate example problem can be separated by scoring both the type of method subjects used to solve the test problem and the accuracy of their solution, with the former providing information about selection and the latter providing information about use.

I used this transfer paradigm in some of my early research on spatial diagrams (Novick, 1990; Novick & Hmelo, 1994). Subjects (college students) began the experiment by solving three short example problems, in the guise of getting materials ready for a later experiment. A partially constructed representation was provided with each problem. Subjects had to complete the representation and then use it to solve the problem. The example problems were pretested to ensure a high degree of success at both of these tasks. In the control condition, all three example problems used representations (some diagrammatic, some not) that were inappropriate for the test problem. In the experimental condition, the second unrelated example problem was replaced with a problem that used the same type of spatial diagram as was appropriate for the test problem.

After finishing the example problems, subjects were told that the real experiment was about to begin, and they were given the test problem to solve. This problem was presented in writing, without any accompanying

diagram, and no mention was made of diagrams or representations. Subjects were simply asked to solve the problem in the time allotted. To assess the accuracy of the diagram selection process, independent of subjects' ability to use the selected representation, I coded the type of diagram subjects constructed for the test problem.

It is important to note that although the critical example problem and the test problem possessed information structures that specified the same type of diagrammatic tool, the problems were solved using very different procedures. For example, one of the matrix example problems was solved using the relative frequency definition of probability. The matrix test problem was solved using deductive reasoning. Thus transfer of the representational format from the example problem to the test problem could not be aided by procedural similarity between the problems.

I first studied representational transfer using the matrix representation (Novick, 1990). The results of this small-scale study were straightforward. Only 21% of the ($n = 14$) subjects in the control condition used a matrix for the test problem, whereas 75% of the ($n = 16$) subjects in the experimental condition did so. Novick and Hmelo (1994, Experiment 3) replicated these results using a larger ($n \approx 27$ per condition) and more heterogeneous sample. In their study, matrix use for the test problem was coded on a 0/0.5/1 scale, with partial credit being given for constructing an inefficient matrix. The average matrix use score was .76 in the experimental condition and .48 in the control condition.

I mentioned earlier that Novick et al. (1999) categorized the matrix examples that subjects generated into five different subtypes. These subtypes are (a) *binary decision* (e.g., restaurant seating chart to keep track of which tables have been seated), (b) *logical consequence* (e.g., multiplication table), (c) *associated characteristic* (e.g., time schedule), (d) *game* (e.g., checker board), and (e) *graphs, grids, and maps* (e.g., grid for locating archaeological artifacts). In the transfer studies just discussed, both the example and test problems belonged to the *binary decision* subtype. This similarity between the two problems might have been an important factor contributing to the success of representational transfer. Research on the transfer of solution procedures indicates that similarity between the example and test problems influences the likelihood of spontaneous transfer (e.g., Bassok & Olseth, 1995; Holyoak & Koh, 1987; Novick, 1988; Ross, 1984).

In a follow-up study, Novick and Hmelo (1994, Experiment 4) found that the likelihood of using a matrix diagram to represent the information in the test problem did not depend on whether the example and test problems required the same subtype of matrix diagram. Across subjects, the relevant example problem belonged to one of three different subtypes:

binary decision, logical consequence, and *associated characteristic*. The test problem remained the same as in the earlier studies. Use of a matrix diagram for the test problem was statistically equivalent in each of the example conditions ($M = .64$) but was higher in those conditions than in the control condition ($M = .45$). In sum, the results of the three studies reviewed here provide strong evidence for the selection of a matrix diagram for a test problem due to prior exposure to an example problem illustrating the use of that type of tool.

B. Constraints on the Use of Exemplar Knowledge

Nevertheless, there are constraints on the use of exemplar-based reasoning for diagrammatic tool selection. Although people can and sometimes do use exemplar-based reasoning for this purpose, that does not appear to be the primary basis for tool selection. In the ensuing subsections, I will summarize evidence for this conclusion from additional studies of representational transfer and from studies of spatial diagram tool selection.

1. Evidence from Studies of Representational Transfer

In contrast to the results for the matrix diagram, Novick and Hmelo (1994, Experiment 1) found no evidence that subjects would spontaneously transfer either a hierarchy or a network diagram from an example problem to a test problem. The reason for the discrepancy in performance across the three diagrammatic tools appears to be due to differences in superficial similarity between the example and test problems. The two matrix problems were similar in that they involved the same number of values on each dimension. In the binary decision example problem, for example, there were five different shirts and five different pairs of pants. In the test problem, there were five men, five diseases, and five hospital rooms. There was no such numerical similarity between the two hierarchy or network problems. For example, the network example problem described 7 objects and 11 paired relations between them, whereas the test problem described 10 objects and 19 paired relations.

Novick and Hmelo (1994, Experiment 3) found that the degree of numerical similarity between the matrix example and test problems systematically affected the rate of transfer. In all conditions, subjects received the test problem used in the other studies of matrix transfer. The example problem differed across conditions in the number of dimensions (zero, one, or two) that matched the test problem in having five values. When there was no numerical similarity between the two problems (e.g., the example problem had three values on one dimension and seven on the other), there was no transfer, replicating the results for the hierarchy and network diagrams.

The mean matrix use scores were .48 for the control condition and .49 for the zero-matching-dimensions condition. Matrix use increased when the example problem had one matching dimension ($M = .65$) and again when it had two ($M = .76$).

Thus, the degree of superficial similarity between the previous example and the current problem places a constraint on the use of exemplar-based reasoning for tool selection. The only type of superficial similarity that has been investigated in representational transfer is numerical similarity of parameters that are relevant for constructing the diagrammatic tool. It is not known, therefore, whether other types of superficial similarity also would affect representational transfer. It seems likely, however, that cover story similarity would be important as well, because that factor has been shown to affect the retrieval of problems and stories in studies of both procedural (i.e., analogical) transfer and reminding (e.g., Gentner, Ratterman, & Forbus, 1993; Holyoak & Koh, 1987; Novick, 1988; Ross, 1984).

2. *Evidence from Studies of Tool Selection*

One might be inclined to argue that some superficial similarity between the two problems was required in the transfer studies because subjects had to overcome a bias (encouraged by the experimental instructions) to encode the example and test problems as belonging to different experiments with different goals. However, this is unlikely to be the case. Novick et al. (1999) found that even when subjects know that the examples are relevant to the main task of selecting a diagrammatic tool, exemplar-based reasoning does not seem to be their method of choice.

Novick et al. (1999) introduced a new task for studying tool selection that more clearly isolated the process of selecting an appropriate diagram from the subsequent use of the selected diagram to solve the problem. In the main task, subjects were asked to select the most appropriate type of diagram for each of 12–15 short story problems. Solving these problems would have required using analytical or mathematical reasoning. The subjects' task, however, was simply to analyze the relations in the problems sufficiently to be able to select the best type of representation from among the choices provided: one of the three spatial diagrams or "none of the above." Good performance required subjects to analyze the relations in the problems, because the cover stories of the problems did not provide cues as to the appropriate diagram. By constraining subjects to select from among the three spatial diagrams (or none-of-the-above), the new methodology provided a better assessment of people's competence in this arena.

Novick et al. (1999) manipulated the type of information subjects were cued to use to help them make their tool selections. This manipulation was

accomplished by varying, jointly, (a) the preliminary task given to subjects before they completed the selection task and (b) the method of presenting the multiple-choice options on the selection task. In this section, I will only consider one of the three knowledge cue methods. The others will be discussed in the subsequent section on the existence of abstract tool schemas. In one condition, subjects were shown (one at a time) a labeled diagram denoting each tool. These diagrams were similar to the hierarchy, matrix, and network diagrams shown in Figure 1b. The labels were respectively, ''a hierarchy or branching structure,'' ''a matrix with rows and columns,'' and ''a network or system of paths.'' In the preliminary task, subjects were given 4–5 min for each tool to generate as many examples as they could of situations for which that type of representation could be used to organize information. Subjects were told that the purpose of this task was to familiarize them with the types of representations they would need to consider in the subsequent task. Thus, subjects clearly knew that they were expected to use this information later. On the selection task, the multiple-choice options were indicated by the labeled diagrams subjects had seen while generating their examples (plus the phrase ''none of the above'').

I reported earlier that subjects were quite successful at generating a variety of appropriate examples for all three diagrams. If subjects used these examples to guide their subsequent tool selections, then variability in characteristics of the examples generated should predict variability in performance on the selection task. This general hypothesis leads to several specific predictions that were tested by Novick et al. (1999).

First, one would predict that subjects who generated a larger number of appropriate examples should do better at selecting diagrammatic tools because they have more possibilities for successful representational transfer. Second, to the extent that different subtypes of examples can be identified for each diagram, subjects who generated a larger proportion of the possible subtypes should do better because it is more likely that they will have generated a subtype that matches one of the selection problems. The results of two studies ($N = 38$ for Experiment 2 and $N = 20$ for Experiment 1) failed to find evidence for exemplar-based reasoning from either of these predictions. For each of the three diagrams, the correlations between selection accuracy and both the number of examples generated and the number of different subtypes of examples generated ranged from slightly positive to slightly negative and were never reliably different from zero.

In a final effort to find evidence for exemplar-based reasoning, Novick et al. (1999) categorized the test problems into the same subtypes that were used to classify the examples that subjects generated. Then they examined whether generating examples of the subtypes that were used in the selection task improved selection accuracy for the test problems of those subtypes.

For example, they examined whether the number of hierarchy descent examples generated was correlated with selecting the hierarchy tool for the descent problem on the selection task. Seven such correlations were computed (2–3 for each type of diagram) for each study. Again, none were reliably different from zero.

VI. The Existence in Memory of Abstract Tool Schemas

A. Overview and Predictions

The results of these studies indicate that even when the experimental situation makes it easy to use exemplar-based reasoning—indeed, one might argue that it strongly suggests using such a strategy—subjects are unlikely to do so. Rather, it appears that college students possess schemas for the three spatial diagram tools that include at least rudimentary information about their applicability conditions, and that they use that information to guide their tool selections. The information contained in these schemas is relatively abstract, in the sense that it is not tied to specific content domains. For example, one might know that matrices are useful for displaying the relations between items in two sets, whereas networks are useful for displaying the relations between items in a single set, independent of the identities of the items in those sets. I will have more to say about the applicability conditions, and students' knowledge of them, in subsequent sections. Here, I will discuss the results from the remainder of the experiments reported in Novick et al. (1999), which provide preliminary evidence for the existence of abstract tool schemas.

As I mentioned earlier, Novick et al. (1999) manipulated the type of information subjects were cued to use to help them make their diagram selections. Subjects in the general category condition were cued to use whatever general information they had in memory about the three spatial diagrams. The specific tasks used for this purpose will be described shortly. In this condition, the multiple-choice options on the selection task were indicated by labeled diagrams for the three tools (as described earlier), plus the phrase "none of the above." Subjects in the specific example condition were cued to use a single, detailed example problem for each representation that was provided for them. On the selection task, the multiple-choice options were indicated by short labels referring to the content of the example problems they had seen earlier (e.g., "fantasy game example"), plus the phrase "none of the above."

To assess performance on the selection task, Novick et al. (1999) computed what they referred to as a discrimination score. For each representa-

tion, this score was defined as the proportion of times that response was selected when it was appropriate minus the proportion of times it was selected when it was inappropriate. Thus the possible scores ranged from −1 to 1, with 0 indicating chance performance. A comparison of these discrimination scores in the general category and specific example conditions, for problems whose structures do versus do not fit one of the three spatial diagrams, allows inferences about the source of knowledge subjects used to make their tool selections. I will discuss the predictions for the spatial diagram problems first, followed by the predictions for the ''none-of-the-above'' problems.

Consider first the possibility that subjects do not have abstract schemas for the three spatial diagrams that include information about their applicability conditions, so that their tool selections must be guided by exemplar-based reasoning. In this case, subjects in the general category condition should perform poorly on the spatial diagram problems, because whatever information they retrieve from memory does not include knowledge about the applicability conditions for those tools. Moreover, performance should be better in the specific example condition than in the general category condition, because subjects should be able to compare the selection task problems to the readily available, and appropriate, example problems and make their selections based on similarity (Novick, 1990; Novick & Hmelo, 1994).

Now consider the possibility that subjects do have (rudimentary) abstract schemas for the three spatial diagrams that include some information about their applicability conditions. In this case, the opposite pattern of results should be obtained. Although relying on a single appropriate example for each representation to guide tool selection may be effective, relying on more general knowledge should be more effective. As discussed by Holyoak (1985), abstract schemas contain structurally relevant information unattached to specific content domains. Thus, when reasoning using abstract schemas, people only need to consider whether the current problem fits the structure specified by the schema. In contrast, when reasoning with reference to a specific example, they must take into account both the structural similarities and the superficial differences between the two problems. Hence schema-based reasoning should be more effective than exemplar-based reasoning (Holyoak, 1985).

Finally, let us consider the selection task problems for which the correct representation choice is ''none of the above.'' These problems do not belong to a coherent category. The only thing they have in common is that none of the spatial diagrams can be used to describe their structures. Moreover, a task designed to cue knowledge of the three spatial diagrams (general category condition) will not help subjects access any abstract schemas they might have that would be relevant for these problems. Thus, for this set

of problems, there is no basis for predicting any difference between the general category and specific example conditions.

B. Experimental Results

Novick et al. (1999) conducted two experiments, using somewhat different methods for cuing subjects' general knowledge in memory about the three spatial diagrams, to test these predictions. The specific example condition was the same in both studies. In that condition, subjects were given 5 (Experiment 1) or 6 (Experiment 3) min for each tool to study and solve a single detailed example problem illustrating the use of that tool. These example problems were identical or similar to the example problems used in the studies of representational transfer discussed earlier (Novick, 1990; Novick & Hmelo, 1994).[1] In the general category condition used in Experiment 1, subjects' knowledge in memory about the three spatial diagram tools was cued by having them generate as many examples as they could of specific situations for which each type of representation is used to organize information. They were given 5 min to generate examples for each diagram.

As predicted under the hypothesis that subjects possess at least rudimentary abstract schemas for the spatial diagram tools, there was an interaction between problem type and condition. For all three spatial diagrams, accuracy was higher in the general category condition than in the specific example condition (mean discrimination scores of .47 vs. .22, respectively, collapsed across the three diagrams). For the none-of-the-above problems, however, there was no difference between the conditions (means of .42 and .43, respectively).

Although this pattern of results is consistent with the existence of rudimentary abstract schemas for the three spatial diagrams, it is also open to several alternative interpretations. One obvious alternative is that, in the general category condition, subjects used the examples they generated to help them select appropriate representations for the problems on the selection task. According to this hypothesis, these subjects performed better on the selection task than did subjects in the specific example condition, because they had more examples and more different kinds of examples available to guide their reasoning. As I have already discussed, however, the absence of reliable correlations between the examples subjects generated and the diagrams they selected for the selection task problems, both in this experiment and in Experiment 2, refutes this alternative hypothesis.

[1] When subjects in Novick and Hmelo's (1994) transfer studies were told to use these problems to help them construct an appropriate diagram for the test problem, they were typically able to do so. Thus, it is clear that these problems are useful for the present task.

Novick et al. (1999) argued, instead, that the example generation task helped subjects to access the relevant part of their abstract knowledge base, and that they used that abstract knowledge about the properties of the three tools to guide their later diagram selections. Because the abstract properties of a representation (e.g., for the matrix, two sets of items, factorial combination of items across sets, and nondirectional links between items—see Novick & Hurley, 2000) apply equally to any of the (canonical) examples of that representation (e.g., multiplication table, restaurant seating chart, calendar, or police map of a city), retrieving more examples (or more subtypes of examples) should not translate into more abstract properties accessed. The equal applicability of abstract properties to all subtypes of a diagram also would account for why Novick and Hmelo's (1994, Experiment 4) manipulation of subtype similarity had no effect on the likelihood of transferring a matrix diagram from the example problem to the test problem.

Novick et al.'s (1999) Experiment 3 tested two additional alternative explanations for the superiority of tool selection in the general category condition compared to the specific example condition while providing a conceptual replication of Experiment 1. One explanation is that the subjects in the general category condition of Experiment 1 induced some of the abstract structural properties of the diagrams by generalizing across the examples they generated. According to this explanation, representation selections were based on abstract knowledge, but that knowledge was learned in the experiment rather than pre-existing. To test this explanation, subjects in the general category (generate) condition were given only 90 (rather than 300) sec to generate examples of each representation in use. Novick et al. reasoned that if the induction explanation were correct, performance would be worse in this condition than in the similar condition in Experiment 1, and likely worse than in the specific example condition as well, because there would be few examples over which induction might occur and very little time to do any induction.

The second alternative explanation tested in Experiment 3 was a transfer-appropriate processing explanation. According to this explanation, performance was higher in the general category condition in Experiment 1 because the example generation task provided those subjects with practice at linking specific situations to the abstract representation forms (i.e., the labeled diagrams), which arguably is the main goal of the selection task. To test this explanation, subjects in the general category (familiarize) condition were shown each of the labeled diagrams for 20 sec and were asked to familiarize themselves with the diagrams "so that you'll have clearly in mind what each one is like when you do the next task." Novick et al. (1999) reasoned that if the transfer-appropriate processing explanation were correct, performance would be worse in the familiarize condition than in the

specific example condition because the familiarize knowledge cue task and the selection task require very different types of processing. In the example condition, however, subjects worked with story problems in both tasks and should have been able to transfer the knowledge gained from the example problems to the selection task problems (see Novick, 1990; Novick & Hmelo, 1994).

The results of Experiment 3 supported Novick et al.'s (1999) hypothesis that students are able to reason about the applicability of the three spatial diagrams to story problems based on abstract schematic (i.e., general category) knowledge. In addition, they argued strongly against both the induction and transfer-appropriate processing explanations. Performance on the spatial diagram problems in the general category (generate) condition was comparable to that in Experiment 1 despite the reduction in example generation time from 5 min to 1.5 min. Moreover, accuracy was similar in the general category (generate) and general category (familiarize) conditions, and reliably higher in those two conditions than in the specific example condition. Novick et al. provided several compelling arguments against the possibility that this pattern of results could have been due to the use of misleading examples in the specific example condition. As in Experiment 1, there were no reliable differences in selection accuracy across conditions for the none-of-the-above problems.

These results are shown in Figure 2. The filled squares and solid line indicate performance on the problems requiring one of the spatial diagrams. All three tools showed the same pattern of performance. The open squares and dashed line indicate performance on the problems for which none-of-the-above was the correct answer. It is particularly noteworthy that cuing subjects' knowledge by having them think about the labeled diagrams for 20 sec each (general category familiarize condition) led to much greater accuracy at selecting spatial diagram tools for the story problems than did spending 6 min studying and successfully solving a relevant example story problem for each diagram (specific example condition).

C. Conclusions

Novick et al.'s (1999) results strongly suggest that typical college students—those without advanced training in mathematics or computer science—possess rudimentary abstract schemas for hierarchies, matrices, and networks that contain some information about their applicability conditions. These schemas presumably were induced over the course of normal school and life experiences, as these diagrams are fairly common in American culture. For example, hierarchies are used for family trees, military power structures, chain letters, mathematical factoring, decision trees, biological

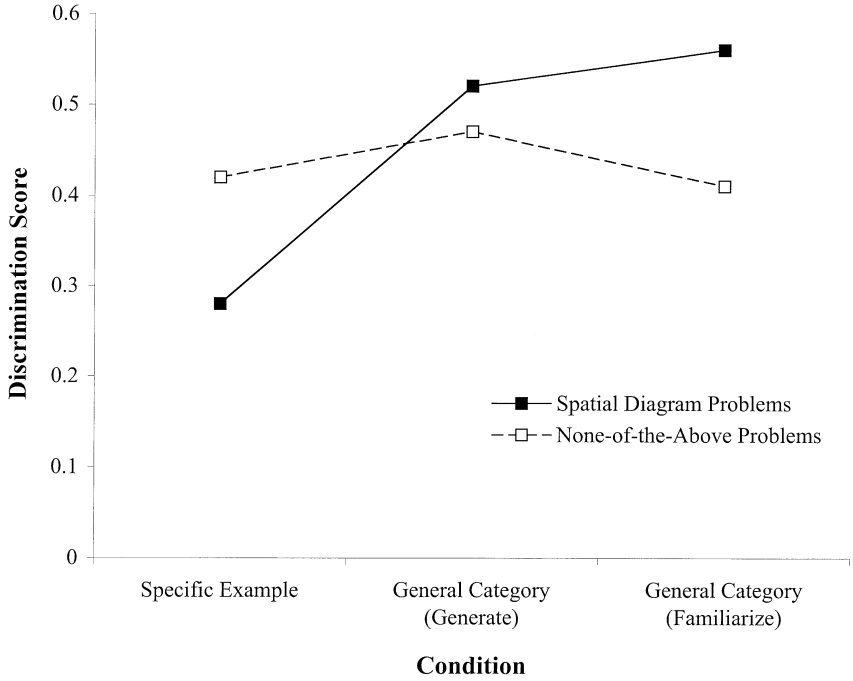


Fig. 2. Average discrimination scores for problems requiring one of the three spatial diagrams or not as a function of knowledge-cuing condition in Experiment 3 of Novick et al. (1999). (From Evidence for abstract, schematic knowledge of three spatial diagram representations, by L. R. Novick, S. M., Hurley, & M. Francis, 1999, *Memory & Cognition, 27,* p. 303. Copyright 1999 by the Psychonomic Society, Inc. Adapted with permission.)

classification, and tournaments; matrices are used for multiplication tables, time schedules, grade books, seating charts, and police maps of a city; and networks are used for airline route maps, computer networks, alliances among nations, flow charts, and football plays.

Like the permission schema studied by Cheng and Holyoak (1985) and the schemas for discrete and continuous models of change studied by Bassok and Olseth (1995), which are also believed to be induced over ordinary life experiences, the diagrammatic tool schemas seem to be at an intermediate level of generality (Novick et al., 1999). They are neither highly general, as are algebra and logic, nor are they specific to a narrow content domain, as are Newton's Second Law and a circuit diagram. Rather, each type of spatial diagram is best suited for a particular type of relational structure, regardless of the content domain in which that structure is embedded. For example, a hierarchy is appropriate whenever items in a domain can be

organized into levels and there is only one route between pairs of items (Novick & Hurley, 2000).

The problem structure best suited for each type of tool can be easily summarized by a ''prototype'' spatial diagram (Novick et al., 1999). Thus solvers could begin to construct rudimentary schemas by using the prototype diagrams as ''hooks'' to which other relevant information could be attached. This information might include the applicability conditions for the diagram, examples of situations in which the diagram has proven useful in the past, and rules for constructing the diagram. I will consider the applicability conditions in the next section and the possibility of construction rules in a later section.

VII. A Structural Analysis of the Three Spatial Diagram Tools

A. Overview

As noted earlier, because superior reasoning is observed when the display format and the structure of the environment are consistent (Sanfey & Hastie, 1998), effective problem solving depends on being able to select a representational tool whose structure matches that of the information to be represented. Selection of an appropriate representational tool, in turn, depends on knowing the applicability conditions for a variety of different tools. Just as hammers, screwdrivers, and wrenches work best in certain types of situations, so do hierarchies, matrices, and networks. McKim (1980) made this point more generally:

> To be sure, as screwdrivers are sometimes used as hammers, chisels, or prybars by the ill-equipped or inept, [visual-thinking] strategies can also be applied inappropriately and ineffectively. As with tools, [visual-thinking] strategies must be skillfully chosen and applied in order for them to be effective. (p. 186)

In this section, I will summarize Novick and Hurley's (2000) structural analysis of hierarchies, matrices, and networks. They proposed that the three spatial diagrams are defined by 10 structural properties. These properties can be thought of as variables, and their values for the three diagrams constitute the applicability conditions for those tools. For example, value ''m'' on property 1 is one of the applicability conditions for a matrix representation. To my knowledge, no one has previously attempted a comprehensive structural analysis of these diagrams. A number of theoreticians have proposed a variety of structural features of individual diagrams, but there has been little discussion of the relations among the properties or the diagrams. Some of the influences of prior work on Novick and Hurley's

comprehensive structural analysis of the three spatial diagrams will be mentioned here. Full details may be found in Novick and Hurley.

It is important to note that the 10 structural properties of hierarchies, matrices, and networks are interrelated rather than orthogonal. Each property specifies an identifiable aspect of representation structure that can, to some extent, be discussed and investigated separately. But because the representations are integral wholes, the property values cannot be separated from each other and mixed and matched at will. Similarly, one cannot easily combine, for example, the head of a hammer, the handles of a pair of pliers, and the blade of a saw. To facilitate discussion of the spatial diagram properties, Novick and Hurley (2000) grouped them into three categories: basic structure, details about items and links, and the potential for movement of information through the diagram.

B. Properties Related to the Basic Structure of the Diagram

Four properties provide information about the basic structure of the diagram. The first of these, *global structure,* specifies the general form of the representation. Most abstract diagrams have a general form that allows one to quickly identify the type of diagram, without conducting a detailed analysis of its individual parts. For example, we can easily distinguish a bar graph from a line graph or a pie chart from a Venn diagram based on their general forms. Novick and Hurley (2000) argued that the same is true of matrices and hierarchies. In particular, the matrix representation has a general structure in which there is a cross-classification of all of the values of one variable with all of the values of another variable (also see Bertin, 1980; Day, 1988; Inhelder & Piaget, 1964). In contrast, the hierarchy is organized into levels (Bertin, 1981). A single root node, usually located either at the top or right of the diagram, branches out to subsequent levels such that the identities of the objects at one level depend on the identities of the objects at the preceding level (Vessey & Weber, 1986). The network is unlike most other abstract diagrams in that it is characterized by the lack of any formal structure. Novick and Hurley argued that this makes sense, because mathematically the network representation is the general form (Chartrand, 1985), of which both matrices and hierarchies are special cases.

Although the different diagrams may be recognized by their global structures without careful attention to the individual elements of which they are composed, actually constructing the diagrams requires knowledge of their basic units. Thus the second property specifies the *building block* of each diagram. The building block of a matrix is a box (or cell) that denotes the combination (i.e., intersection) of a specific value on one variable with

a specific value on the other variable (Inhelder & Piaget, 1964). The building block of a network is two items and a (directional or nondirectional) relation between them. The building block of a hierarchy is three items connected by two directional relations, arranged as an upside-down or sideways V (Elmasri & Navathe, 1994).

The third property in the basic structure category specifies the *number of sets* of objects or concepts for which the diagram is optimally suited. This property is important for distinguishing the matrix and network diagrams, but it is less diagnostic of the hierarchy. Whereas a matrix is most appropriate when there are two sets of objects or concepts (Bertin, 1981; Day, 1988), the network is most appropriate when there is only one set (Bertin, 1981). Hierarchies can accommodate a variety of numbers of sets, depending on how one construes a set and on the specific content being represented (see Novick & Hurley, 2000).

Finally, the *item/link constraints* property specifies any constraints that the diagrams themselves impose on which items may be linked together, independent of the content being represented. This property is closely related to the global structure property, because the differing global structures are due in part to the different constraints on linking items. The network diagram places no constraints on which items may be linked: In principle, any item may be linked to any other item. Of course, the semantics of a particular situation being represented will specify certain links and not others, but the constraints come from the content domain rather than from the diagram. In contrast, both the matrix and the hierarchy inherently place some constraints on the linking of items. The structure of the matrix diagram specifies that the items within a set (e.g., the items labeling the rows) may not be linked. The structure of the hierarchy diagram specifies that there may not be direct links between items at the same level or between items in nonadjacent levels.

C. Properties Providing Details about the Items and Links in the Diagram

Three properties provide details about the items and links depicted in the three diagrams. The *item distinguishability* property specifies how the items in a set differ. In particular, it specifies whether they differ in name only or in some other more structured way. For the network diagram, all of the items have identical status; that is, they are indistinguishable except by name (Bertin, 1981). That is also true of the items in each set for a matrix. (Elmasri & Navathe, 1994). This is not to say that some order cannot be imposed on the items depicted in these two diagrams. Clearly, that is possible. However, the ordering is imposed by the semantics of the situation

being represented; it is not inherent in the diagrams themselves. In contrast, the hierarchy does inherently impose an ordering on the items. In particular, the items at different levels differ in status, whereas those at the same level have identical status.

The *link type* property specifies differences in the types of relations between items that the three diagrams are best suited to convey. A matrix is best suited to situations in which the links between items are purely associative (Elmasri & Navathe, 1994). That is, the links in a matrix specify some type of nondirectional association between pairs of items. A hierarchy, on the other hand, is well suited to representing directional relations. In particular, the links in a hierarchy are directional such that processing flows from one end of the diagram to the other—that is, from the root to the leaves or the reverse. Because the network is the general form that subsumes both a matrix and a hierarchy, it is more flexible in terms of the types of relations it can accommodate. The links between items in this representation can be associative or unidirectional or bidirectional (Epp, 1995).

Not only do the three diagrams differ with respect to the types of relations they are best able to convey, but they also differ with respect to their ability to convey information about the *absence of a relation* between items. The matrix diagram is superior in this regard, because it can represent the absence of a relation between two items directly through the placement of a special mark (e.g., an "X") in the relevant cell (Bertin, 1980, 1981; Tversky, in press). At the other extreme, the network diagram is not at all suited to providing this information. Rather, the absence of a relation between two items must be computed because there are no constraints on which items may be linked. The hierarchy diagram is intermediate in its ability to convey information about the absence of a relation between two items. In certain cases, the absence of a relation is indicated implicitly due to the constraints on which items may be linked. For example, the absence of links between items at the same level is not explicitly marked, but it is predictable *a priori.* But in other cases—namely when items in adjacent levels are not linked—the absence of a relation must be computed.

D. Properties Related to the Potential for Movement of Information through the Diagram

The remaining three properties are related to providing information about the potential for the movement of information through the diagram. The *linking relations* property is a local property that focuses on the links going into and out of individual nodes in the diagram. It specifies whether the links between items are one-to-many, many-to-one, or both. This property is especially important for distinguishing the network and hierarchy repre-

sentations. In a network diagram, the links between items may be both one-to-many and many-to-one (Elmasri & Navathe, 1994; D. Schwartz, 1993). Thus, a given node in a network may have multiple lines (i.e., relations) both entering and leaving it. In contrast, in a hierarchy diagram, the relations are either one-to-many (Elmasri & Navathe, 1994; D. Schwartz, 1993) or many-to-one, but not both. That is, it is either the case that one line enters each node and multiple lines leave it or that multiple lines enter and one line exits. Whether the relations are encoded as one-to-many or many-to-one depends on whether one moves through the hierarchy from the root to the leaves or vice versa. Finally, in a matrix diagram, both one-to-many and many-to-one relations are possible for a given item, but this does not seem to be a particularly salient aspect of this diagram.

In contrast to the linking relations property, the *path* property is a global property that focuses on the representation as a whole. Quite simply, it specifies whether the diagram is optimized for representing paths connecting subsets of at least three items. The hierarchy and network are (Chartrand, 1985); the matrix is not.

Finally, the *traversal* property specifies structural differences in the types of paths that are possible. It is important for distinguishing the hierarchy and network diagrams, because these are the types of diagrams that can easily convey information about paths. It does not really make sense to talk about traversing the matrix diagram. For the hierarchy diagram, there is only one path connecting any pair of nodes, ''A'' and ''B'' (Bertin, 1981; Chartrand, 1985; Epp, 1995). Thus, closed loops are not possible. In contrast, for the network diagram, closed loops are possible and thus there may be multiple paths connecting pairs of nodes (Epp, 1995). The traversal property is highly related to the linking relations property. In particular, the local (item-level) differences in linking relations generate the global differences in traversal.

E. Concluding Remarks and Unresolved Issues

The 10 property values for each diagram constitute the proposed applicability conditions for that diagram. The uniqueness (in most cases) of the property values across diagrams supports the claim that hierarchies, matrices, and networks are distinct representation types that are optimized for different problem structures. The idea of optimization here is critical. Although it is possible, for example, to use a matrix to represent the relations among objects in a single set or a network to represent the relations between objects in two sets, there are disadvantages to doing so (Chartrand, 1985; Novick & Hurley, 2000). Thus, Novick and Hurley's (2000) structural analysis of the three spatial diagrams specifies the representation prototypes (e.g., the prototypical matrix).

Although many everyday situations and formal problems fit the prototypes, some situations likely include structural features that specify more than one type of diagram. It is possible to construct hybrid diagrams for these situations, such as a hierarchical tree with just a few links that create closed loops or a matrix with some links drawn between items within a set (e.g., between items labeling the columns). It will be an important topic for future research to examine problem solvers' behavior when they are presented with situations whose structural features overlap two or more of the diagrams.

It is also important to note that the properties likely are not equally important for each type of diagram. A property that is very important for determining that a matrix would be the most appropriate diagram to use for a particular situation may be less important for determing that a hierarchy is most appropriate, and vice versa. Put another way, the properties likely vary in their importance (or diagnosticity) for specifying the three diagrams. For example, Novick and Hurley (2000) suggested that the number-of-sets property is more important than the traversal property for specifying a matrix diagram, whereas the reverse is the case for the hierarchy diagram.

Diagnosticity, as just defined, specifies the strength of the directed link from the property to the diagram. It may be important to consider whether the links between the properties and the diagrams are symmetric or asymmetric: Is the link going from a property to a particular diagram equal in strength to the link going from that diagram to that property? Novick and Hurley (2000) hypothesized that, in general, the links are asymmetric, although their data cannot address this issue. For example, the link from the matrix diagram to the matrix value of the linking relations property may not be equivalent in strength to the link from that property value to that diagram. Although a matrix is easily applicable to a situation with both one-to-many and many-to-one relations, that property value does not seem highly diagnostic of the matrix diagram. That is, knowing that a situation involves both one-to-many and many-to-one relations does not seem to strongly suggest the use of matrix. More generally, a particular diagram might be very good at accommodating a certain property value (applicability), yet one might not place high weight on that property value when making a decision about whether to use that type of diagram (diagnosticity). Additional research is needed to address this issue.

VIII. Empirical Investigation of the Structural Analysis and Students' Diagram Knowledge

A. Overview and Methods

Novick and Hurley (2000) argued that for the purpose of assessing the adequacy of their structural analysis, it was important to test students who

would be expected, *a priori,* to be relatively knowledgeable about the three types of diagrams (e.g., because of their coursework in college). Such students would provide the best chance of (a) finding support for, or contradictions to, the structural analysis, as well as (b) uncovering additional distinguishing properties that had not been considered. Therefore, they tested 23 students who were nearing the end of their degree programs (i.e., mostly juniors and seniors) in computer science (n = 12) or in the combination of secondary education and mathematics (n = 11). Novick and Hurley's data also provided evidence concerning (a) variability in these students' knowledge across the properties and the property values for each type of diagram and (b) how the properties are organized in these students' knowledge bases.

Although the computer science and math education students were a good choice for evaluating the structural analysis, their data likely do not provide a good view of the knowledge possessed by the typical college student. Therefore, I will also report here, for the first time, data from a group of more typical undergraduates who are more similar to the subjects used in the previous studies reviewed here (Novick, 1990; Novick & Hmelo, 1994; Novick et al., 1999). These students were 12 juniors and seniors whose most advanced math class was first-year calculus, which is the modal level of math coursework completed by Vanderbilt undergraduates.

The materials for this study consisted of 18 short (3–5 sentence) scenarios set in a medical context. This context was chosen for two reasons. First, it is broad enough to enable scenarios to be written for all three types of diagrams while still providing some comparability of content across diagrams. Second, most specific medical contents do not preferentially cue one type of diagram versus another. Thus, subjects would have to attend to the structural information, or guess, to make their response. Structural information was provided for each scenario by writing one or two sentences to focus on the value for one property for one diagram. For example, the following scenario focuses on the matrix value for the global structure property. For illustrative purposes here, the property-focusing statement is italicized.

> In the psychiatric ward of a certain hospital, each patient sees only one doctor, who is responsible for diagnosis and treatment. A researcher is interested in determining whether patients would receive different diagnoses from different doctors. *Therefore, she selected a group of newly admitted patients and asked all of the staff psychiatrists to submit a diagnosis for each patient in the group*. The department chair would like a diagram showing each doctor's diagnosis for each patient.

Overall, 6 of the 10 properties were focused: global structure, number of sets, item/link constraints, link type, linking relations, and traversal. One

criterion for choosing a property was that it had ''positive'' values for at least two of the three diagrams, so that it was possible to write at least two scenarios focusing on that property (this rules out the absence of a relation property). A second criterion was that at least two of the positive values had to be different (this rules out the path property). The building block property was excluded because it was difficult to write scenarios that focused purely on that property, without very strongly suggesting either or both of the global structure and link type properties. The item distinguishability property was excluded because it seemed less important for discriminating among the three diagrams than the six remaining properties. The six selected properties were used in 2–5 scenarios each. Across the 18 scenarios, each diagram had one of its property values focused six times.

Each scenario ended with a sentence stating someone's desire for a diagram to represent the information presented. Subjects were given a choice between two of the three representations of interest, which were presented as labeled diagrams as in Novick et al. (1999). The two representations had different values on the property used in the scenario so that the structural analysis would make a strong prediction as to the most appropriate diagram. One dependent measure was the percent of scenarios for which subjects' choices conformed to the predictions of the structural analysis. After selecting a diagram for a scenario, subjects were asked to justify their choice of the selected diagram and to explain why they thought the unselected diagram was less good. These verbal responses were audiotaped, transcribed, and coded.

A strength of the choice methodology is that it allowed Novick and Hurley (2000) to tap into diagrammatic knowledge that might be to some extent implicit, whereas the verbal protocols required explicit access to such knowledge. One strength of the protocol methodology is that it provided quite strong evidence for the structural properties that subjects did in fact use in their diagram justifications. Another strength of this methodology is that it allowed for the possibility of discovering additional structural properties that were not proposed initially. It is important to note that no additional properties were mentioned by subjects.

B. Support for the Structural Analysis

Novick and Hurley's (2000) study yielded strong support for the proposed structural analysis with respect to both subjects' choices of the best type of diagram for each scenario and their verbal justifications for their choices. The subjects overwhelmingly selected the diagrams predicted by the structural analysis, with 88% of their choices conforming to the predictions. This conclusion holds for the individual diagrams as well as for the individual

focused properties. These results indicate that subjects' implicit knowledge about the applicability conditions for the three diagrams is consistent with the structural analysis.

In examining subjects' verbal justifications for their selections, Novick and Hurley (2000) found more direct support for the structural analysis. They were able to reliably code 9 of the 10 hypothesized structural properties. Subjects either did not know the item distinguishability property or did not consider it to be relevant for choosing a diagram for the scenarios. For the nine codable properties, subjects almost never made statements contradicting the structural analysis (e.g., saying that the network is most useful when there are two sets of objects rather than one). Moreover, all of the subjects mentioned a majority of the 10 properties proposed in the structural analysis somewhere in their protocols ($M = 80\%$). All nine codable properties were mentioned by more than half of the subjects, and eight of the nine were mentioned by at least 70% of the subjects. The darker bars in Figure 3 show the proportions of the computer science and math education subjects (combined) who (appropriately) referred to each of the nine codable properties for any of the three diagrams. For example, 74% of these subjects referred to the number of sets property for at least one of the three types of diagrams.

Overall, 21 of the 27 codable property values (9 properties $\times$ 3 representations; 78%) were supported by evidence from the verbal justifications. For the matrix, the supported properties are global structure, number of sets, building block, link type, absence of a relation, and path. For the network, the supported properties are global structure, number of sets, building block, link type, item/link constraints, linking relations, path, and traversal. For the hierarchy, the supported properties are global structure, building block, link type, item/link constraints, linking relations, path, and traversal. These results indicate that the computer science and math education students' explicit knowledge about the applicability conditions for the three spatial diagrams supports Novick and Hurley's (2000) analysis of the structures of those diagrams.

The representation-choice and verbal-justification results provide strong support for the structural analysis as a whole and no disconfirming evidence for any of the proposed property values.[2] However, the status of 9 of the 30 proposed property values remains uncertain. Three of these property values were for the item distinguishability property, which could not be coded reliably (in part due to the paucity of references to it by subjects).

[2] However, the results did suggest a boundary condition on the number of sets value for the network representation. In particular, a single set of objects implies a network diagram, as predicted, but only if the number of items in the set is not too large. Further discussion of this boundary condition may be found in Novick and Hurley (2000).

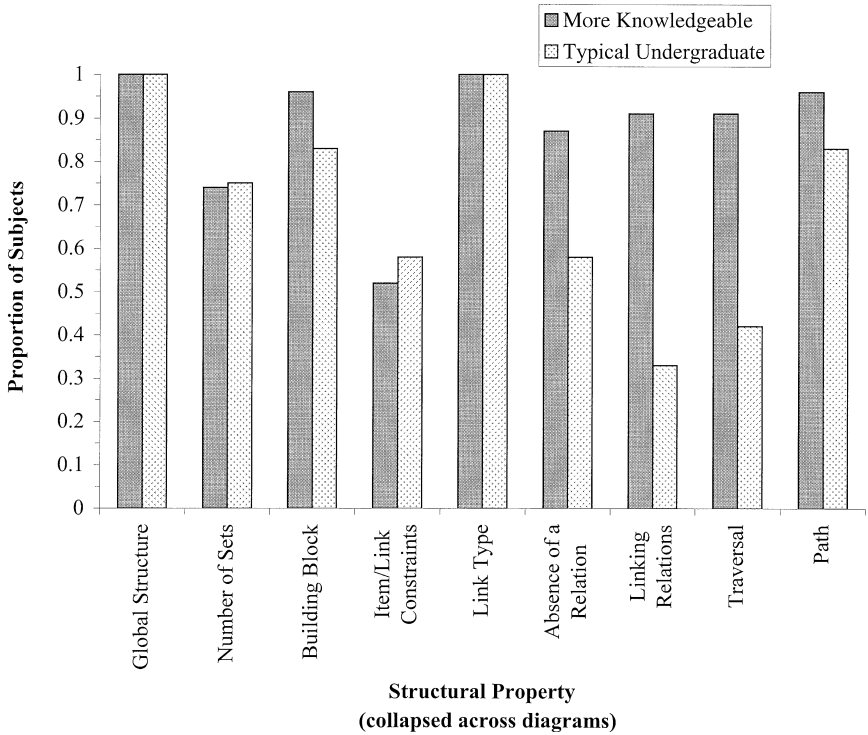


Fig. 3. The proportion of more knowledgeable and typical students who referred to each of nine structural properties, collapsed across the three spatial diagrams.

The remaining six property values were (a) hierarchy number of sets and absence of a relation, (b) matrix item/link constraints, linking relations, and traversal, and (c) network absence of a relation. The uncertain status of these property values could be resolved in a number of different ways. Novick and Hurley (2000) argued that two resolutions seem most likely. One possibility is that the uncertain property values are important in some situations, but they were not needed to justify the diagram choices for their scenarios. A second possibility is that not all of the properties provide important applicability conditions for all of the diagrams. Novick and Hurley forced their structural analysis to have a factorial structure, in which each of the properties specified an applicability condition for each of the diagrams. It may be the case, however, that this factorial structure is only a useful approximation. Additional studies that directly examine the diagnosticity of each property for each diagram will be needed to resolve this issue.

C. Comparison of Diagram Knowledge as a Function of Expertise

It is clear that the computer science and math education students tested by Novick and Hurley (2000) are quite knowledgeable about the structural properties and about the individual applicability conditions for hierarchies, matrices, and networks. It seems reasonable to believe, however, that this level of diagrammatic knowledge may not be typical of college students in general. Most students have encountered these three diagrams in the media and sometimes in (K–12) classes, but any abstract knowledge they have is likely to have been induced from these ordinary, everyday experiences rather than to have been received through direct instruction. Thus, it is likely to be less extensive. To address this issue, I now report the data from 12 more typical students who completed the diagram selection and justification tasks. These students did quite well on these tasks, although not as well as the computer science and math education students.

In selecting the most appropriate type of diagram for the 18 scenarios, these students' choices conformed to the predictions of the structural analysis 84% of the time. This is slightly, but not reliably, lower than the performance of the more knowledgeable students ($M = 88\%$), $F(1, 33) = 2.27$, $p > .14$, $MSE = .0041$. Table I shows the mean level of conformity to the predictions for each type of diagram and for each focused property separately for the two more knowledgeable subject groups combined and for the typical undergraduates. It is striking how similar the two groups of subjects are in most respects. Even the typical undergraduates have fairly good implicit knowledge of the applicability conditions that were directly tested in the 18 scenarios.

Nevertheless, there are a few notable differences between the two groups. The largest difference occurred for the linking relations property. Although the typical students' diagram selections for these scenarios conformed to the predictions of the structural analyses more often than would be expected by chance, $t(11) = 3.45$, $p < .01$, $SE = .25$, suggesting that they do have some implicit knowledge of this property, their performance was considerably worse than that of the more knowledgeable students (75% vs. 94%). This group difference will reappear in the justifications data. The typical undergraduates also performed less well on the scenarios whose structures indicated either a hierarchy or a matrix diagram, although the differences were not as extreme. Whereas the performance of the more knowledgeable students was near ceiling for these diagrams ($M = 96\%$), that of the typical students was merely very good ($M = 88\%$).

In sum, the subject groups are roughly comparable in terms of their tacit knowledge of five of the six structural properties that were focused on in

TABLE I

MEAN LEVELS OF CONFORMITY TO THE PREDICTIONS FOR EACH DIAGRAM AND EACH STRUCTURAL PROPERTY SEPARATELY FOR THE MORE KNOWLEDGEABLE AND TYPICAL STUDENTS, AND COMPARISONS OF THE TWO SUBJECT GROUPS

Category	No. of scenarios	Subject population		Comparison of subject groups[a]
		Math ed./comp. sci.	Typical undergraduate	
Representation				
Hierarchy	6	.96	.87	$F = 9.04$, $MSE = .0069$, $p < .01$
Matrix	6	.96	.89	$F = 3.60$, $MSE = .0123$, $p < .07$
Network	6	.70	.76	$F = 1.06$, $MSE = .0276$, $p > .30$
Property				
Global structure	5	.94	.90	$F < 1$, $MSE = .0168$
Number of sets	2	.80	.83	$F < 1$, $MSE = .0617$
Link type	2	.93	.96	$F < 1$, $MSE = .0267$
Item/link constraints	3	.68	.75	$F < 1$, $MSE = .0377$
Linking relations	3	.94	.75	$F = 9.04$, $MSE = .0322$, $p < .01$
Traversal	3	.91	.86	$F < 1$, $MSE = .0248$

[a] All F statistics have 1 and 33 degrees of freedom.

the scenarios, at least as measured by performance on this diagram selection task. With respect to explicit knowledge about the properties, however, as indicated by subjects' verbal justifications for their diagram choices, it is clear that the typical undergraduates are at a disadvantage.

The lighter bars in Figure 3 show the justifications given by the more typical students, collapsed across the three diagrams. For example, 75% of these subjects referred to the number of sets property in their justifications for at least one of the three diagrams. It is clear that even the typical students possess some rudimentary information about the applicability conditions for the three spatial diagrams, as all of the properties were mentioned by at least one-third of these subjects. However, their knowledge is not as extensive as that possessed by the math educators and computer science majors. On average, the more knowledgeable students referred to 8.00 different properties in their justifications, whereas the typical undergraduates referred to only 6.33 different properties, $F(1, 33) = 20.85$, $p < .001$, $MSE = 1.0505$.

It is clear from Figure 3, however, that the magnitude of the difference between the two subject groups depends on the property. There are sizable differences between the proportions of subjects in each group who referred to three of the properties—absence of a relation, linking relations, and traversal—and smaller differences for references to two properties—building block and path. There is little or no difference between the groups for four properties—global structure, number of sets, link type, and item/link constraints. Using properties as the random factor, therefore, the difference between the more knowledgeable and typical students (means of 0.87 and 0.70, respectively) is only marginally reliable, $t(8) = 2.23$, $p < .06$, $SE = .077$.

The five properties that show differences between the two groups are pretty much evenly split between those whose values were focused on for some of the diagrams and those whose values were never focused on. In fact, the two properties with the largest differences—linking relations and traversal—were focused on in several scenarios. These results lend credence to the claim that the typical undergraduates really are less knowledgeable about at least some of the applicability conditions for the three spatial diagrams. Even when important applicability conditions were isolated and focused on in the scenarios, the typical students did not necessarily use them to justify their diagram selections. This finding persists when the property justifications are examined for each diagram separately, rather than collapsed across diagrams as in Figure 3.

Figures 4–6 show subjects' use of the hypothesized property values for the hierarchy, network, and matrix diagrams, respectively, separately for subjects in the two groups. The data for the computer science and math education students were previously reported in Novick and Hurley (2000).

I discussed earlier in this chapter several possible interpretations of the property values that were not mentioned. Here, I will focus on a comparison of the two subject groups. There are some striking similarities in the results across diagrams.

For the hierarchy diagram, the more knowledgeable students referred to an average of 5.91 property values, compared to a mean of 4.67 for the typical students. This difference is statistically reliable, $F(1, 33) = 6.91$, $p < .02$, $MSE = 1.7725$. Looking at Figure 4, differences between the two subject groups are evident for four properties—link type, linking relations, traversal, and path, with linking relations and traversal showing the largest differences. All four of these properties also showed differences between the subject groups for the network diagram (see Figure 5), with the largest differences again being observed for the linking relations and traversal properties. These two properties are critical for distinguishing the hierarchy and network diagrams, so it makes sense that if the typical students referred to them less often for one of these diagrams, they would refer to them less

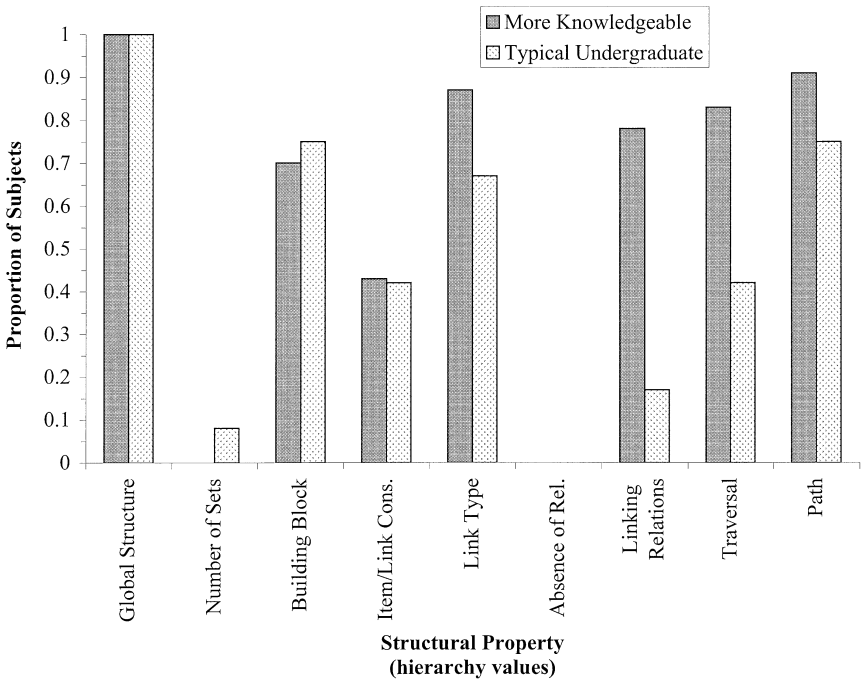


Fig. 4. The proportion of more knowledgeable and typical students who referred to each of nine structural properties for the hierarchy diagram.

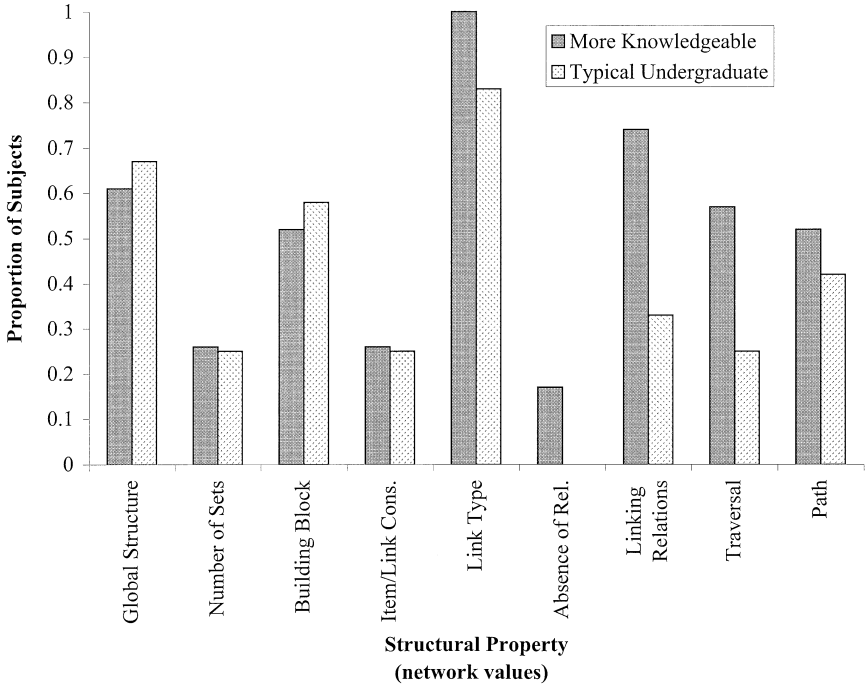


Fig. 5. The proportion of more knowledgeable and typical students who referred to each of nine structural properties for the network diagram.

often for the other as well. Recall that the typical students also were less likely to select the appropriate diagram when the scenario focused on the linking relations property. Consistent with the justifications results, the difference in selection performance between the two subject groups was more pronounced for the hierarchy- and network-focused linking-relations scenarios than for the matrix-focused scenario (differences of .33, .16, and .08, respectively).

Focusing on the network diagram now, the more knowledgeable students referred to an average of 5.09 property values compared to a mean of 4.17 for the typical students. Although this difference is only marginally reliable, $F(1, 33) = 3.70$, $p < .07$, $MSE = 1.8028$, I suspect that there is a real difference between the subject groups in their knowledge of this diagram as well. This is because all of the group differences for individual properties that were observed for this diagram also were observed for one of the other two diagrams. I have already noted the similarities in the group differences for the hierarchy and network diagrams (for link type, linking relations,

traversal, and path). The remaining property that showed a group difference for the network—absence of a relation—also showed a group difference for the matrix diagram (see Figure 6).

The other properties that showed a group difference for the matrix diagram are global structure, building block, item/link constraints, and path. On average, the more knowledgeable students referred to 4.74 matrix property values in their justifications, compared to a mean of 3.67 for the typical students, $F(1, 33) = 7.65$, $p < .01$, $MSE = 1.1849$.

In sum, all of the properties except number of sets showed greater frequency of use by the computer science and math education students compared to the typical students for at least one of the three diagrams. Four properties showed the typical students to be at a disadvantage for two diagrams–link type, absence of a relation, linking relations, and traversal. The path property showed small but consistent differences for all three diagrams. Although both the selection and justification data indicate that the typical students have some knowledge concerning a majority of the

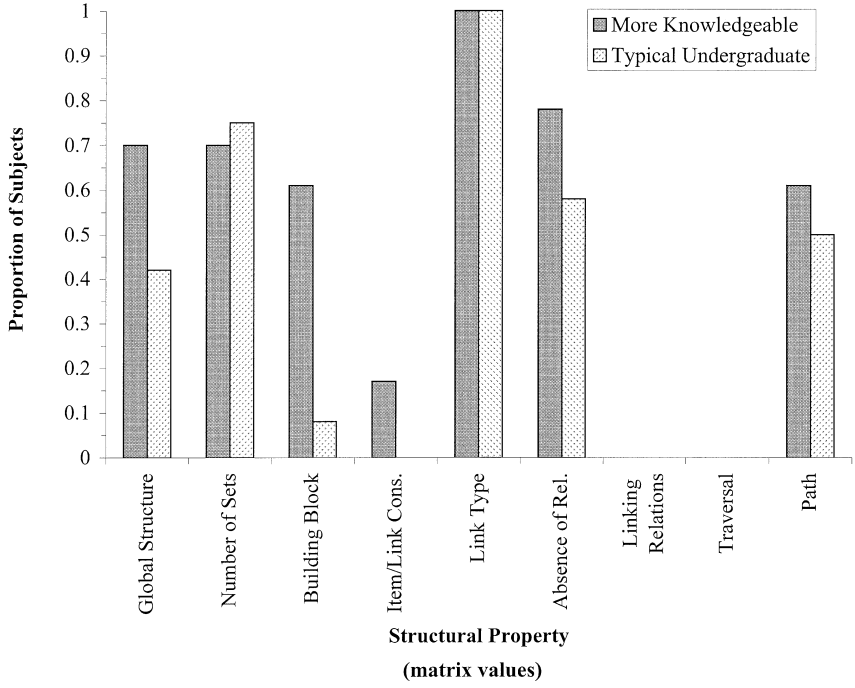


Fig. 6. The proportion of more knowledgeable and typical students who referred to each of nine structural properties for the matrix diagram.

property values that were coded, it appears to be less extensive than that possessed by the math education and computer science students.

IX. The Structure of Students' Knowledge about the Three Spatial Diagrams

A. Overview

The data concerning students' knowledge of the structural properties of the three spatial diagrams that I have reported so far consider each property in isolation. As I noted earlier, however, the structural properties are not orthogonal. Rather, many of the properties are conceptually related. For example, linking relations is the local property that enables the global differences in traversal, and traversal provides a more detailed specification of the types of paths that are possible. Thus, analyses that consider the properties in isolation provide an incomplete picture of students' knowledge. To investigate the organization of property knowledge (i.e., the interrelations among the properties), Novick and Hurley (2000) conducted a factor analysis of the justifications given by the 23 more knowledgeable subjects. I will summarize their results for these students and present the results of a comparable analysis for the typical undergraduates.

B. Computer Science and Math Education Students

The factor analysis conducted by Novick and Hurley (2000) included nine dependent measures, with each measure being references to one of the structural properties (all except item distinguishability). The replications factor in this analysis was the different scenarios. For each scenario for each subject, Novick and Hurley counted the total number of references to each property across the two diagrams presented with that scenario. Then, for each property (i.e., each dependent measure), they computed a mean for each scenario collapsed across subjects. Thus, the data entering into the factor analysis were the frequencies with which the nine properties were used to justify the representation choice for each scenario.

Because there were 9 dependent variables and only 18 scenarios, Novick and Hurley (2000) computed property means separately for the math education and computer science students. This enabled the factor analysis to be conducted on 36 cases rather than only 18. They obtained a four-factor solution that accounted for 72% of the variability in the data. Interpretations of the individual factors are presented in the Novick and Hurley paper. Figure 7 summarizes the observed relations among the properties. In this figure, properties that loaded together on the same factor with the same

sign (with a magnitude of at least .30 in absolute value) are connected by a solid black line. Properties that loaded together on the same factor but with the opposite sign (again with a magnitude of at least .30 in absolute value) are connected by a dotted grey line. This network representation of the factor analysis results clarifies the underlying organization of the properties.

As Figure 7 indicates, the results of the factor analysis of the more knowledgeable students' justifications for their diagram choices supported several of the expected interrelations among the properties. Consider, first, the right side of the figure. As predicted, linking relations, traversal, and path form a coherent cluster. Traversal is positively related to both path and linking relations. As mentioned earlier, the traversal property provides a detailed specification of the types of paths that are possible, and differences in the local linking relations property are responsible for the more global differences in traversal. These three properties are perceived as being quite distinct from the remaining properties, as the remaining 10 links coming out from those nodes represent negative relations.[3] Moreover, those negative links go to four of the six remaining properties (all except global structure and item/link constraints). According to the structural analysis, this cluster of properties is most important for distinguishing the network and hierarchy diagrams. The factor scores for the scenarios indicate that the computer science and math education students used this cluster of properties for this purpose, hence the question written above this cluster on Figure 7.

Absence of a relation, link type, and number of sets also form a fairly coherent cluster. These three properties were all positively linked to each other at least once in the factor analysis, and there were no negative relations among them. The positive association between link type and absence of a relation was expected. In some respects, these two properties are opposite sides of the same coin. One side tells the types of relations between items that a diagram can easily represent; the other tells whether the diagram can represent explicitly the absence of a relation between items. The factor scores for the scenarios indicated that this cluster of properties was used to determine whether a matrix would be the most appropriate type of diagram, hence the question written below this cluster in the figure.

Note, however, that this cluster of properties is not as distinct as the linking-relations/traversal/path cluster, because both absence of a relation

[3] Note that there are two negative links between traversal and link type. This means that these two properties loaded together, with the opposite sign, on two of the four factors. Thus, this negative relation between the two properties is redundantly coded in the factor analysis results.

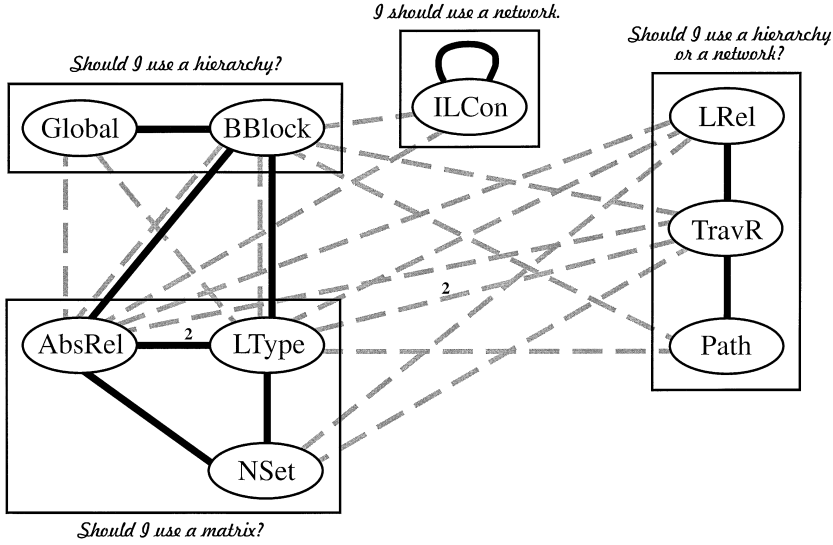


Fig. 7. The organization of property knowledge in memory for the computer science and math education students. The property labels are abbreviated as follows: Global, global structure; BBlock, building block; ILCon, item/link constraints; AbsRel, absence of a relation; LType, link type; NSet, number of sets; LRel, linking relations; TravR, traversal; and Path, path.

and link type have positive as well as negative relations to building block. The positive relation between building block and link type was expected. The building block property indicates the arrangement of items and relations that is used as the basic unit for constructing the diagram. The link type property specifies the nature of the relations appearing in the basic unit. These two properties were used together in justifications when a matrix provided the best structural model for the scenario. The positive link between absence of a relation and building block does not have as coherent an explanation based on the factor scores for the scenarios.

Global structure and building block form a third, smaller cluster. This association was expected, because building block specifies how the global structure is constructed. The factor scores for the scenarios indicated that these properties were used together by subjects to justify the hierarchy as being the most appropriate type of diagram, hence the statement written above this cluster.

Finally, item/link constraints forms a cluster all to itself. This property was used almost solely to justify the choice of a network diagram, hence the statement written above that property in the figure. Item/link constraints was expected to be grouped with global structure, because the global struc-

tures of the diagrams are determined in part by the constraints on which items can be linked. I will discuss the results for this property after presenting the factor analysis for the typical undergraduates.

C. Typical Undergraduates

The factor analysis for the typical undergraduates included the same nine dependent measures, which were the frequencies with which each of the structural properties was used to justify the diagram selections. The replications factor was the 18 scenarios, with means computed across the 12 typical students for each property. Clearly, 18 cases are not very much data for conducting a factor analysis on nine dependent variables. Thus, one must be somewhat more cautious in interpreting the results of this analysis compared to the earlier one.

I obtained a four-factor solution that accounted for 79% of the variability in the data. As Novick and Hurley (2000) did for the more knowledgeable subjects, I will only consider property variables that have factor loadings of at least .30 in absolute value. Figure 8 summarizes the results for the typical students using the notation adopted for Figure 7: Properties that loaded on the same factor with the same sign are connected by a solid black line, whereas those that loaded on the same factor but with the opposite sign are connected by a dotted grey line.

The first thing to notice is that the clusters indicated in Figure 8 are highly similar to those indicated in Figure 7. In particular, for both subject groups, linking relations, traversal, and path form a coherent cluster; absence of a relation, link type, and number of sets form a cluster; and global structure and building block group together. These similarities provide increased confidence that the properties that are grouped together really do have important conceptual similarities, as predicted by the analysis of the diagram structures.

There are some notable differences between the figures, however. First, although item/link constraints forms a cluster by itself for the more knowledgeable students, it is grouped with global structure and building block for the typical students. As indicated earlier, the positive link between item/link constraints and global structure was expected. Novick and Hurley (2000) interpreted the results from one of the factors for the more knowledgeable students as providing suggestive evidence for such a link for the network diagram. The factor analysis for the typical students suggests that this link may be general rather than diagram specific.

A second difference concerns the summary questions or themes attached to each cluster in the two figures, which were derived by examining the characteristics of the scenarios that had high versus low scores on each

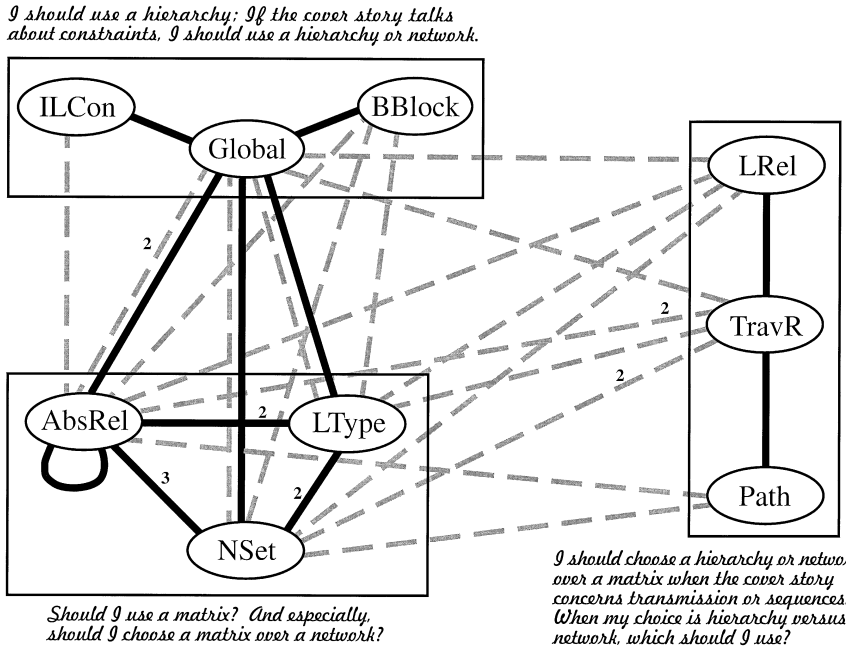


Fig. 8. The organization of property knowledge in memory for the typical undergraduates. The property labels are abbreviated as follows: Global, global structure; BBlock, building block; ILCon, item/link constraints; AbsRel, absence of a relation; LType, link type; NSet, number of sets; LRel, linking relations; TravR, traversal; and Path, path.

factor. For the computer science and math education students (Figure 7), all of the questions and themes are focused on the diagrams. They involve (a) consideration of the merits of a particular diagram regardless of whether it is selected or rejected (matrix), (b) justification of a specific diagram (network, hierarchy), and (c) comparison of two specific diagrams (hierarchy vs. network). In contrast, the questions and themes for the typical students (Figure 8) sometimes are focused on the diagrams but sometimes are more contextualized. For example, two of the themes include specifying the type of diagram to select when the content of the scenario involves constraints or transmission. This difference between the two groups is consistent with what is known about the acquisition of expertise (e.g., Chi et al., 1981).

Third, there are more links (both positive and negative) between the two clusters on the left side of Figure 8 for the typical students than there are for the comparable clusters for the more knowledgeable students (Figure 7). Overall, there are 35 links in Figure 8 compared with 26 in Figure 7.

Moreover, many more links are redundant (i.e., were generated by the loadings on more than one factor) for the typical students than for the more knowledgeable students (seven vs. two). Thus, extra redundant links account for half of the additional links (five of nine) observed for the typical students. These differences between the subject populations suggest that the knowledge representation for the computer science and math education students is more distinct than that for the typical students. This is consistent with the differences in the summary questions and themes attached to the clusters for the two groups of students.

D. Concluding Remarks and an Unresolved Issue

Consistent with Novick and Hurley's (2000) structural analysis of hierarchies, matrices, and networks, the results of the factor analyses indicate that both subject populations view the structural properties as being interrelated rather than orthogonal. Moreover, as shown in Figures 7 and 8, the organization of property knowledge was quite similar (although not identical) for the two subject groups in terms of the specific properties that were grouped together. The observed clusters support several interrelations among the properties that were expected *a priori,* thereby providing additional support for the structural analysis.

An important unresolved issue, discussed in some detail by Novick and Hurley (2000), concerns the generality of the relations depicted in Figures 7 and 8 across the three diagrams. As Novick and Hurley noted, the results of the factor analyses likely reflect those property relations that generalize across at least two of the three diagrams, because those are the ones that are most likely to survive when the data are collapsed across representations. Although some pairs of properties probably are closely related for all three diagrams (e.g., global structure and building block), other pairs of properties may be related for only one or two diagrams. For example, traversal may be closely related to path for the network and hierarchy but not the matrix, and item/link constraints and number of sets may be related only for the matrix. Diagram-specific relations are less likely to emerge from factor analyses of the collapsed data. Given that each type of diagram was used in only 12 scenarios and not all of the properties were mentioned for all of the diagrams, it is not possible to address this issue with the present data.

X. Constructing Spatial Diagrams

The bulk of this chapter has concerned the process of selecting a spatial diagram that fits the structure of the situation to be represented. As I

noted at the outset, however, simply selecting an appropriate diagrammatic representation is not sufficient for reasoning and problem solving. One must also be able to construct a specific instance of the selected type of diagram so that the important concepts and relations in the problem are accurately mapped onto the components of the diagram. There has not been much research on the diagram construction process to date.

Novick and Hmelo (1994) argued that even if the appropriate type of diagram is selected using exemplar-based reasoning, people construct the diagram from scratch from general principles. The diagram use errors from all four of their experiments are consistent with this hypothesis. If the test problem diagrams were constructed by adapting the diagram provided with an earlier example problem (i.e., using exemplar-based reasoning), one would expect to see a variety of errors of commission, in which inappropriate features from the example diagram would be incorporated into the new diagram. Contrary to this prediction, however, subjects' errors in constructing the diagrams for the test problems were primarily ones of omission, in which items or relations from the test problems were simply missing from their diagrams.

Novick and Hmelo (1994) suggested that the principles or rules for constructing each diagram might be stored in abstract diagram schemas. For example, one's network schema might specify two ''rules'' for constructing such a diagram: (a) Draw a single circle to represent each object or concept in the problem; (b) whenever there is some kind of link or relation between two objects or concepts, draw a line connecting the corresponding two circles. This proposal is consistent with the results of more recent research (Novick et al., 1999), reviewed earlier, supporting the existence of abstract schemas for the three spatial diagrams. It makes sense that if people have such abstract schemas, those schemas would include (rudimentary) information about construction principles as well as (rudimentary) information about applicability conditions. Obviously, one conclusion does not necessarily entail the other, but they are mutually consistent.

XI. Discussion

A. Initial Issues in Diagrammatic Reasoning

The cognitive toolbox contains a variety of representational and procedural tools for reasoning and problem solving. A number of researchers have argued that pictorial formats for problem representation, especially abstract diagrams, are key instruments in the toolbox for thought (e.g., Day, 1988; Dufour-Janvier et al., 1987; Kindfield, 1993/1994; Larkin & Simon, 1987;

Polya, 1957). In this chapter, I have discussed three abstract spatial diagrams—hierarchies, matrices, and networks—that are known to facilitate comprehension, reasoning, and problem solving (e.g., Bartram, 1980; Day, 1988; Guri-Rozenblit, 1988; Novick & Hmelo, 1994; Polich & Schwartz, 1974; Scanlon, 1989; S. Schwartz, 1971; Vessey & Weber, 1986).

Diagrammatic tools only enhance performance, however, when the display format and the structure of the environment are consistent (Sanfey & Hastie, 1998). Thus, effective problem solving depends on being able to select a representational tool whose structure matches that of the information to be represented. It is important, therefore, to understand both the structures of diagrams and the information people use to select the most appropriate diagram for particular situations. These two issues have been the primary focus of my research. I have adopted a competence perspective with respect to assessing students' knowledge. That is, I have sought to determine what people know about these tools and when to use them, under conditions that are relatively favorable for them to exhibit their knowledge.

B. Summary of Research on Diagram Selection

I reported several studies that examined whether people's decisions concerning the most appropriate type of representation to use typically are made by comparing the current situation to (a) one or more specific situations encountered previously or (b) abstract information contained in one or more diagram schemas. The results indicated that although college students sometimes use exemplar-based reasoning to select an appropriate spatial diagram (Novick, 1990; Novick & Hmelo, 1994), typically the selections are based on information contained in rudimentary diagram schemas (Novick et al., 1999). Novick et al. (1999) suggested that these schemas were induced over the course of normal school and life experiences, as hierarchies, matrices, and networks are fairly common in American culture.

If people use diagram schemas to help them select appropriate representations, the schemas presumably would need to contain information about the applicability conditions for their respective diagrams (Novick & Hurley, 2000), as well as rules for constructing the diagrams (Novick & Hmelo, 1994). Novick and Hurley (2000) identified 10 structural properties on which they hypothesized that hierarchies, matrices, and networks differ. These properties can be thought of as variables, and their values for the three diagrams constitute the applicability conditions for those tools. For example, the global structure property specifies the general form of each diagram. The matrix has a general structure in which there is a cross-classification of all of the values of one variable with all of the values of another variable. In contrast, the hierarchy is organized into levels such

that the identities of the objects at one level depend on the identities of the objects at the preceding level. The network is unlike most other abstract diagrams in that it is characterized by the lack of any formal structure. To take another example, the traversal property specifies structural differences in the paths connecting subsets of items. For the hierarchy diagram, there is only one path connecting any pair of nodes, because closed loops are not possible. In contrast, for the network diagram, there may be multiple paths connecting pairs of nodes, because closed loops *are* possible. Finally, it does not really make sense to talk about traversing the matrix diagram, because that diagram is not good for representing paths.

Novick and Hurley (2000) reported diagram choice and verbal justifications data from students who were nearing the end of their degree program in either computer science or the combination of secondary education and mathematics. These data provided strong support for their structural analysis of the three spatial diagrams. First, the subjects overwhelmingly selected the diagrams whose property values were focused on in the scenarios. Second, all of the subjects mentioned a majority of the 10 properties at some point in justifying their diagram choices. Looking at the data from the perspective of the properties, all nine properties that Novick and Hurley were able to code from the protocols were mentioned by more than half of the subjects. Third, the results of a factor analysis of subjects' justifications were largely consistent with the structural analysis, which proposes that the properties are interrelated rather than orthogonal. In particular, the clusters of properties that emerged from the factor analysis primarily reflected expected interrelations among the properties.

In addition to validating Novick and Hurley's (2000) structural analysis of the three spatial diagrams, these data indicated that the computer science and math education students were quite knowledgeable about both the structural properties in general and the diagram-specific applicability conditions. I also reported data in this chapter for a new set of students who were more typical of the undergraduate population. Both the selection and justification data indicated that the typical students had some knowledge concerning a majority of the applicability conditions (i.e., property values). However, their knowledge was less extensive than that possessed by the math education and computer science students. Also, although the results of the factor analysis for the typical students were largely consistent with those for the more knowledgeable students, the typical students' knowledge of the structural properties appeared to be more contextualized and less distinct. These comparisons across the two subject groups are consistent with observed expertise differences in other domains (e.g., Chi et al., 1981; Schoenfeld & Herrmann, 1982).

C. Issues on the Horizon

The research presented here has provided an important core of information concerning hierarchies, matrices, and networks as well as students' knowledge about these diagrams. Nevertheless, the results are preliminary in some respects. As discussed by Novick and Hurley (2000), additional research along three lines would be especially useful in terms of extending the frontiers of our knowledge in this domain. First, more focused testing of the proposed structural analysis of the three spatial diagrams is needed, using new methodologies to complement the existing evidence from the representation-choice and verbal-protocol tasks. This research also should investigate the possibility of differences across the three diagrams in (a) the diagnosticity of the properties and (b) the interrelations among the properties.

Having a clear understanding of the applicability conditions for hierarchies, matrices, and networks is necessary for furthering our understanding of two additional important issues. As I noted earlier, it is likely to be the case that many complex real-world situations are hybrids in the sense that their structures overlap with more than one of the spatial diagrams. A holistic comparison of such a situation to one's diagram schemas is unlikely to suffice for diagram selection because of the conflicting cues. It makes sense to predict that in the face of such a conflict, both the relative number of properties favoring each diagram and the relative diagnosticities of the property values would influence people's diagram selections. Specifying such predictions for particular situations obviously depends on knowing how diagnostic each property is of each diagram.

Finally, returning to the idea of abstract diagrams as cognitive tools, it is clear that the whole point of selecting a diagram to use for representing the structural relations in a particular situation is that using diagrams often facilitates reasoning and problem solving. Once we have a clear understanding of the applicability conditions for the three spatial diagrams, as well as students' knowledge of them, we can begin to ask how problem solvers use the information in their (rudimentary) diagram schemas to construct spatial diagrams. And we can begin to understand the factors that affect success at constructing diagrams so that the relations in the problem appropriately map onto components of the diagram.

D. Concluding Remarks

Diagrams have long played an important role in human societies. For example, the Bayeux Tapestry, created in the eleventh century, records the events leading up to and including the Battle of Hastings. In the fifteenth century, da Vinci contributed literally thousands of drawings dealing with

anatomical, mechanical, geographical, and other subject matters. The first edition of Wilson's biology textbook, *The Cell,* published in 1896, has almost 43 illustrations per 100 pages (Maienschein, 1991). The third edition, published in 1925, has 47 illustrations per 100 pages, and these diagrammatic representations are more abstract and schematic. McKim (1980) has noted that diagrams are found on the blackboards of almost every academic department. In our increasingly technical and technological society, diagrams are likely to become even more important. For example, even the traditionally text-driven domain of computer programming now includes visual languages, as a number of researchers have argued that visual representations need to be integrated into programming languages (see Myers, 1990, and Whitley, 1997, for recent reviews of this literature). The LabVIEW™ language (National Instruments, 1998), for example, expresses the flow of information through the program using a notation that resembles circuit diagrams. Ultimately, the knowledge gained from research on diagrammatic reasoning, including investigations of abstract spatial diagrams, may help us to design better diagrammatic tools and to teach people how to use them effectively.

Acknowledgments

I would like to thank Sean Hurley for helpful discussions of many of the issues considered in this chapter and for comments on an earlier version of the manuscript. I would also like to thank Doug Morse and Kirsten Whitley for examples of the use of diagrams in society throughout history.

References

Bartram, D. J. (1980). Comprehending spatial information: The relative efficiency of different methods of presenting information about bus routes. *Journal of Applied Psychology, 65,* 103–110.

Bassok, M., & Olseth, K. L. (1995). Object-based representations: Transfer between cases of continuous and discrete models of change. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1522–1538.

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science, 4,* 372–378.

Bertin, J. (1980). The basic test of the graph: A matrix theory of graph construction and cartography. In P. A. Kolers, M. E. Wrolstad, & H. Bouma (Eds.), *Processing of visible language, Vol. 2* (pp. 585–604). New York: Plenum Press. [Translated from the French by P. G. Allen.]

Bertin, J. (1981). *Graphics and graphic information-processing* (W. J. Berg & P. Scott, Trans.). Berlin: de Gruyter. (Original work published 1977)

Butler, D. L. (1993). Graphics in psychology: Pictures, data, and especially concepts. *Behavior Research Methods, Instruments, & Computers, 25,* 81–92.

Chartrand, G. (1985). *Introductory graph theory*. New York: Dover.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17,* 391–416.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152.

Day, R. S. (1988). Alternative representations. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 261–305). San Diego, CA: Academic Press.

Dufour-Janvier, B., Bednarz, N., & Belanger, M. (1987). Pedagogical considerations concerning the problem of representation. In C. Janvier (Ed.), *Problems of representation in the teaching and learning of mathematics* (pp. 109–122). Hillsdale, NJ: Erlbaum.

Duncker, K. (1945). On problem-solving, *Psychological Monographs, 58* (Whole No. 270).

Elmasri, R., & Navathe, S. B. (1994). *Fundamentals of database systems* (2nd Ed.). Menlo Park, CA: Addison-Wesley.

Epp, S. S. (1995). *Discrete mathematics with applications* (2nd Ed.). Pacific Grove, CA: Brooks/Cole.

Gentner, D., Ratterman, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology, 25,* 524–575.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12,* 306–355.

Greeno, J. G. (1977). Process of understanding in problem solving. In N. J. Castellan Jr., D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive theory* (Vol. 2, pp. 43–83). Hillsdale, NJ: Erlbaum.

Greeno, J. G. (1983). Conceptual entities. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 227–252). Hillsdale, NJ: Erlbaum.

Guri-Rozenblit, S. (1988). The interrelations between diagrammatic representations and verbal explanations in learning from social science texts. *Instructional Science, 17,* 219–234.

Hayes, J. R., & Simon, H. A. (1977). Psychological differences among problem isomorphs. In N. J. Castellan, D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive theory* (Vol. 2, pp. 21–44). Hillsdale, NJ: Erlbaum.

Hegarty, M., Carpenter, P. A., & Just, M. A. (1991). Diagrams in the comprehension of scientific texts. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 641–668) New York: Longman.

Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language, 32,* 717–742.

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 19, pp. 59–87). NY: Academic Press.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15,* 332–340.

Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child*. (E. A. Lunzer & D. Papert, Trans.). New York: Harper & Row. (Original work published 1959).

Iwasaki, Y. (1995). Problem solving with diagrams. In J. Glasgow, N. H. Narayanan, & B. Chandrasekaran (Eds.), *Diagrammatic reasoning* (pp. 657–667). Menlo Park, CA:AAAI Press/MIT Press.

Kindfield, A. C. H. (1993/1994). Biology diagrams: Tools to think with. *The Journal of the Learning Sciences, 3,* 1–36.

Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science, 14,* 511–550.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11,* 65–99.

Levin, J. R. (1989). A transfer-appropriate-processing perspective of pictures in prose. In H. Mandl & J. R. Levin (Eds.), *Knowledge acquisition from text and pictures* (pp. 83–100). Amsterdam: Elsevier.

Lohse, G. L., Biolsi, K., Walker, N., & Rueter, H. H. (1994). A classification of visual representations. *Communications of the ACM, 37 (12),* 36–49.

Maienschein, J. (1991). From presentation to representation in E. B. Wilson's *The Cell. Biology and Philosophy, 6,* 227–254.

Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology, 82,* 715–726.

McGuinness, C. (1986). Problem representation: The effects of spatial arrays. *Memory & Cognition, 14,* 270–280.

McKim, R. H. (1980). *Thinking visually: A strategy manual for problem solving*. Belmont, CA: Wadsworth.

Myers, B. A. (1990). Taxonomies of visual programming and program visualization. *Journal of Visual Languages and Computing, 1,* 97–123.

National Instruments (1998). *LabVIEW™ User Manual* [Software manual]. Austin, TX: Author.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 510–520.

Novick, L. R. (1990). Representational transfer in problem solving. *Psychological Science, 1,* 128–132.

Novick, L. R. (1995). Some determinants of successful analogical transfer in the solution of algebra word problems. *Thinking and Reasoning, 1,* 5–30.

Novick, L. R., & Hmelo, C. E. (1994). Transferring symbolic representations across nonisomorphic problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1296–1321.

Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 398–415.

Novick, L. R., & Hurley, S. M. (2000). *To matrix, network, or hierarchy, that is the question*. Manuscript submitted for publication.

Novick, L. R., Hurley, S. M, & Francis, M. (1999). Evidence for abstract, schematic knowledge of three spatial diagram representations. *Memory & Cognition, 27,* 288–308.

Novick, L. R., & Morse, D. L. (in press). Folding a fish, making a mushroom: The role of diagrams in executing assembly procedures. *Memory & Cognition*.

Paige, J. M., & Simon, H. A. (1966). Cognitive processes in solving algebra word problems. In B. Kleinmuntz (Ed.), *Problem solving: Research, method, and theory* (pp. 51–119). New York: Wiley.

Polich, J. M., & Schwartz, S. H. (1974). The effect of problem size on representation in deductive problem solving. *Memory & Cognition, 2,* 683–686.

Polya, G. (1957). *How to solve it* (2nd ed.). Princeton, NJ: Princeton University Press.

Reed, S. K. (1987). A structure-mapping model for word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 124–139.

Ross, B. H. (1984). Remindings and their effects in learning a cognitive skill. *Cognitive Psychology, 16,* 371–416.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 629–639.

Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 456–468.

Sanfey, A., & Hastie, R. (1998). Does evidence presentation format affect judgment? An experimental evaluation of displays of data for judgments. *Psychological Science, 9,* 99–103.

Scanlon, D. A. (1989). Structured flowcharts outperform pseudocode: An experimental comparison. *IEEE Software, 6 (Sept.),* 28–36.

Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 484–494.

Schwartz, D. L. (1993). The construction and analogical transfer of symbolic visualizations. *Journal of Research in Science Teaching, 30,* 1309–1325.

Schwartz, S. H. (1971). Modes of representation and problem solving: Well evolved is half solved. *Journal of Experimental Psychology, 91,* 347–350.

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load as a factor in the structuring of technical material. *Journal of Experimental Psychology: General, 119,* 176–192.

Tversky, B. (in press). Spatial schemas in depictions. In M. Gattis (Ed.), *Spatial schemas and abstract thought.* Cambridge: MIT Press.

Vessey, I., & Weber, R. (1986). Structured tools and conditional logic: An empirical investigation. *Communications of the ACM, 29 (1),* 48–57.

Wertheimer, M. (1982). *Productive thinking* (enlarged ed.). Chicago: The University of Chicago Press. (Original work published 1959)

Whitley, K. N. (1997). Visual programming languages and the empirical evidence for and against. *Journal of Visual Languages and Computing, 8,* 109–142.

Winn, W. (1989). The design and use of instructional graphics. In H. Mandl & J. R. Levin (Eds.), *Knowledge acquisition from text and pictures* (pp. 125–144). Amsterdam: Elsevier.

This Page Intentionally Left Blank

REINFORCEMENT AND PUNISHMENT IN THE PRISONER'S DILEMMA GAME

Howard Rachlin, Jay Brown, and Forest Baker

I. Introduction

The prisoner's dilemma is a much studied game (Axelrod, 1980, dubs it the *E. coli* of social psychology) that opposes the interests of each individual player to the interests of the players as a group. The prisoner's dilemma is important because it mirrors many dilemmas of everyday life. The name, ''prisoner's dilemma,'' comes from the common occurrence wherein two criminal suspects are arrested and separated by the police. Either suspect has the choice of confessing or holding out. The evidence against them is such that if both suspects hold out (*cooperate* with each other) each will receive a moderately short sentence while if both suspects confess (*defect* from their mutual interests) each will receive a moderately long sentence. However, if one confesses (and testifies against the other) while the other holds out, the confessor will get off scot-free, while the one who had held out will receive a very long sentence. In such a situation, each prisoner may reason as follows:

> If the other prisoner holds out I should confess because then I will get off scot-free rather than serve a moderately short sentence. If the other prisoner confesses I should also confess because then I will receive a moderately long sentence rather than a very long sentence. Thus, whatever the other prisoner does, I should confess.

However, if both prisoners (or their lawyers) reason this way, both will confess and each will receive a moderately long sentence, whereas if they

THE PSYCHOLOGY OF LEARNING AND MOTIVATION, VOL. 40

Copyright © 2001 by Academic Press.
All rights of reproduction in any form reserved.

0079-7421/00 $35.00

had both held out they would each have received a moderately short sentence. Hence, the dilemma.

Figure 1a shows a contingency matrix for a hypothetical two-person prisoner's dilemma game as it is studied in the laboratory. The players (A and B) each choose between cooperation, C, and defection, D. (In the laboratory, the value-laden labels for choice responses, "cooperation" and "defection," are not generally used. Participants choose "X" or "Y," "left" or "right," as in the experiments to be described here.) If both players choose cooperation, each earns 3 points. If both players choose defection, each earns 2 points. However, if one player cooperates while the other defects, the cooperator earns only 1 point while the defector earns 4 points. Player A, in deciding which alternative to choose, may reason as follows:

> If Player B cooperates I should defect because then I will earn 4 points instead of 3; if Player B defects I should also defect because then I will earn 2 points instead of 1; thus, in either case, regardless of what Player B chooses, I should defect.

As in the actual prisoner's dilemma described above, this reasoning is perfectly logical. It is what A's lawyer would advise. However, as above, if Player B reasons in the same way as Player A, both players will defect and earn 2 points each, whereas if both had cooperated (disobeying their lawyers) each would have earned 3 points.

Behavior in a prisoner's dilemma game such as that illustrated in Figure 1a depends strongly on factors extrinsic to the game itself. Mutual cooperation is more likely when the players know each other well or when they are permitted to come to an agreement in advance (Caporael, Dawes, Orbell, & van de Kragt, 1989). That is, there is a strong interpersonal component to the game. Consider the prototypical prisoner's dilemma game described above. The prisoners would be less likely to confess if each knew that the other had previously held out in prior arrests and if they were both likely to be faced with the same dilemma in the future. In other words, behavior in the prisoner's dilemma game depends strongly on experience in prior games and on expectations about future games. For this reason most empirical studies of the prisoner's dilemma game expose participants to repeated rather than one-shot games (Komorita, 1994; Rapoport & Chammah, 1965). In fact, given people's expectations based on prior experience, there is no such thing as a one-shot game. All choices, even those made in the laboratory, occur within a social context.

Studies of the repeated prisoner's dilemma by economists and mathematicians have focused on the interplay of various strategies. The theoretical basis for these studies has been maximization of expected utility. Theories have frequently been tested by performance, not of human subjects, but

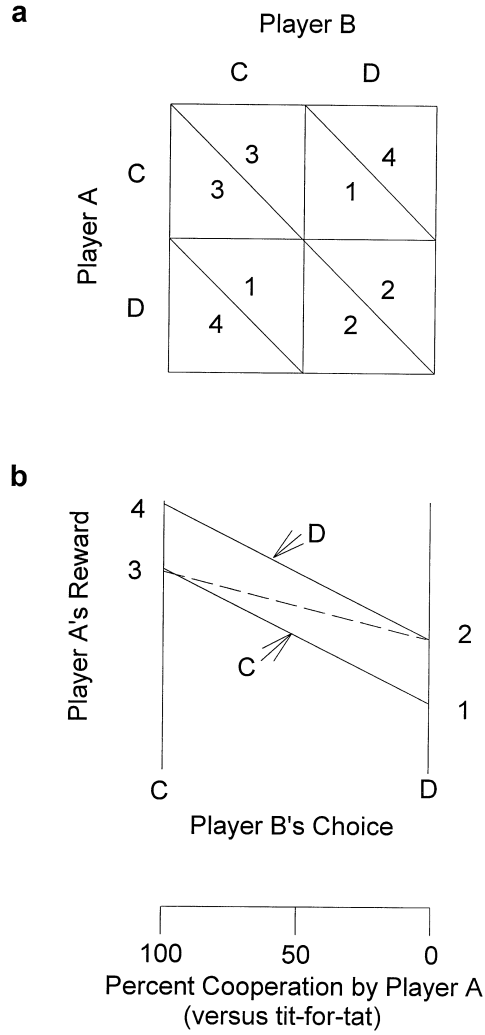


Fig. 1. (a) Matrix of rewards in a typical two-person prisoner's dilemma game. (b) The same game from player A's viewpoint. The left and right axes show A's alternatives if B cooperates (C) or defects (D). The upper and lower diagonal solid lines connect A's rewards for defecting or cooperating. The diagonal dashed line shows A's average reward if B plays tit-for-tat over a series of trials as a function of A's percent cooperation (scale goes from right to left).

of mathematical or computer models. Cognitive and social psychologists, on the other hand, have tended to study the behavior of human subjects in the laboratory but still focus on strategy formation. Communication, agreement, and trust between players are central in cognitive and social studies (Caporael et al., 1989). Prior experience with the game itself is important in these studies only to the extent that it is said to result in central representations of agreement and trust. Verbal assurances are said to give rise to central representations which then cause behavior. The present studies take a different tack, analyzing laboratory studies with human subjects in terms of contingencies of reinforcement and punishment. From the behavioral viewpoint taken here, the relevant context of a choice is not an internal event at all but prior experience or "history of reinforcement." Verbal agreements or assurances, like laboratory instructions to subjects, themselves represent past experience and need not be internally represented in any coherent way to have an effect. That is, such agreements and assurances are treated as discriminative stimuli functionally the same as (albeit more complex than) the red or green lights on the key that a pigeon pecks.

Figure 1b illustrates the contingencies of the prisoner's dilemma game of Figure 1a from Player A's viewpoint. The left vertical line shows Player A's alternatives if Player B cooperates; the right vertical line shows Player A's alternatives if player B defects. The upper solid line connects the pair of defection (D) alternatives; the lower solid line connects the pair of cooperation (C) alternatives. The fact that D is higher than C shows again that on the current trial the reward for defection always exceeds that for cooperation. Figure 1b shows that the game presents Player A with two motives: first, *Player A wants to defect;* second, *Player A wants Player B to cooperate.* The figure also reveals that, if and when these two motives conflict, the latter should dominate the former. If Player B cooperates, the minimum Player A can earn is 3 points, whereas if Player B defects the maximum Player A can earn is 2 points. In other words, Player B's present cooperation reinforces Player A's prior choice, whereas Player B's present defection punishes Player A's prior choice. In a much-studied repeated-game strategy called tit-for-tat, Player B reinforces Player A's cooperation on trial n by cooperating on trial $n + 1$ and punishes Player A's defection on trial n by defecting on trial $n + 1$. In laboratory experiments with repeated games, with one player (Player B) playing tit-for-tat, the other player (Player A) generally learns to consistently cooperate.

The tit-for-tat strategy essentially converts the social dilemma to an intrapersonal dilemma resembling problems of individual self-control. This can be seen by again referring to Figure 1b. If Player B is playing tit-for-tat (mimicking Player A's trial-n choice on trial $n + 1$) then the label on

the horizontal axis, ''Player B's Choice,'' can be replaced by the label, ''Player A's Previous Choice.'' Player A would always gain one point on the *current* trial by defecting but would pay for this gain on the next trial by having to choose between 1 and 2 points rather than 3 and 4 points. That is, the opposing motives: *Player A wants to defect* and *Player A wants Player B to cooperate,* are replaced, under tit-for-tat, by *Player A wants to defect on this trial* and *Player A wants to have cooperated on the previous trial.* Of course, over a succession of trials, these motives are incompatible. Similarly, the alcoholic's motives: *I want to drink today* and *I wish I had abstained from drinking yesterday* are incompatible over succeeding days. The alcoholic's body (in the form of hangovers and health in general) and social environment (in the form of family relationships and job performance) are essentially playing tit-for-tat with him.

The dashed line in Figure 1b represents Player A's average earnings as a function of the percent of trials on which Player A cooperates, given that Player B plays tit-for-tat. Cooperating on 100% of the trials (left axis) earns Player A 3 points per trial; defecting on 100% of the trials (right axis) earns Player A 2 points per trial; alternating between defection and cooperation earns Player A 4 points for defection and 1 point for cooperation, averaging 2.5 points per trial. Even a single defection (except on the very last trial of a series) will cost Player A points in the long run. A way to increase self-control, therefore, would be to increase the salience of average or long-term benefits relative to immediate benefits. An effective way to do this would be to group decisions in patterns or chunks rather than to make them on a case-by-case basis (Rachlin, 1995, 2000).

When Player B plays tit-for-tat over repeated games, Player A is in the position of a learner: Player A's cooperation is reinforced and defection punished. A more complex strategy, called ''Pavlov,'' which has been found to dominate other strategies, including tit-for-tat (that is, to earn the most points) in computer simulations (Nowak & Sigmund, 1993), puts Player A in the role of a teacher rather than a learner. Using Pavlov, Player B cooperates in two cases: first, if Player B had previously cooperated and that cooperation was reinforced by Player A's cooperation; second, if Player B had previously defected and that defection was punished by Player A's defection. Player B defects in two cases: first, if Player B had previously cooperated and that cooperation was punished by Player A's defection; second, if Player B had previosuly defected and that defection was reinforced by Player A's cooperation. Player B's Pavlov strategy thus puts Player A in the position of a teacher, reinforcing or punishing Player B's choices. (Because the putative learning process is instrumental, the strategy should really have been called, ''Thorndike'' or ''Skinner'' rather than ''Pavlov.'') When the prisoner's dilemma is played in real life, the roles of

teacher and learner are naturally intermixed. The final study to be reported in this article compares the effects of these two strategies on the behavior of subjects playing against them.

II. Playing Card Paradigm (Single Trials)

In the first phase of this experiment (Silverstein, Cross, Brown, & Rachlin, 1998), subjects (80 female undergraduates at the University of Rhode Island) played a repeated prisoner's dilemma game against the experimenter; in the second phase, the subjects played against each other. The experimenter, in the first phase, used one of four fixed strategies: always cooperate (10 subjects); always defect (10 subjects); play randomly (10 subjects); tit-for-tat (50 subjects). The experimenter's strategy was explained to each subject beforehand and the subject was required to describe it. The number of trials varied from 20 to 40; within this range, play was stopped when a subject cooperated or defected 10 times in a row.

As previously indicated, the prisoner's-dilemma subject has two motives: to defect and to get the other player to cooperate, the latter motive dominating the former. The reward for cooperating is greater than the reward for defecting but probabilistic (since it depends on the other player's reciprocation) and delayed (until the next trial). However, when the experimenter played all-cooperate, all-defect, or randomly, there was nothing the subject could do to influence the other player's behavior. That is, while defection by the subjects in these groups was reinforced, cooperation was not reinforced (an extinction procedure). Subjects in these groups were therefore expected to defect. On the other hand, when the experimenter played tit-for-tat, the reward for cooperation, although delayed until the next trial, was both certain and greater than the reward for defection. Subjects in the first-phase tit-for-tat group were therefore expected to cooperate.

Although the three first-phase experimenter-strategy groups: all-cooperate, all defect, and random, were alike in that cooperation was extinguished, they were unlike in the degree to which extinction could be discriminated from reinforcement. In experiments explicitly directed to determination of extinction effectiveness (Herrnstein, 1966), extinction of positive reinforcement (defined as a zero correlation between response and reinforcer) is most effective when the rate of (noncontingent) reinforcer presentation is zero, less effective when the rate of noncontingent reinforcer presentation is very high, and still less effective when noncontingent reinforcement is randomly related to responding. In general, extinction is most effective when it is most discriminable from conditioning. Because there was no conditioning prior to the first phase of the present experiment,

performance of the three extinction groups may be indicative of subjects' reinforcement history.

In the second phase of the experiment, subjects played each other. All subjects played against a subject who had previously played versus tit-for-tat. That is, 10 of the 50 first-phase tit-for-tat subjects played against a first-phase all-cooperate subject; 10 played against a first-phase all-defect subject; 10 played against a first-phase random subject; 20 played against another first-phase tit-for-tat subject. The game ended after 40 trials, but the subjects were not told how long the game would be. Because, at the end of the first phase, subjects playing against tit-for-tat were expected to be cooperating, whereas those in other groups were expected to be defecting, the second phase would test (a) whether subjects would carry over their behavior from play against the experimenter to play against another subject, (b) whether cooperating subjects would continue to cooperate when playing against each other, and (c) whether cooperating subjects would cause defecting opponents to cooperate or vice versa.

The prisoner's dilemma game was played with playing cards. Only the color of the cards, red (standing for cooperate) or black (standing for defect), had any significance. Throughout the game each subject had before her the payoff matrix of Figure 1a, except the alternatives were labeled, ''You play RED,'' ''You play BLACK,'' ''Experimenter plays RED,'' ''Experimenter plays BLACK,'' rather than ''C'' or ''D.'' The rewards were white poker chips that the subject were asked to imagine were dollars. Each subject was given two half-decks of cards, one all red, one all black. A trial began with both the experimenter and subject choosing a card and placing it face down on a table between them. Then both cards were turned over simultaneously. Recall that the experimenter fully described her strategy prior to the game. In addition, for the random group, the experimenter shuffled her cards in front of the subject and played them blindly from the top of the deck. For the tit-for-tat group, the experimenter played red on the first trial and on subsequent trials mimicked the subject's prior-trial choice. After each trial subjects were given a number of poker chips corresponding to the reward. Chips were stored in a clear plastic container visible to the subject at all times. In the second phase two subjects, each with identical reward matrices, played each other, with the experimenter signaling when to place cards on the table and when to turn them over. Each subject could see the other's accumulated chips as well as the chips earned on each trial. One important characteristic of this game was that both players chose simultaneously. Neither knew the others' choice until the trial was over. This simultaneous choice is characteristic of prisoner's dilemma games as typically studied in the laboratory. (Experiments to be described later will study successive play.)

Figure 2 shows the average percent cooperation for the first eight trials and the last eight trials for each group in the first phase. (The last eight trials for the all-defect group is at zero with no variance.) As expected, the group playing against tit-for-tat cooperated the most and increased cooperation over the course of the experiment, whereas all of the other groups as a whole cooperated less initially than the tit-for-tat group and decreased cooperation over the course of the experiment. Despite the fact that when playing randomly the experimenter shuffled her cards and played them blindly (her play was clearly random and not influenced by the subject's play), the random group cooperated more than the all-cooperate and all-defect groups. The results were in the order expected on the basis of typical extinction procedures. It is somewhat surprising that subjects in the random group cooperated to the extent that they did and showed no significant tendency to defect more as the experiment progressed. Subsequent experiments with the random procedure (not shown here) using small amounts of money rather than poker chips produced corresponding results. Clearly most of the variance between the groups was present at the very beginning of the experiment.

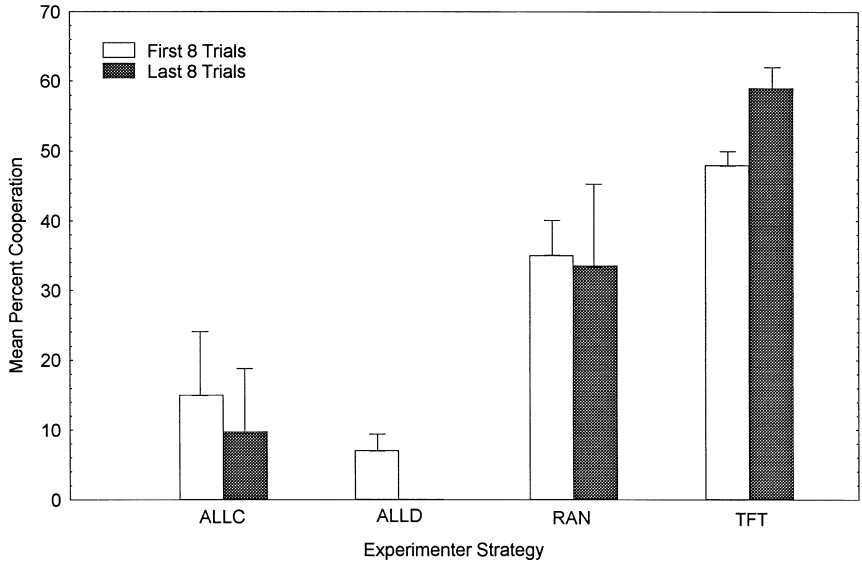


Fig. 2. Percent cooperation and standard error for each group in the first phase (versus the experimenter) of Experiment 1. Average of first and last eight trials. ALLC, all cooperate; ALLD, all defect; RAN, random; TFT, tit-for-tat.

Figure 3 shows the results of the second phase in which subjects played each other rather than the experimenter. The open squares and circles show average percent cooperation of each tit-for-tat and other group for the last four trials of the first phase. The closed squares and circles show average percent cooperation for successive four-trial blocks of the second phase. To some extent, subjects did carry over behavior from the first to the second phases. Subjects who had played against tit-for-tat cooperated more on average in the first four trials of the second phase than subjects who had played against the other strategies. However, regardless of their cooperation or defection at the end of the first phase, all groups cooperated at the beginning of the second phase within a range of 25 to 50%. Even the all-defection group, which had been defecting on every trial at the end of the first phase, cooperated within this range at the beginning of the second phase.

By the second block of four trials in the second phase, the slight difference between tit-for-tat and other groups reversed for two of the groups. By the third block the difference reversed for all groups. Thereafter, there was a steady deterioration of cooperation for all groups including tit-for-tat sub-

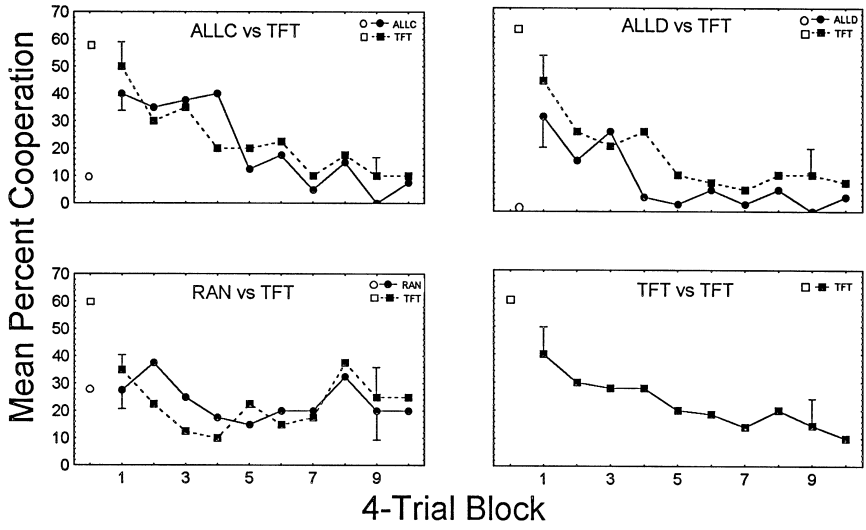


Fig. 3. Average percent cooperation for each group of Experiment 1 of the last four trials of the first phase (subject vs. experimenter) and all of the second phase (subjects versus each other) plotted in blocks of four. Standard errors are shown for blocks 1 and 9. The dashed lines show performance of the tit-for-tat (TFT) subjects who played against each other group. The solid lines show performance of that group. ALLC, all cooperate; ALLD, all defect; RAN, random.

jects playing against each other. An exception to the pattern of deterioration of cooperation is the apparent rise in cooperation at the end of the second phase of both the random group and the tit-for-tat subjects playing against them. However, a replication of the first and second phases with random versus tit-for-tat groups and tit-for-tat versus tit-for-tat groups, run in the second phase for 100 trials, failed to find a rise in cooperation at the end of the second phase. The clearest finding of the second phase was that the cooperation of subjects versus an experimenter who played tit-for-tat was highly specific to the situation. Almost as soon as the situation was changed and subjects began to play each other, those who had been cooperating at a high rate began to defect, whereas those who had been defecting at a high rate began to cooperate. Thereafter, subjects began to defect together and, by the end of the second phase, were generally defecting.

In summary, there was a small degree of transfer from one phase to the next, but subjects' behavior at the beginning of the second phase was largely determined by the change in opponent from experimenter to other subject. The crucial factor underlying this change is undetermined at this point, but later experiments will show that it is an expected probability of reciprocation of both cooperation and defection. For tit-for-tat subjects, expected probability of reciprocation declined when the opponent changed from experimenter to other subject; cooperation therefore fell at that point. For the all-cooperate and all-defect groups, expected probability of reciprocation rose when the opponent changed from experimenter to other subject; cooperation therefore rose at that point. Initial expectation of the other subjects' reciprocation seems to have been higher than it actually turned out to be because, as subjects continued in the second phase, cooperation mutually declined.

III. Playing Card Paradigm (Grouped Trials)

Recall that playing a prisoner's dilemma game against another player who uses a tit-for-tat strategy is like a self-control situation in that the alternatives are a smaller reward on the present trial versus a larger reward on the next trial. A tit-for-tat strategy essentially sets probability of reciprocation equal to 1.0. But playing a prisoner's dilemma game against another player whose probability of reciprocation is less than 1.0 is also like a self-control situation, except the next-trial reward for cooperation and the next-trial punishment for defection are probabilistic as well as delayed. Because probabilistic and delayed rewards are similarly discounted (Rachlin, Logue, Gibbon, & Frankel, 1986), prisoner's dilemma games, from the individual player's viewpoint, may be said to be very much like self-control situations. That

is, self-control and social cooperation are not just vaguely analogous but closely related.

In real-life self-control situations, as in real-life social cooperation situations, an immediate and certain, but small, reward is opposed to a distant and probabilistic, but large, reward. An alcoholic, for example, often chooses between an immediate, certain drink and a distant and very uncertain sobriety (entailing better health, social and job success, etc.). The crucial distinction between self-control and social cooperation situations is that, given present cooperation, in a self-control situation the probability of obtaining the larger more distant reward depends crucially on the person's own future reciprocation (the alcoholic will not attain sobriety unless he refrains from drinking in the future as well as in the present), whereas in a social cooperation situation the probability of obtaining the larger more distant reward depends on the reciprocation of others.

In prior laboratory studies of self-control where the larger reward was probabilistic (Rachlin & Siegel, 1994) or both distant and dependent on the subject's own future reciprocation (Kudadjie-Gyamfi & Rachlin, 1996), temporal grouping or patterning of trials tended to increase self-control (self-cooperation). This makes sense if it is noted that the patterning of trials—three or four quick trials and then a long intertrial interval—tends to encourage subjects to decide their play in advance for the trials within the group. This is a kind of commitment (Siegel & Rachlin, 1995, call it ''soft commitment'') and commitment tends to increase self-control. The experiment to be described in this section (Silverstein et al., 1998, Experiment 2) examines the effect of patterning on cooperation in the prisoner's dilemma.

The procedure with patterned trials was the same as in the single-trial experiment except for the following:

1. The experimenter played only randomly or tit-for-tat.
2. Each subject played exactly 40 trials with the experimenter.
3. Subjects in each group were divided into a single- or patterned-trial condition. The former repeated the condition in the single-trial experiment. The latter required subjects to select a sequence of four cards and place them face down in a row in front of them. In the first phase, when the experimenter played randomly, she blindly placed four cards from the shuffled deck opposite to the subject's cards. When the experimenter played tit-for-tat, she duplicated the subject's prior four-trial pattern. Then pairs of experimenter–subject cards were turned over one at a time with the subject receiving the designated reward after each pair was turned over.
4. In the second phase, subjects played against each other under the same condition (single or patterned trials) as in the first phase. Under each

condition, 10 pairs of first-phase tit-for-tat subjects played against each other and 10 pairs of first-phase random subjects played against each other. In addition, under each condition, 10 first-phase random subjects played against 10 first-phase tit-for-tat subjects.

5. Subjects played against each other in the second phase for 50 trials.
6. Subjects traded their chips for money at the rate of 2 cents per chip.

During the first phase, cooperation by the groups playing with patterned trials differed neither from the single-trial subjects in this experiment nor from those of the previous experiment. Tit-for-tat subjects cooperated more than random subjects both at the beginning and at the end of this phase. Tit-for-tat subjects tended to increase cooperation over trials, whereas the random subjects tended to decrease cooperation over trials. That is, patterning of trials had no measured effect on subjects' behavior when they were playing against the experimenter under either single-trial or patterned-trial conditions.

In the second phase, however, when subjects played against each other, patterned trials increased cooperation for all groups. Figure 4 shows the second-phase results. As in Experiment 1, in the single-trial condition, tit-for-tat subjects decreased cooperation from the end of the first phase to the beginning of the second phase. In this experiment, for both conditions,

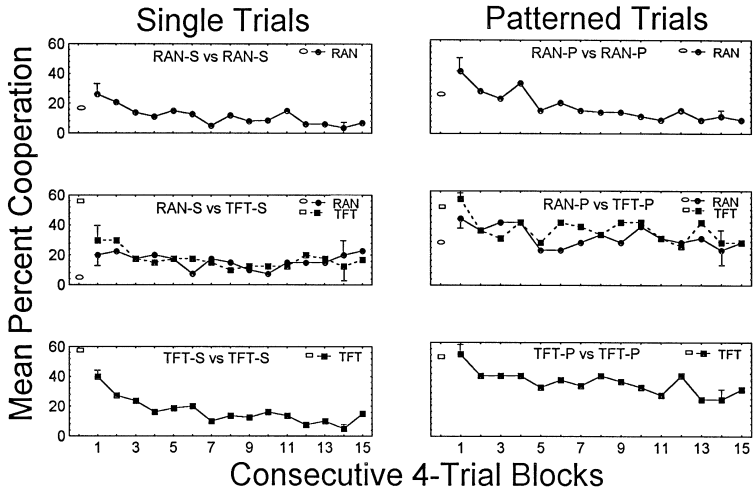


Fig. 4. Average percent cooperation for each group of Experiment 2 for the last four trials of the first phase in which subjects played versus the experimenter (leftmost open points) and all of the second phase (plotted in blocks of four). Standard errors are shown for blocks 1 and 14. RAN, random; TFT, tit-for-tat.

random subjects increased cooperation from the end of the first phase to the beginning of the second. However all patterned-trial groups, whether random or tit-for-tat, increased cooperation from the end of the first phase to the beginning of the second phase. And, as before, there was a steady decline in cooperation as the second phase progressed for all pairs, except for the random/tit-for-tat pairs with single trials (but recall that when this condition was repeated with 100 second-phase trials cooperation did decline).

The clearest and most reliable finding was the attenuation of deterioration of cooperation throughout the second phase by the patterning of trials. Thus, at least when playing each other, patterning increased cooperation. Why did patterning have no effect in the first phase? For the random group, patterning should have had no effect or should have increased defection in the first phase—because cooperation was then unreinforced. For the tit-for-tat group, patterning may have obscured, or at least not clarified, the fixed 100% reciprocation contingency characteristic of tit-for-tat. In the single-trial condition, each of two alternatives was consistently reinforced or consistently punished on the next trial. In the patterned-trials condition there were 24 possible patterns and four levels of reinforcement or punishment. Even though the tit-for-tat contingency was explained to the subjects, the response–reinforcer relationship may still have been obscure. In the second phase, however, when playing against another subject with a less-than-unity probability of reciprocation, the grouping of trials would reduce the variance of the probability (in the same way that the variance between groups of individuals is less than the variance between all of the individuals) and made subjects more confident that their opponents would reciprocate their cooperation.

Why did cooperation between individuals deteriorate in the second phase of both experiments? Remember, what generates cooperation by a player is not the cooperation of the other player (if it were, subjects in the all-cooperate group of Experiment I would have learned to cooperate), it is the other player's tendency to reciprocate—that is, to play tit-for-tat or some probabilistic approximation to it. Although subjects who played against tit-for-tat learned to cooperate, a trial-by-trial analysis of their first-phase play revealed no tendency to play tit-for-tat. Postexperimental interviews with subjects in Experiment 1 elicited a diverse group of strategies that subjects claim to have followed. A few subjects claimed to have followed a tit-for-tat strategy, but no such tendency was found in their actual play. (The only strategy for which a correspondence was found between verbal report and behavior was defect on all trials.)

Recall that the best possible state of prisoner's dilemma outcomes from an individual subject's viewpoint is to defect while the other player cooperates.

Many players expressed this as their goal. But, as previously noted, it is not an achievable goal in repeated play; it is this that may have led to mutual defection in the second phase of both experiments.

Another factor that may have led to mutual defection is the incompatibility of the tit-for-tat and Pavlov strategies. The choices of a player in a prisoner's dilemma game are simultaneously being reinforced or punished and reinforcing or punishing the choices of the other player. That is, a player is both a learner and a teacher. Although either role leads to mutual cooperation, a single player cannot consistently assume both roles (except at the point where both players are cooperating on all trials). This confusion of strategies may have led to the mutual defection that we observed. The final experiment of this chapter will explore further the effect of these strategies on behavior. Meanwhile, the next series of experiments attempted to obtain more consistent cooperation by clarifying the prisoner's dilemma contigencies and having subjects choose sequentially rather than simultaneously.

IV. Game Board Paradigm (Single Trials)

Despite explicit instructions and the availability of the reward matrix throughout the experiment, subjects in the playing card experiments, playing versus tit-for-tat, did not cooperate, on the average, on more than 60% of the trials. Although subjects in tit-for-tat groups could verbally express the experimenter's tit-for-tat contingency, and were sensitive to the larger reward obtainable by cooperating, in the sense that their cooperation increased over trials, most of them did not learn to maximize reward by cooperating on every trial when playing against tit-for-tat, and did not maintain mutual cooperation when playing against each other. Accordingly, a game board was constructed with the aim of clarifying the tit-for-tat contingency (Brown & Rachlin, 1999).

The game board is diagramed in Figure 5a. It consisted of a rectangular plastic tray divided into four compartments (''boxes''). Each box contained three items:

1. A red or green index card with a picture of a door (''red doors'' or ''green doors'')
2. A red or green key
3. A number of nickels (one, two, three, or four nickels in Experiment 3; one, two, five, or six nickels in subsequent experiments).

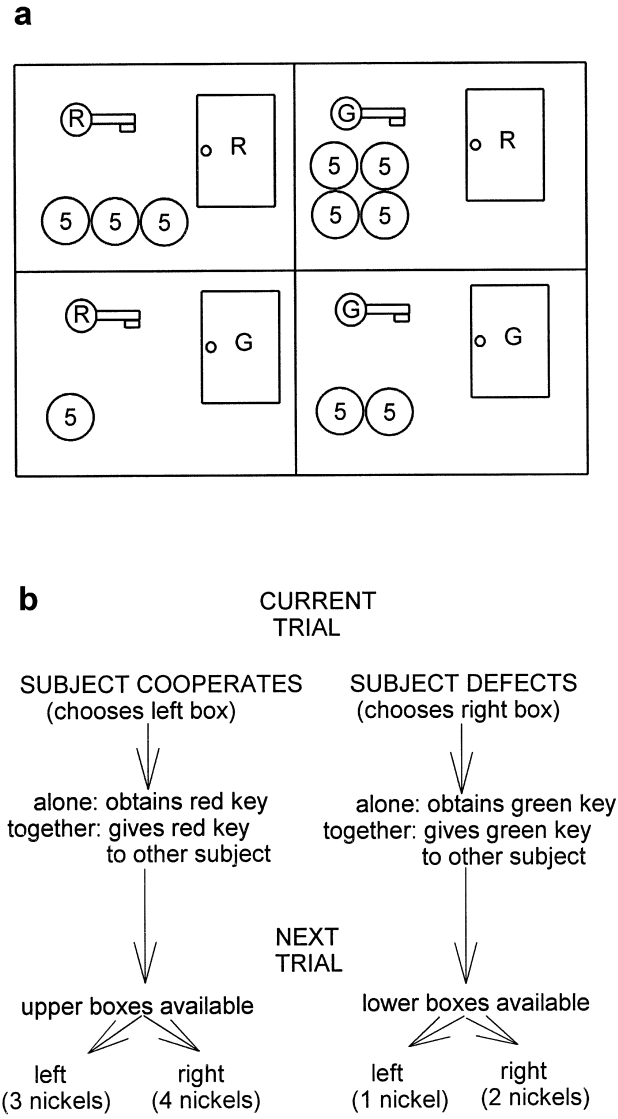


Fig. 5. (a) Diagram of apparatus used in Experiment 3. (b) Sequence of alone and together conditions of Experiment 3.

The upper boxes both contained red doors; the lower boxes both contained green doors. The left boxes both contained red keys; the right boxes both contained green keys. Note that each right box held one more nickel than the box to its left, and each upper box held two more nickels than the one below it (4 more in Experiment 2). All the items in the boxes were visible to the players.

The rules of the game were as follows: Keys were used to "open a door" of the apparatus. A red key opened one of the two red doors; a green key opened one of the two green doors. Opening a door allowed the player to take both the key and the nickels from that compartment. (Subjects kept the money they earned.) The key used to open the door was taken from the player, and new nickels and the appropriate colored key were replaced in the apparatus by the experimenter. That is, after each trial the configuration shown in Figure 5a was restored. Figure 5b shows the sequence of choices and their consequences.

The game could be played in two ways: by a single player ("alone" condition) or by a pair of players ("together" condition). Each trial in either condition began with the apparatus as pictured in Figure 5a. In the alone condition, the player was given a red key. The player could use that key to "open" one or the other red door (to choose either the upper left or upper right box). The used key was then surrendered. If the upper left box was chosen, the player was permitted to take the three nickels and the red key from that box. If the upper right box was chosen, the player was permitted to take the four nickels and the green key from that box. Then the nickel(s) and key taken were replaced by the experimenter and the next trial began. If a red key had been received on the previous trial, the player could again choose between the two red doors as before; if a green key had been received on the previous trial, the player could use the key to "open" one or the other green door (to choose a key and nickels from either the lower left or lower right box).

In the alone procedure the apparatus is essentially playing tit-for-tat with the subject. If the subject "cooperates" by choosing the lower of the two available rewards, the apparatus also "cooperates" by giving the subject a red key that allows the subject to choose between the two higher rewards on the next trial. If the subject "defects" by choosing the higher of the two available rewards, the apparatus "defects" by giving the subject a green key that forces the subject to choose between the two lower rewards on the next trial. Because this is a tit-for-tat procedure, it is also a self-control procedure, in the sense that the behavior leading to the higher current reward (choosing the right box) conflicts with the behavior that maximizes overall reward (choosing the left box—the one with the red key). Choosing the right box always earned the player one more nickel than did choosing

the left box, but at the cost of obtaining a green key. With the green key the player payed for the 1-point gain (for choosing the right box) on the previous trial with an average 2-point loss (having to choose between the lower boxes) on the present trial. The subject's best overall strategy, as in any prisoner's dilemma game against tit-for-tat, is to always choose the left box (''cooperate''), always receive a red key, and always earn three nickels.

The tit-for-tat contingency was broken in the together condition by having each subject hand the key she had obtained on any given trial to her opponent who then used the key to make a choice on the next trial. In other words, in the together condition, two players, playing on a single game board, as diagrammed in Figure 5a, made choices on alternate trials. This procedure differs from that of Experiments 1 and 2 in that the players chose sequentially with full knowledge of their opponents' choice and their own alternatives. The experimenter arbitrarily designated one player to choose first, and gave that player a red key. After the first player chose one of the upper boxes, the key from that box was handed to the second player, who chose between the upper boxes if the key was red (if the first player had chosen the left box) or between the lower boxes if the key was green (if the first player had chosen the right box). The key obtained by the second player was then handed to the first player, and the first player made the permitted choice, and so forth. Thus, after the first trial, whether a player chose between the upper boxes or between the lower boxes depended on the other player's choice on the previous trial.

In this game, income would be maximized (at three nickels per trial) for each player if both players repeatedly chose the left box (cooperated). However, the individual player would always gain more on the present trial by choosing the right box (defecting) than by choosing the left box. The penalty for defecting, of having to choose between the lower boxes, is suffered not by the player who defects (as it was in the first phase) but by the other player, who inherits the green key. As in the playing-card game, cooperating is the very worst strategy, unless the other player also cooperates. Therefore, the only reason to cooperate (within the demands of the game) is to influence the other player to cooperate subsequently.

Ten pairs of subjects (female Stony Brook undergraduates) played alone in the first 20-trial phase of the experiment and together in the second 20-trial phase; 10 pairs played together in the first phase and alone in the second phase; 10 pairs played alone in both phases; 10 pairs played together in both phases. The experimental design is a typical transfer design. Figure 6 summarizes the results. Pairs who played alone in both phases (tit-for-tat) began by cooperating on about 35% of the trials, gradually increased cooperation and ended by cooperating on about 60% of the trials, about the same rate as subjects in the tit-for-tat group of the playing-card para-

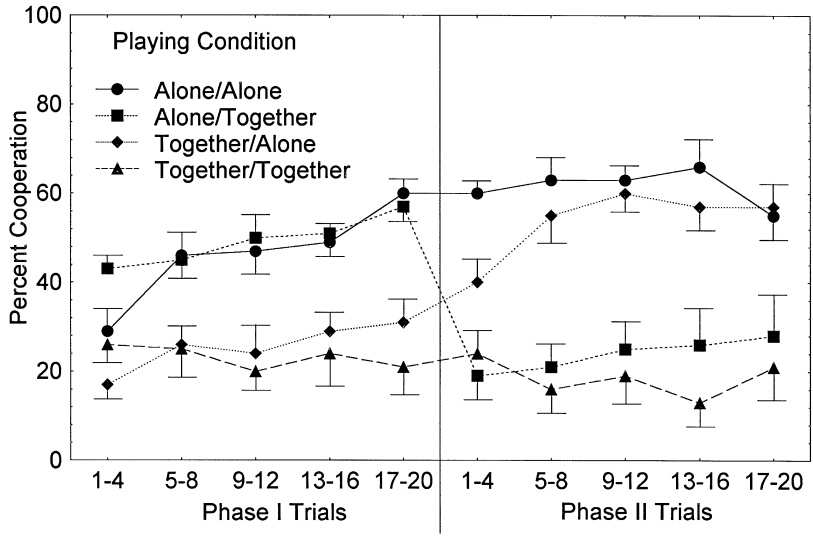


Fig. 6. Average percent cooperation of each group for both phases of Experiment 3 (four-trial blocks).

digm. Thus, the (attempted) greater clarity of the game board as compared with the playing card paradigm did not increase cooperation. Pairs who played together in both phases began cooperating on about 20% of the trials and did not increase cooperation over the course of the experiment. This is a slightly higher cooperation rate than the final rate of subjects playing each other with the playing card paradigm. Also, there was no initial elevation and gradual decline with the game board as there was with the playing cards. Perhaps this was an effect of the procedure's clarity or the successive rather than simultaneous trials. For the two groups that were switched, no transfer at all was observed. Cooperation by the group that played first alone and then together fell to the level of the group playing together in both phases as soon as the procedure was changed. The group that played first together and then alone learned to cooperate no faster in the second phase than the groups that played alone in the first phase.

In summary, even more than with the playing-card paradigm, the subjects' situation rather than their experience within the experimental setting determined their behavior. What exactly about the alone and together situations generated such different degrees of cooperation? At this point we suspected, and later experiments (to be described) confirmed, that the crucial factor was probability of reciprocation: probability of reinforcement of cooperation (by the opponent's cooperation) and probability of punishment

of defection (by the opponent's defection). This probability was 1.0 in the alone (tit-for-tat) game. Although the game board should have made this clear, subjects seem to have expected it to be lower and hence they had to learn to cooperate as the game progressed. This learning was never complete, reaching asymptote at about 60%. Expected probability of reciprocation was apparently much less than 1.0 in the together game, hence the subjects' expectations were self-fulfilling; if each subject expected the other to cooperate one-fifth of the time, then each might have cooperated in turn only one-fifth of the time, accounting for the stable 20% cooperation of all groups in the together condition.

A question might arise as to the extent to which motivational forces played any part in these experiments. After all, there *is* a right and a wrong way to behave in the tit-for-tat conditions of the previous experiments. These conditions may be said to present subjects with a purely cognitive problem that only some of them solved. Within-group distributions of percent cooperation (not shown here but to be presented in later experiments) of subjects playing alone, were far from normal. Some subjects did cooperate on all of the last several trials, some seemed to play randomly, whereas others alternated between cooperation and defection for short stretches of trials. Certainly, to some degree, there was a cognitive element to the tit-for-tat contingency of the alone condition. The next experiment was designed to see whether there was any motivational element to this condition at all.

V. Game Board Paradigm (Single Trials, Increased Reward for Cooperation)

Experiment 4 was a replication of Experiment 3 except, in this experiment, the number of nickels in the upper boxes was increased (from 3 and 4) to 5 and 6.[1] In experiments on self-control, defined as choice of a larger-later reward over a smaller-sooner reward, increasing the magnitude of the larger-later reward tends to increase self-control (Logue, 1988). If the tit-for-tat contingency is truly a self-control contingency, then increasing reward for cooperation should increase cooperation. If on the other hand

[1] Subjects in some of the experiments reported in this chapter earned hypothetical monetary rewards, some earned real but small amounts of money, and some simply earned points. Results did not vary in any discernable way with type of reward. Where reward magnitude was varied, small amounts of money were used as rewards. But, we believe, variation of points earned would have had the same effect. This is not to say that magnitude or type of reward makes no difference. Rather, the small amounts of money that could practically be used as rewards were no more effective as incentives in these experiments (or in others) than points or larger but hypothetical amounts of money.

the alone game is purely a cognitive problem, like a problem in practical arithmetic or a question on the SAT, increasing the numbers would not be expected to make the problem any easier to solve.

Figure 7 shows the results of increasing the reward for cooperation. The asymptotic level of cooperation of groups playing in the alone condition rose from 60% with the smaller rewards to about 80% with the larger rewards. With all else held constant, increasing reward for cooperation clearly increased cooperation of subjects playing alone.

It may be argued that the tit-for-tat contingency of the alone game is still essentially a cognitive problem, and that increasing the reward for cooperation merely increased the subjects' motivation to solve the problem. This may well be the case. But from a behavioral viewpoint (our viewpoint) the internal mechanism by which reward acts is irrelevant. (It may be that all self-control problems are basically cognitive in this very sense.) The fact remains that cooperation in the alone condition was sensitive to the parameters of its reinforcement.

On the other hand, increased reward had no effect on cooperation of subjects playing together. This was expected since the closest beneficiary of cooperation was not the player who cooperated but the other player. Subjects' expected level of reciprocation in the together condition would not have been affected by the increased reward. The next experiment was

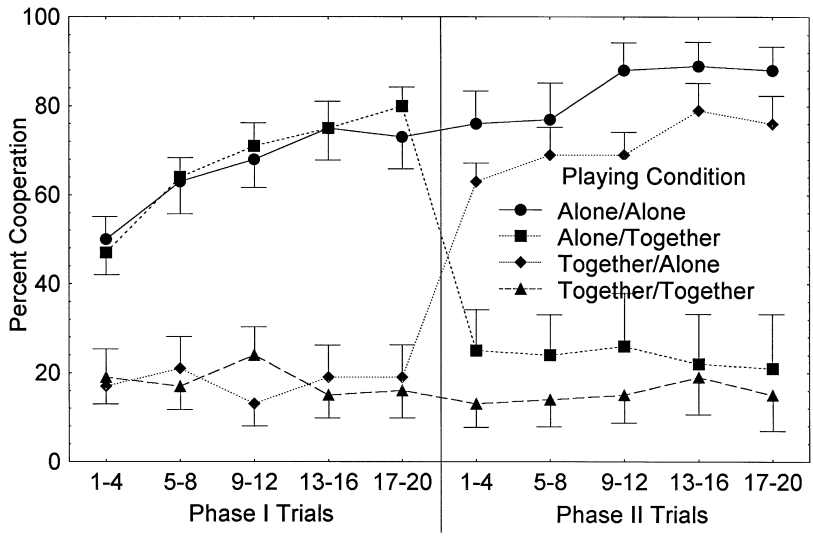


Fig. 7. Average percent cooperation of each group for both phases of Experiment 4, with increased reward for cooperating (four-trial blocks).

an attempt to increase the cooperation of subjects playing together by a method corresponding to that of Experiment 2, patterning of trials.

VI. Game Board Paradigm (Patterned Trials)

In Experiment 5 all subjects played in the together condition in both phases. The rewards were one, two, five, and six nickels as in Experiment 4. The subjects played under two conditions: one trial at a time as in Experiment 4, and four trials at a time as in Experiment 2. Each subject of the pair playing together with patterned trials was asked to make four choices at a time by writing, on a slip of paper, a sequence of four "L" and "R" choices, standing for left-box and right-box choices (the left box always containing one less nickel than the right). After both subjects in a pair recorded their choices, the two slips of paper were exposed and the four trials (per subject) were played out in sequence. Four groups of subjects (10 in each group) were studied in a transfer design. In the first phase, two groups were studied with single trials and two groups with patterned trials; in the second phase two groups switched conditions while two groups continued in the same condition as the first phase.

Figure 8 shows the results for the last four trials of each phase. Patterning of trials, as before, increased the average percentage of trials on which

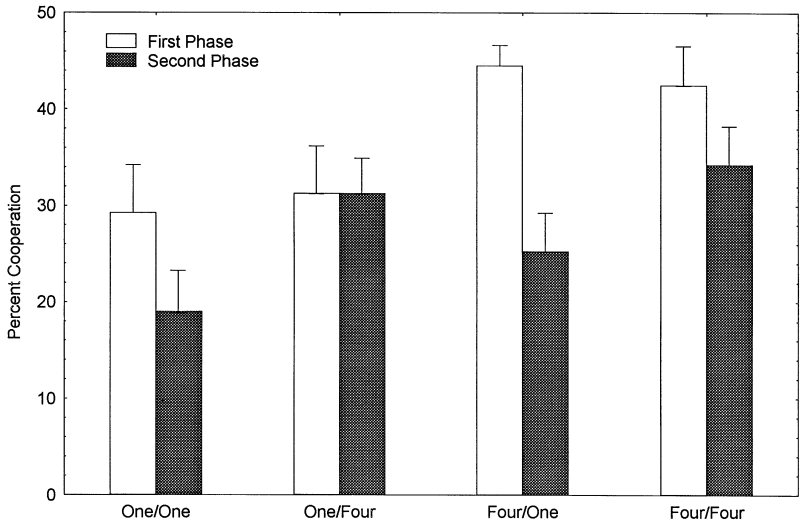


Fig. 8. Percent cooperation and standard error for each group in both phases of Experiment 5.

subjects playing together cooperated. However, the increase was small—from about 20% to about 35%.

As previously indicated, if probability of reciprocation was the crucial determinant of cooperation, patterning of trials of subjects playing together may have had a consolidating effect on the estimation of this probability, reducing risk, just as deciding to buy a group of risky stocks (say, in a mutual fund) reduces risk. Rachlin et al. (1986) argued that risk avoidance in probabilistic choice (a pervasive finding) was essentially the same phenomenon as impulsiveness in standard self-control situations. Rachlin, Castrogiovanni, and Cross (1987) showed that commitment works the same way in probabilistic choice situations as in standard self-control situations. The patterned-trial conditon is essentially a commitment condition; subjects had to commit themselves to a pattern of choices in four-trial chunks. Therefore (assuming that expected probability of reciprocation is the crucial variable determining social cooperation in the prisoners dilemma game), just as commitment increases self-control, patterning of trials increases social cooperation. As a test of this assumption, the next series of experiments examines cooperation and defection as a function of probability of reciprocation where probability of reciprocation is directly varied.

VII. Computer Paradigm (Probability of Reciprocation, Spinners)

The object of Experiment 6 (Baker, 1999) was to directly test the theory that the crucial variable differentiating the high levels of cooperation in the self-control (tit-for-tat) conditions from the low levels of cooperation in the social cooperation (subjects playing together) conditions is the probability of reciprocation. According to the theory, subjects playing alone in Experiments 3 and 4 cooperated more than subjects playing together because the alone player's expectation of the probability that if she cooperated on this trial *she herself* would cooperate on the next trial was higher than the together player's estimation that if she cooperated on this trial *her partner* would cooperate on the next trial. In Experiment 6, probability of reciprocation was explicitly varied. The game was played not on a game board with another subject but on a computer screen with the computer taking the place of the other subject. Figure 9 depicts the screen seen by the subjects on the first trial. On the upper right was a diagram with four boxes like the game board of Figure 5a but without keys and doors and with 5 and 6 points, rather than three and four nickels, in the upper boxes. The bar at the lower right of the screen was filled-in proportional to the money earned. Corresponding to the alone condition of Experiment 4,

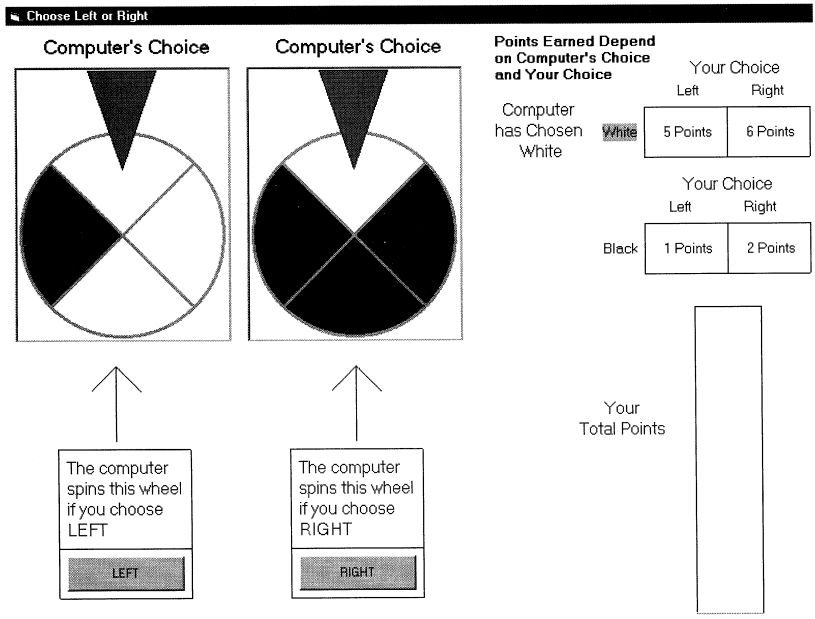


Fig. 9. Computer screen seen by subjects in Experiment 6.

subjects chose between the left box with 1 or 5 points and the right box with 2 or 6 points. Whether subjects could choose between the upper two boxes (5 versus 6 points) or the lower two boxes (1 versus 2 points) depended on the computer. Between subject choices the computer would decide to highlight either the word "white," designating the two upper boxes as available for choice on the next trial, or the word "black," designating the two lower boxes as available for choice on the next trial. The subject could choose only between the designated boxes. The subject's decision determined the computer's decision with a probability, PR. This probability, the probability of reciprocation by the computer, was the crucial experimental variable. Figure 10 illustrates the contingencies.

On the screen, in addition to the four boxes and the bar indicating winnings, were two spinners with white and black sectors. The proportion of the left spinner that was white was equal to *PR,* and the proportion of the right spinner that was white was equal to 1 − PR. That is, the right spinner was the inverse of the left. On a given trial, if the subject chose the left box with the lower number of points (if the subject cooperated) the left spinner was spun. If the spinner stopped with the white sector under the arrow, the upper boxes were designated as available on the next

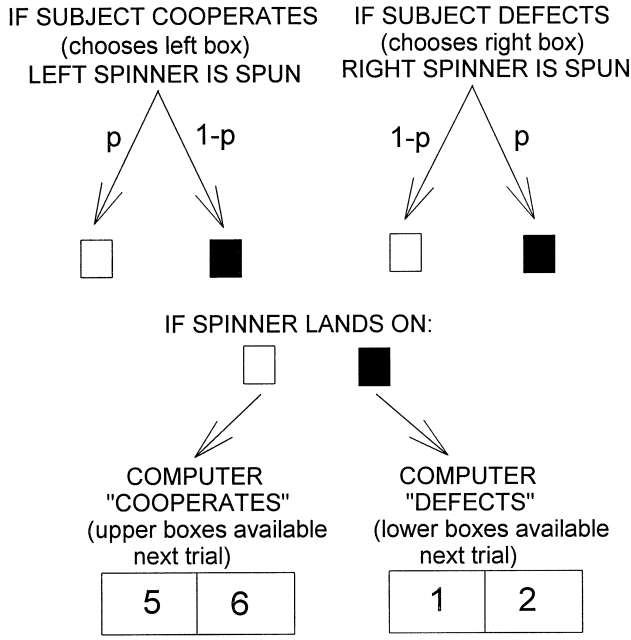


Fig. 10. Sequence of events in Experiment 6. The reciprocation probability (PR) is p.

trial (that is, the computer cooperated); if it stopped with the black sector under the arrow, the lower boxes were designated as available on the next trial (that is, the computer defected). On the other hand, if the subject chose the right box with the higher number of points (if the subject defected), the right spinner was spun. Again, if the spinner landed on white, the upper boxes were designated as available for choice on the next trial (the computer cooperated), whereas if the spinner landed on black, the lower boxes were designated as available for choice on the next trial (the computer defected).

The probability was the probability of reciprocation (PR). If the subject cooperated the computer would cooperate with a probability equal to PR and defect with a probability equal to 1 − PR; if the subject defected the computer defected with a probability equal to PR and cooperated with a probability equal to 1 − PR. Thus, the probability of reciprocation, implicit in prior experiments, was explicit and signaled by a discriminative stimulus in Experiment 6. Three groups of subjects (8 males, 8 females in each group) were each tested for 100 trials at a different probability of reciprocation (PR): 50%, 75%, and 100%. With PR = 100%, the computer was playing tit-for-tat. If the subject cooperated on a given trial, the computer would surely cooperate; if the subject defected the computer would surely defect.

The PR = 100% condition was equivalent to the alone condition in Experiment 4; subjects would maximize earnings by always cooperating. With PR = 50%, the computer essentially flipped a coin to decide whether to cooperate or defect. Subjects would maximize earnings by always defecting (choosing the higher current-trial number of points) since cooperation with a computer would earn them nothing - not even gratitude. The probability at which the expected value of cooperating equaled the expected value of defecting was 63%. For reciprocation probabilities below this value, the subject would earn more points in the long run by defecting on every trial than by cooperating on every trial; for reciprocation probabilities greater than this value, the reverse would obtain.

Figure 11 shows average cooperation for the three groups as the experiment progressed. The PR = 100% and PR = 75% groups both learned to cooperate. The rate of cooperation of the PR = 100% group was higher than that of tit-for-tat subjects in the playing card and game board paradigms. The explicit signal of the 100% reciprocation probability provided by the spinners (one all white, the other all black for this group) evidently was a stronger discriminative stimulus for cooperation than the verbal description or graphical presentation of the tit-for-tat contingency in the other two paradigms. The PR = 75% group increased cooperation more slowly and to a lower asymptote than did the PR = 100% group. The expected value of cooperating (5 points) was higher for the PR = 100% group than for the PR = 75% group (4 points). Thus, comparing these two groups is like comparing high reward (Experiment 4) and low reward

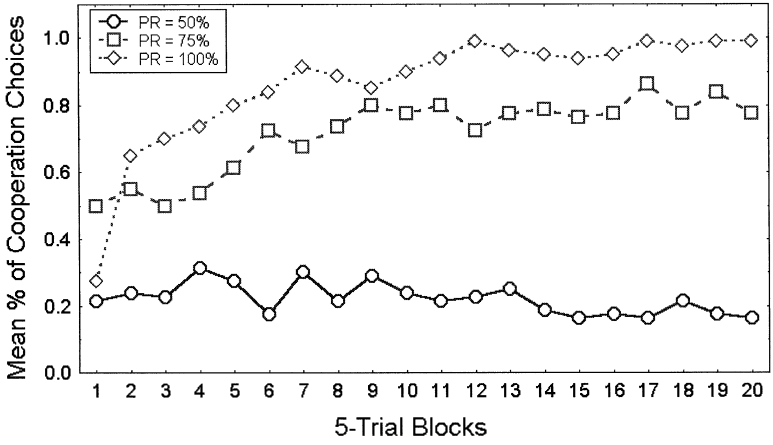


Fig. 11. Average percent cooperation for each reciprocation probability (PR) group over the course of Experiment 6 (five-trial blocks).

(Experiment 3) of the game board paradigm. Again, increasing the reward for cooperating increased cooperation. The PR = 50% group was not rewarded for cooperating and showed no increase from the low level of cooperation (around 20%) that occurred at the beginning of the experiment. This group is comparable to the random groups in the playing card paradigm. Again, the spinners (each identically half white and half black for this group) proved to be a better discriminative stimulus for randomness than was verbal description and shuffling of cards.

Figure 12 shows distributions of the 16 subjects of each group over the last 20 trials (there were no significant differences between males and females). Note the extreme skewness of the data. The modal subject in both the PR = 100% and PR = 75% groups cooperated on all 20 trials and thus maximized reward. The latter group, however, was far more variable than the former. The PR = 50% group was more variable still; however, 5 of the 16 subjects in this group did defect on all 20 of the trials, thus maximizing reinforcement.

In general, cooperation was strongly affected by the PR. This probability, explicitly varied in this experiment, must have implicitly varied in the together conditions of the game board experiments, and been largely instrumental in determining the difference between the alone and together conditions. Even though the actual reciprocation probability for the alone subjects in the game-board experiments was 100%, the lack of a clear discriminative stimulus and the complexity of the key and door contingen-

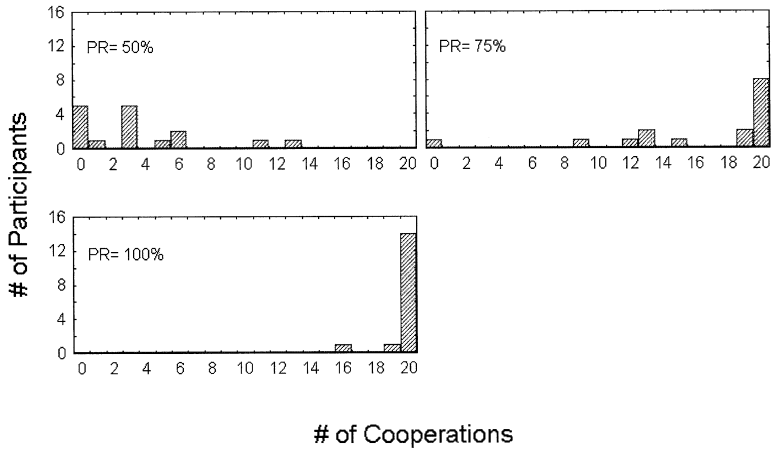


Fig. 12. Distribution of subjects in each reciprocation probability (PR) group of Experiment 6 with respect to number of cooperations in the last 20 trials.

cies may have reduced the subjective reciprocation probability for these subjects below 100%.

VIII. Computer Paradigm (Probability of Reciprocation, No Spinners)

Experiment 6 (Baker, 1999) was replicated with new subjects, the same number and composition of groups, the same procedure, and the same set of prisoner's dilemma parameters. The computer display was also the same except that in Experiment 7 the spinners were not visible to the subjects; they had to learn the probabilities and expected values from their own behavior and its consequences. Figures 13 and 14 show percent cooperation over the course of the experiment and distributions of individual subjects' cooperation over the last 20 trials.

For the PR = 100% and PR = 50% groups, the data were similar to those with spinners, except fewer subjects in the PR = 100% group cooperated on all of the last 20 trials. The good performance of the PR = 50% group (that is, their more frequent defection) relative to the random groups in the playing card experiments was surprising. Something about the playing card procedure aside from the clarity of the discriminative stimulus must have increased the subjects' tendency to cooperate; perhaps the playing of cards encouraged probability matching. The distribution of the PR = 75% group without spinners was the mirror image of the PR = 75% group with

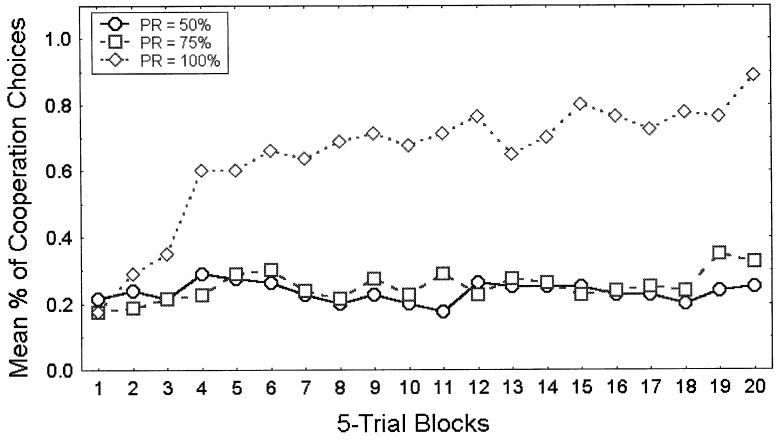


Fig. 13. Average percent cooperation for each reciprocation probability (PR) group over the course of Experiment 7 (five-trial blocks).

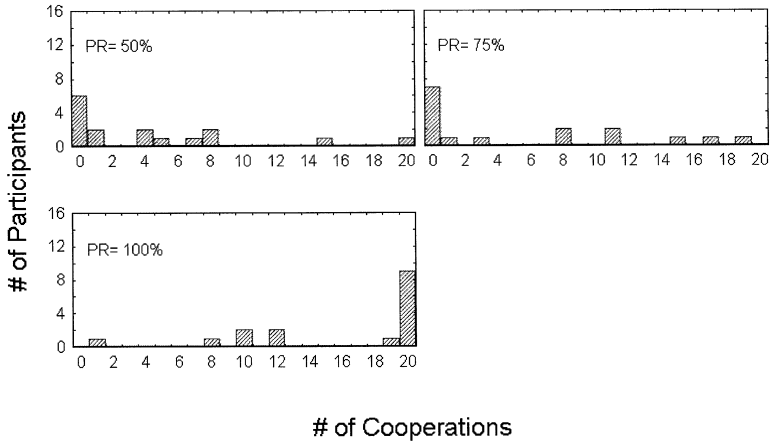


Fig. 14. Distribution of subjects in each reciprocation probability (PR) group of Experiment 7 with respect to number of cooperations in the last 20 trials.

spinners. The modal subject in the PR = 75% group without spinners defected on all of the last 20 trials. Recall that PR = 75% is only marginally greater than the PR = 63% at which the expected value of cooperating equals that of defecting. A difference of 12% may not have been discriminable without the spinners.

IX. Computer Paradigm (Probability of Reciprocation, Other Player)

Experiment 8 (Baker, 1999) replicated experiments 6 and 7 with new subjects, the same number and composition of groups, the same procedure, and the same set of prisoner's dilemma parameters. The computer display was like that of Experiment 7; the spinners were not visible to the subjects and they had to learn the probabilities and expected values from their own behavior and its consequences. However, in Experiment 8, subjects were led to believe that they were playing against another player in another room. Accordingly, the experimenter went through a mild charade at the beginning of the experiment, speaking on the phone to a supposed other experimenter, asking if the other subject were ready, and so forth. Also, an additional reward matrix on the screen displayed the ''other player's'' choices and ''earnings'' on each trial. A frame labeled ''your turn'' or ''other player's turn'' moved back and forth between the two matrices as appropriate. After the experimental session, each subject filled out a

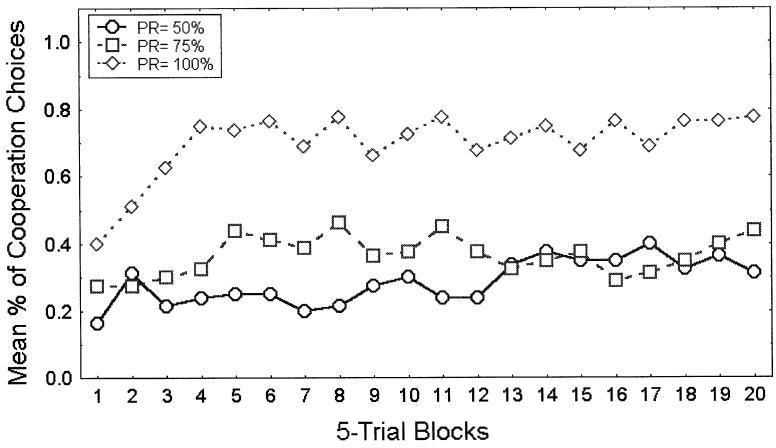


Fig. 15. Average percent cooperation (C) for each reciprocation probability (PR) group over the course of Experiment 8 (five-trial blocks).

questionnaire asking what he or she thought of the other subject. Two subjects, both in the PR = 100% group, expressed doubts that there was another subject. The data from those two subjects were discarded and two additional subjects were tested in that condition.

Figures 15 and 16 show the results of Experiment 8 in the same form as those of Experiments 6 and 7. Average percent cooperation of each group

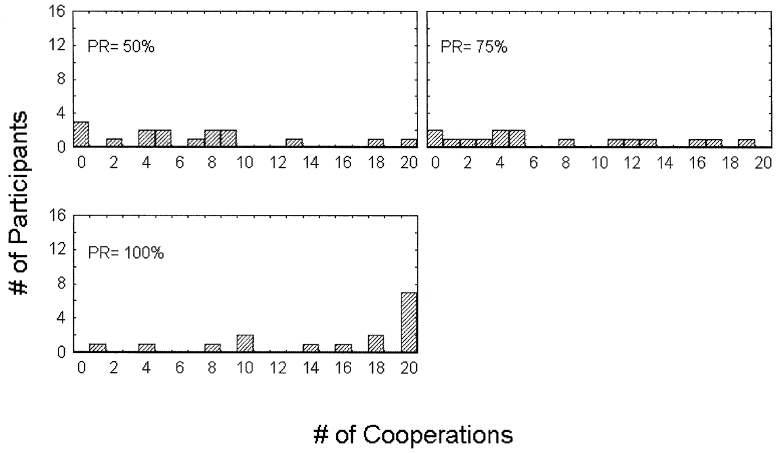


Fig. 16. Distribution of subjects in each reciprocation probability (PR) group of Experiment 8 with respect to number of cooperations in the last 20 trials.

in Experiment 8 was indistinguishable from that of Experiment 7. (That is, Figures 13 and 15 are similar.) There was, however, some noticeable difference in the distributions. In each group, fewer subjects fall at the extremes (all cooperate or all defect) over the last 20 trials. That is, the distributions of Figure 16 are less skewed than those of Figure 14. This is especially true of the PR = 50% and PR = 75% conditions, where more than 40% of the subjects playing against the computer defected on all of the last 20 trials, whereas less than 16% of the subjects did so when they believed they were playing against another subject. Perhaps this was due to a tendency by subjects to distribute reward equitably; consistent defection would penalize an opponent on occasions when the opponent cooperated. On the other hand, equity would be irrelevant in the case of a computer opponent.

X. Computer Paradigm (Probability of Reinforcement and Punishment Varied Separately)

In Experiment 9, the three procedures, spinners, no spinners, and "two players," were repeated, with probability of reciprocation of cooperation and probability of reciprocation of defection varied independently. In prior computer paradigm experiments, reciprocation probability was always equal for defection and cooperation. That is, the two spinners (visible or not) were inverses of each other. In the present experiment the white and black sections of the two spinners could vary independently. For each of the three procedures, there were 9 groups of 6 subjects each, a total of 27 groups and 162 subjects. Figure 17 shows average percent cooperation over the last 20 (of 100) trials for each group. One other difference between this experiment and the previous ones with this paradigm was that in this experiment subjects were paid at a rate of 1 cent per point earned over the 100 trials.

The results are shown in two ways: with the probability of reciprocation of defection as the independent variable and probability of reciprocation of cooperation as the parameter (left column of graphs), and with the probability of reciprocation of cooperation as the independent variable and probability of reciprocation of defection as the parameter (right column of graphs).

Experiments in other laboratories (Bixenstein & Gaebelein, 1971; Komorita, Hilty, & Parks, 1991) found reciprocation of cooperation to have a stronger effect on behavior than did reciprocation of defection. In our experiment there was no significant difference in these effects. Both were strong, as shown by the positive regression lines (the dotted lines). The

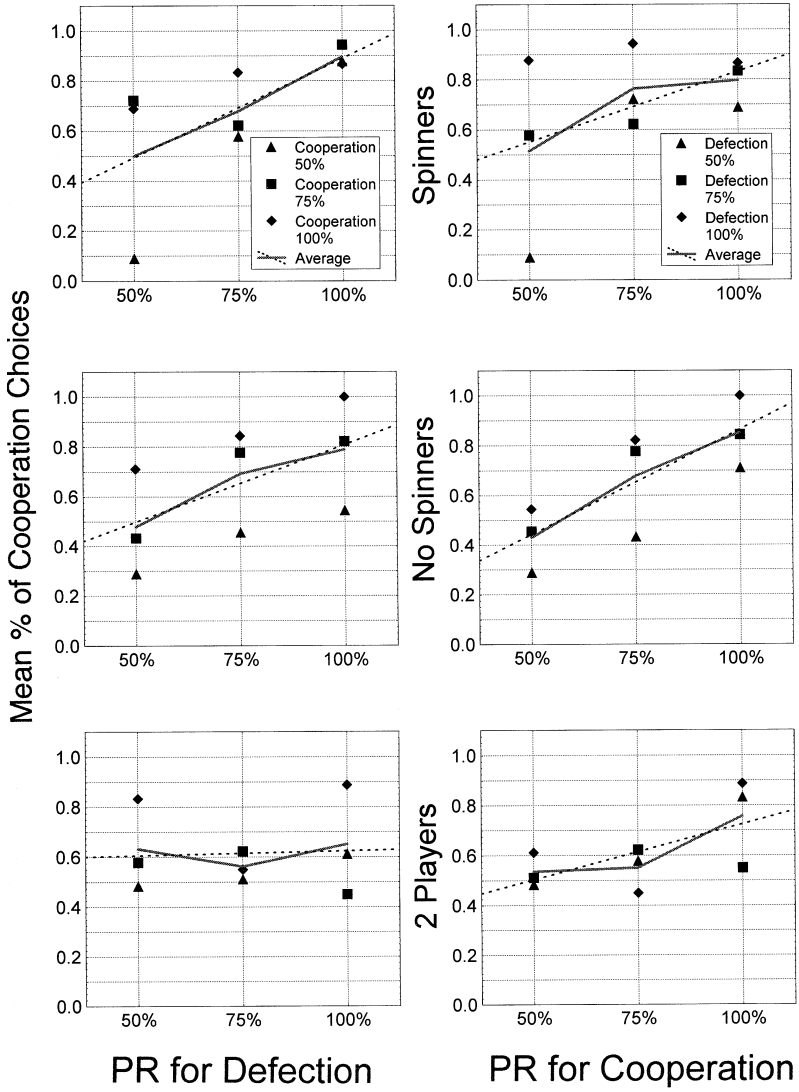


Fig. 17. Results of Experiment 9. Left column of graphs: average percent cooperation as a function of reciprocation probability (PR) of defection, PR of cooperation as parameter. Right column of graphs: average percent cooperation as a function of PR of cooperation, PR of defection as parameter. The upper, middle, and lower graphs show the results for the three conditions: spinners visible; spinners not visible; 2 players.

most noticeable characteristic of these data is the relative weakness of reciprocation probability (flatness of the regression lines) when subjects believed that they were playing against another player (the two bottom graphs). The belief that there was another player making the computer's choices and receiving benefits was the only difference between these groups and the groups playing without spinners visible (the middle graphs). The belief that there was another subject present, sensitive to their cooperation and defection, evidently made these subjects less sensitive to the actual consequences of their choices and perhaps more sensitive to whatever prior interpersonal contingencies were signaled by the instructions. These would have been constant regardless of reciprocation probability.

XI. Computer Paradigm (Tit-for-Tat versus Pavlov)

Experiment 10 (Baker, 1999) compares the two strategies, tit-for-tat (the subject as a learner) and Pavlov (the subject as a teacher). This experiment focused more sharply than did prior computer paradigm experiments on the role of the other player. When playing against tit-for-tat, the subject is a learner while the tit-for-tat player is a teacher: the subject's cooperation is reinforced by the tit-for-tat player's cooperation; the subject's defection is punished by the tit-for-tat player's defection. Axelrod (1984), in a simulated computer tournament, found that tit-for-tat performed better (earned higher reward for the player who used it) than other strategies such as always defect or always cooperate. When playing against Pavlov, on the other hand, the subject is a teacher, whereas the Pavlov player is a learner: the Pavlov player will cooperate if a prior cooperation is reinforced by the subject's cooperation or if a prior defection is punished by the subject's defection; the Pavlov player will defect if a prior cooperation is punished by the subject's defection or if a prior defection is reinforced by the subject's cooperation. Recently, tit-for-tat was outperformed by Pavlov during a computer-simulated tournament (Nowak & Sigmund, 1993).

We are aware of no prior study of human subjects playing against Pavlov. This experiment compared cooperation by subjects playing against tit-for-tat with cooperation by subjects playing against Pavlov. However, rather than using a static reciprocation probability that conformed either to Pavlov or tit-for-tat, the present experiment *changed* the reciprocation probability in 25% steps contingent on subjects' past behavior. Only one spinner was presented, but the black and white quadrants on the spinner could change during the experiment. When the computer's strategy (whether tit-for-tat or Pavlov) would normally dictate cooperation, the computer's probability of cooperation was increased by .25 or (if already

1.0) maintained; when the computer's strategy would normally dictate defection, the computer's probability of defection was increased by .25 or (if already 1.0) maintained. The point of this stepwise variation of reciprocation probability was to more closely simulate the variable contingencies of real-life social dilemmas. As in Experiment 9, subjects were paid 1 cent for each point accumulated.

Each of the two computer strategies (tit-for-tat and Pavlov) was studied under the same three conditions as previously: spinners visible; spinners not visible; spinners not visible plus imagined other player ("2-players" condition). Thus, there were a total of six groups; 8 to 10 subjects in each group played 100 trials each. Subjects were studied in groups of four or six, each at a different computer, screens of other subjects not visible. Subjects in the "two-player" groups were led to believe that one of the other players was his or her opponent, but all subjects actually played independently against the computer.

Figure 18 shows percent cooperation for the six groups over the course of the experiment. Figure 19 shows the distributions of cooperations of subjects over the last 20 trials for each group. Figure 18 shows that all groups began the experiment cooperating on about 40% of the trials. Four groups, two playing against tit-for-tat and two playing against Pavlov, responded to the experimental contingencies by increasing cooperation over the course of the experiment. Two groups, one tit-for-tat group and one Pavlov group, did not increase cooperation but continued to cooperate at the same low level throughout the experiment.

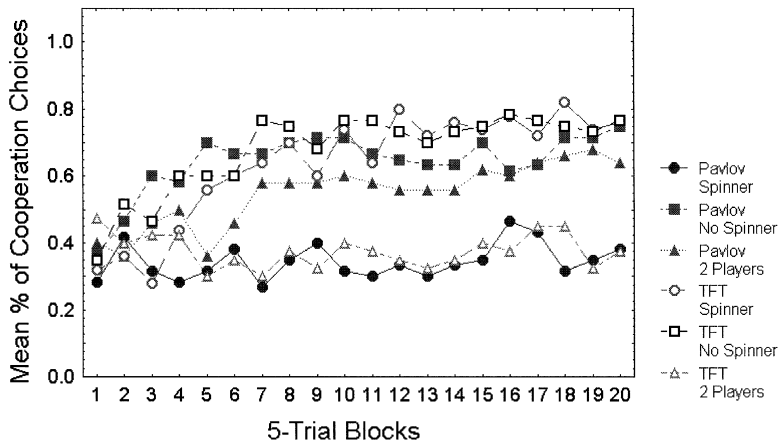


Fig. 18. Average percent cooperation for each group over the course of Experiment 10 (five-trial blocks).

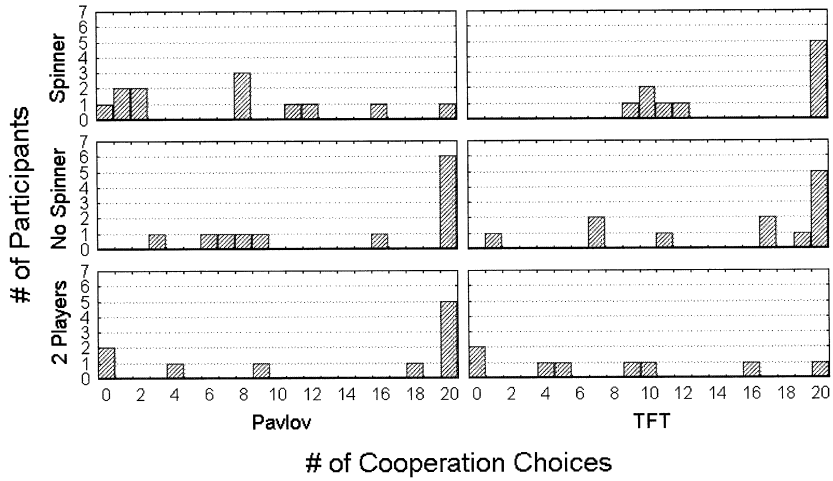


Fig. 19. Distribution of subjects in each group of Experiment 10 with respect to number of cooperations in the last 20 trials. TFT, tit-for-tat.

Figure 19 shows that for the groups that learned to cooperate, the modal subject cooperated on all of the last 20 trials, the remaining subjects being scattered across the distribution. The subjects in the groups that did not learn to cooperate (playing against Pavlov with spinners visible and playing against tit-for-tat with an imagined other player) were all scattered across the distribution. Thus, by the end of this experiment, as in the earlier experiments, there seem to be two types of players: those who learned to maximize earnings by consistently cooperating (in the present experiment and in some conditions of prior experiments) or by consistently defecting (in some conditions of prior experiments), and those who did not learn to maximize. Differences between the groups in percent cooperation reflect the number of players in a group who learned to maximize rather than the degree to which the average player learned.

What explains the pattern of results obtained in this experiment? Without further research we can only speculate. First, consider the subjects in the group playing against tit-for-tat who believed they were playing against another player (the lower right distribution of Figure 19). In Experiment 8, subjects also believed they were playing against another player. In that experiment, with PR = 100%, the modal subject cooperated on all of the last 20 trials. The difference between that experiment and this is that in the present experiment reciprocation probability changed in steps of 25%, whereas for the PR = 100% group of Experiment 8, reciprocation was certain. That is, for the tit-for-tat groups of this experiment, there were

occasions when cooperation (and defection) were unreciprocated. The proper comparison group for the tit-for-tat groups of the present experiment may thus be the PR = 75% group of Experiment 8 (Figure 16, PR = 75%), where subjects did not learn to cooperate. But this explanation raises two other problems: first, why did subjects in the no-spinners, tit-for-tat group of the present experiment learn to cooperate when the corresponding group of Experiment 7 (Figure 14, PR = 75%) actually learned to defect? Note that a constant reciprocation probability of 75% occasionally fails to reciprocate regardless of the subject's behavior, whereas a reciprocation probability that changes in 25% steps reaches 100% after an average of two cooperations in a row. The subjects in the no-spinners, tit-for-tat condition of the present experiment who (disregarding the computer's occasional failure to reciprocate cooperation) put together a string of cooperations, reached this point. There they would effectively be in a PR = 100% condition. In the present experiment's ''two player'' condition, on the other hand, when playing against another player who seemed to profit by defecting when the subject cooperated, occasional failures to reciprocate cooperation could not be ignored. Subjects in that condition would have resisted being so manipulated and tried to punish these defections by defecting (Skinner's, 1953, ''countercontrol'').

Now consider the group playing against Pavlov with spinner visible. In all previous experiments, with subjects playing against tit-for-tat, the presence of spinners enhanced whatever behavior was reinforced by the experimental contingencies. What, then, about playing against the Pavlov contingency, could cause the presence of spinners to actually retard cooperation? One possibility is that the spinners, with their clear indications of reciprocation probability, narrowed the time span of the subjects' sensitivity to reinforcement and punishment to the computer's choice on the very next trial, the choice signaled by the spinner. With the tit-for-tat contingency, the computer's response depends only on the subject's current choice; with the Pavlov contingency the computer's response depends on the computer's prior choice as well as the subject's current choice. The Pavlov contingency is thus more complicated than the tit-for-tat contingency and requires a wider time span to be played out. The spinner, with its vivid next-trial information, might have hindered the development of this wider sensitivity.

Second, subjects might not have expected a mechanical spinner to be responsive to their attempts to reward or punish it; hence they played passively, their behavior reinforced or punished (as would be appropriate against tit-for-tat), instead of actively reinforcing or punishing the computer's choices (as was appropriate against Pavlov). Without the strong discriminative stimulus provided by the spinner, on the other hand, subjects playing against Pavlov may have been more sensitive to the overall positive correla-

tion between cooperation and reward than when spinners were present. These speculations would, of course, have to be confirmed by further experimentation. For example, the number of prior trials influencing reciprocation may be varied (see Kudadjie-Gyamfi & Rachlin, 1996). As this number increases, signals of probability of cooperation or defection by the computer on the current trial may be less effective in generating cooperation by the subject.

XII. Conclusions

Although both tit-for-tat and Pavlov were effective in generating cooperation, they did it in different ways and under different conditions. Trial-by-trial analysis of subjects' choices in all of the experiments reported here occasionally found patterns of behavior that conformed more to tit-for-tat or more to Pavlov. However, in no case was either of these strategies used consistently by subjects who learned to cooperate. Instead, those subjects seem to have been sensitive to the correlation over time between cooperation and reward (which, in the prisoner's dilemma, is opposed by immediate reinforcement of defection). Where that correlation was signaled clearly (as when playing against tit-for-tat with spinners visible), cooperation was enhanced. Where that correlation was degraded (as in the PR = 75% conditions), or unsignaled (as with spinners not visible), or obscured by a signal (as when playing against Pavlov with spinners), or overshadowed by subjects' expectancies (as in the 2-player conditions), cooperation tended to deteriorate. In some cases this correlation was imposed within the experiment itself. In other cases, contextual contingencies (such as those imposed by extra-experimental interpersonal interactions or interactions with computers) were signaled by the instructions. In no case is there any evidence that subjects learned the computer's strategy and countered it with a strategy of their own so as to maximize reinforcement. That is, prisoner's dilemma paradigms were treated by subjects not as puzzles to be solved but, in the same way that self-control situations are treated (Rachlin, 2000), as conflicts between higher valued but abstract contingencies and lower valued but immediate and certain rewards.

What counts for cooperation is not the absolute probability of future cooperation or even the subject's subjective estimation of that probability. It is in the nature of the prisoner's dilemma that it pays to defect regardless of the *absolute* probability that others will cooperate (or, in self-control situations, the absolute probability of an individual cooperating with his or her own future interests). These experiments show that what counts is the *relative,* or *conditional,* probability, the probability of reciprocation over

a series of opportunities to cooperate. What counts, in other words, is the probability of reinforcement of cooperation and the probability of punishment of defection.

The important question for both social cooperation and self-control is not, Will others cooperate (or will I cooperate) in the future? but, *If I cooperate now,* will others cooperate (or will I cooperate) in the future? When self-control breaks down, as in cases of addiction, the probability of reciprocation plummets. The alcoholic who has quit a thousand times, must also have started up again (defected) at least 999 times. For an alcoholic there may be no reason to cooperate today with the person that the alcoholic himself will be tomorrow. The probability that *that* future person, that incarnation of today's decision maker, will reciprocate today's drink refusal may well be low. It is not possible to change that probability at a single moment any more than it is possible, in Aristotle's analogy, for a single retreating soldier to turn and stop an enemy advance. But, particular acts have to be done. This is why the establishment of self-control, as well as social cooperation, when none was there before, requires a sort of faith. This faith is embodied in an act of imagination—acting as if something were true that is not in fact true—acting as if the probability of reciprocation is high when it has been low in the past. Enough such acts will create the very thing imagined. Such acts are part of a still wider pattern of behavior, the pattern we refer to when we talk about the concept of ''self.''

Acknowledgments

The research reported here was supported by grants from the National Institute of Mental Health and the National Institute for Drug Abuse. We thank James C. Baker for programming the computer paradigm experiments.

References

Axelrod, R. (1980). Effective choice in the prisoner's dilemma. *Journal of Conflict Resolution, 24,* 3–25.

Axelrod, R. (1984). *The evolution of cooperation.* New York: Basic Books.

Baker, F. (1999). *Cooperation during individual and social dilemmas.* Unpublished thesis, Psychology Department, State University of New York at Stony Brook.

Bixenstein, V. W., & Gaebelein, J. W. (1971). Strategies of ''real'' opponents in eliciting cooperative choice in a prisoner's dilemma game. *Journal of Conflict Resolution, 15,* 157–166.

Brown, J., & Rachlin, H. (1999). Self-control and social cooperation. *Behavioural Processes, 47,* 65–72.

Caporael, L. R., Dawes, R. M., Orbell, J. M., & van de Kragt, A. J. C. (1989). Selfishness examined: Cooperation in the absence of egoistic incentives. *Behavioral and Brain Sciences, 12,* 683–739.

Herrnstein, R. J. (1966). Superstition: A corollary of the principles of operant conditioning. In W. K. Honig (Ed.), *Operant behavior: Areas of research and application.* Englewood Cliffs, NJ: Prentice-Hall.

Komorita, S. S. (1994). *Social dilemmas.* Dubuque, IA: Wm. C. Brown Communications, Inc.

Komorita, S. S., Hilty, J. A., & Parks, C. D. (1991). Reciprocity and cooperation in social dilemmas. *Journal of Conflict Resolution, 35,* 494–518.

Kudadjie-Gyamfi, E., & Rachlin, H. (1996). Temporal patterning in choice among delayed outcomes. *Organizational Behavior and Human Decision Processes, 65,* 61–67.

Logue, A. W. (1988). Research on self-control: An integrating framework. *Behavioral and Brain Sciences, 11,* 665–710.

Nowak, M., & Sigmund, K. (1993). Expecting continued play in Prisoner's Dilemma games: A test of several models. *Journal of Conflict Resolution, 27,* 279–300.

Rachlin, H. (1995). The value of temporal patterns in behavior. *Current Directions in Psychological Science, 4,* 188–192.

Rachlin, H. (2000). *The science of self-control.* Cambridge, MA: Harvard University Press.

Rachlin, H., Castrogiovanni, A., & Cross, D. (1987). Probability and delay in commitment. *Journal of The Experimental Analysis of Behavior, 48,* 347–354.

Rachlin, H., Logue, A. W., Gibbon, J., & Frankel, M. (1986). Cognition and behavior in studies of choice. *Psychological Review, 93,* 33–45.

Rachlin, H., & Siegel, E. (1994). Temporal patterning in probabilistic choice. *Organizational Behavior and Human Decision Processes, 59,* 161–176.

Rapoport, A., & Chammah, A. (1965). *Prisoner's dilemma.* Ann Arbor: University of Michigan Press.

Siegel, E., & Rachlin, H. (1995). Soft commitment: Self-control achieved by response persistence. *Journal of The Experimental Analysis of Behavior, 64,* 117–128.

Silverstein, A., Cross, D., Brown, J., & Rachlin, H. (1998). Prior experience and patterning in a prisoner's dilemma game. *Journal of Behavioral Decision Making, 11,* 123–138.

Skinner, B. F. (1953). *Science and human behavior.* New York: The Free Press.

INDEX

A

Absence of relation, 299
Abstraction, in causal status effect, 39–40
Abstract tool schema
  in memory
    experiments, 292–294
    predictions, 290–292
  for spatial diagrams, 291
Activity theory, 85–86
Act–outcome expectation, 267–269
Act–outcome representation, 258
Adjacency matrix, 107
AFR, *see* Contour frame was absent
Aging
  adult
    lag effects, 157–162
    primary results, 154–157
  and amnesia, 173–175
Alliesthesia, 237–239
γ-Aminobutyric acid receptor, 255–256
Amnesia
  and aging, 173–175
  anterograde, sequence learning, 149–154
  recollection and priming, 131
Annoyer, in Law of Effect, 225–226
Appreciation, for intentionality, 35–36
Artifacts
  contrast with natural kinds, 24
  vs. natural kinds, causal status hypothesis
    appreciation for intentionality, 35–36
    category use effect, 34–35
    developmental studies, 36–37
    related phenomena, 33–34
Associative recognition, 192
Assumptions, LTM, 3
Asymmetry, functional brain puzzle, 2–3
Atheoretical behaviorism, 224–225
Attenuation, cooperation deterioration, 339
Attractiveness, faces, 18
Auditory modality, 136–137
Autoshaping
  conditioned incentive stimulus, 246–248
  in incentive theory, 235

B

Bartlett, Frederic C., 72–74
Behavior
  atheoretical and radical, as reinforcement theory, 224–225
  in prisoner's dilemma, 328
Beliefs, causal, plausibility, 46–47
Bidirectional relations, causal status hypothesis, 58–59
Bindra–Toates theory, 263–265
Biological logic, hemispheric LTM
  processing experiences, 17
  processing facial attractiveness, 18
Bolles–Bindra–Toates theory
  conditioned incentives, 237–239, 245–246
  CS as incentive, 236
  drive and motivation, 244–245
  sensory preconditioning of hedonic value, 250–252
  S–S reward *vs.* S–R reinforcement, 234–235

Brain
  functional asymmetry, 2–3
  incentive value cognitive expectation, 269–271
  mesolimbic dopamine, 267–269
  mesolimbic dopamine substrates, 265–267
  mesolimbic dopamine systems, 265
  reward learning mechanism, 264
  substrates of cognitive incentive learning, 267
Brain waves, 135

C

Categorization
  causal status effect
    abstraction levels, 39–40
    application, 52–57
    appreciation for intentionality, 35–36
    bidirectional relations, 58–59
    category use effect, 34–35
    causal *vs.* isolated features, 49–52
    common-effect structure
      real-life example, 41–43
      Rehder and Hastie's study, 43–46
    computational modeling, 37–38
    developmental studies, 36–37
    directional relations, 59–61
    domain generality, 47–48
    empirical results, 31–33
    essentialism, 29–31
    fast *vs.* slow judgments, 48–49
    feature centrality determinants, 38–39
    name *vs.* concenceptual centrality, 39–40
    overview, 27–29
    plausibility of causal beliefs, 46–47
  feature centrality
    content-based approach, 23–26
    statistical approach, 26–27
    theory-based approach, 27
  logical
    family resemblance *vs.* logical membership, 10–11
    parity membership, 11–12
  natural
    effect of different instances, 6–9
    representativeness effect, 9–10
  natural superordinate, in LTM, 5–10
Category use effect, 34–35
Causal beliefs, 46–47
Causal status hypothesis
  abstraction levels, 39–40
  application, 52–57
  bidirectional relations, 58–59
  causal *vs.* isolated features, 49–52
  common-effect structure
    real-life example, 41–43
    Rehder and Hastie's study, 43–46
  computational modeling, 37–38
  directional relations, 59–61
  domain generality, 47–48
  empirical results, 31–33
  and essentialism, 29–31
  fast *vs.* slow judgments, 48–49
  feature centrality, other determinants, 38–39
  name *vs.* conceptual centrality, 39–40
  natural kinds *vs.* artifacts
    appreciation for intentionality, 35–36
    category use effect, 34–35
    developmental studies, 36–37
    related phenomena, 33–34
  overview, 27–29
  plausibility of causal beliefs, 46–47
Central motivational state of the hedonic stimulus, 236
Central tendency, 5–6
Clique
  size, 114
  social network analysis, 110–111
Clue, from familiar schema distortion, 4
Cognition, understanding with neuroscience, 122–124
Cognitive expectation
  incentive value, 258–260, 269–271
  and Pavlovian incentive salience
    pulling apart, 259–260
    semipermeable boundary, 261–264
Cognitive incentive learning, brain substrates, 267
Cognitive neuroscience
  human memory, 138–140
  for understanding, 122–124
Cognitive psychology, prisoner's dilemma analysis, 330
Cognitive systems, information-processing system
  group, 81–83
  individual, group, and society, 83–85

Collaborative inhibition
in other tasks, 99–105
in remembering, 92–99
Collaborative remembering
fundamental questions, 91
inhibition, 92–99
inhibition in other tasks, 99–105
Collective memory
fundamental questions, 91
remembering as social process, 86–90
social network analysis, 105–111
story, 111–115
Collective Memory, The, 70–72
Commissurotomy, 3–5
Common-effect structure, 41–43
Communication, in prisoner's dilemma, 330
Computational modeling, causal status hypothesis, 37–38
Computer paradigm, prisoner's dilemma, PR
with other player, 354–356
reinforcement and punishment, 356–358
with spinners, 348–353
without spinners, 353–354
Computer science student, spatial diagram knowledge, 312–315
Conceptual centrality *vs.* name, 39–40
Conditioned incentives
autoshaping, 246–248
CS consumption, 248
in incentive theory, 237–239
motivational properties, 245–246
sensory preconditioning
hedonic value, 250–252
incentive value, 250
Conditioned stimulus
auditory, for sucrose, 240
consumption, 248
as incentive, 236
Conjoint recognition
identifiable parameters, 204
process dissociation extension, 202
recall-to-reject processing, 212–216
ROC curves, 201–202
system of equations, 202–203
Conscious recollection, 125
Consequences, incentive model, 239–243
Consolidation, in declarative memory, 129
Consumption, CS, 248
Content-based approach, to feature centrality, 23–26
Contingencies, in prisoner's dilemma, 330
Contour frame was absent, 14–16
Contour frame was present, 14–16
Cooperation
in prisoner's dilemma, 327
in prisoner's dilemma computer paradigm, 356–358
in prisoner's dilemma game board paradigm, 345–347
in prisoner's dilemma playing card paradigm, 332–335, 338–339
CS, *see* Conditioned stimulus
Cued recall, 104–105
Cue specificity, 97–98

D

Declarative memory, 124–131
Defection, prisoner's dilemma, 327, 340, 356–358
Degree, in relational matrix, 109
Density
network cohesion, 110
standard deviation, 113–114
Deterioration, cooperation in prisoner's dilemma, 339
Dickinson–Balleine theory
cognitive expectation, 260–261
cognitive expectation and Pavlovian incentive salience, 261–264
incentives, 257–258
incentive value cognitive expectation, 258–260
Directional relations, 59–61
Distortion
familiar schema, 4
famous painting, in commissurotomy, 4–5
Domain, in causal status effect, 47–48
Dopamine, mesolimbic system
and cognitive act–outcome expectation, 267–269
incentive salience, 265
Pavlovian incentive salience, 265–267

Drive cue
in incentive theory, 237–239
as motivation, 244–245
Drive reduction theory
in human motivation, 231–232
in physiological psychology, 232–234
as reinforcement theory, 228–231
in social psychology, 231–232
as strong reward, 243–244
DSM-IV, causal status hypothesis application, 52–57
Durkheim, Émile, social nature of memory, 69–70

E

Economists, prisoner's dilemma analysis, 328–329
Electrophysiology
perceptual priming, 137–138
recollection, 131–137
Elementary Forms of the Religious Life, The, 69–70
Equations, conjoint recognition, 202–203
ERP, *see* Event-related potentials
Essentialism, 29–31
Event-related potentials
facial photographs, 136
memory strength, 134
perceptual priming, 137–138
recognition, 133
recollection, 131–137
relative familiarity, 134–135
spoken words, 136–137
visual words, 132–133
Evoked response potential, 184
Evolution, for facial attractiveness processing, 18
Exemplar knowledge, in diagrammatic tool choice
evidence for use, 285–287
representational transfer studies, 287–288
tool selection studies, 288–290
Exhaustive search process, in recall, 186–187
Experiences, hemispheric LTM, 17
Experimental conditions, for recall, 205–206
Expertise, spatial diagram knowledge, 306–312

F

Faces
attractiveness, processing, 18
normal, schemas in LTM, 14–16
photographs, ERPs, 136
Familiarity
relative, ERPs, 134–135
ROC curves, 197–198, 216–217
Family resemblance, 10–11
Feature centrality
in causal status effect, 38–39
content-based approach, 23–26
statistical approach, 26–27
theory-based approach, 27
Features, causal *vs.* isolated, 49–52
Free recall, collaborative inhibition, 104
Frequency, reversals, effect on sequence learning
experimental methodology differences, 164
explicit knowledge, 164–165
general method, 162–163
sequence position effects, 168–172
SRT performance, 165–168

G

GABA receptor, *see* γ-Aminobutyric acid receptor
Game board paradigm, prisoner's dilemma
patterned trials, 347–348
reward for cooperation, 345–347
single trials
apparatus, 340–342
income maximization, 343–344
rules, 342
subject situation effect, 344–345
tit-for-tat play, 342–343
Group differences
adult aging effects
lag effects, 157–162
primary results, 154–157
as information-processing system
group only, 81–83
individual, group, and society, 83–85
sequence learning in amnesic patients, 149–154
Group-level memory, transactive memory system, 88–90
Groupthink, definition, 87

H

Halbwachs, Maurice, 70–72
Hedonic experience, 258–260
Hedonic modulation, 237–239
Hedonic state, 240
Hedonic value, 250–252
Hemi-field technique, 5–6
Hemispheric specialization–LTM interaction, 16–17
Hierarchy diagram, 309–310
Hutchins, Edwin, 75–76
Hypothesis, causal status, *see* Causal status hypothesis

I

Implicit memory tests, 125
Incentive expectations, 257–258
Incentive model
  cognitive learning, brain substrates, 267
  conditioned, *see* Conditioned incentives
  motivation, by Pavlovian cues, 248–249
  motivation and reward
    consequences, 239–243
    drive reduction, 243–244
  Pavlovian, *see* Pavlovian incentive
  primary stimulus, motivational properties, 245–246
  reward and motivation, drive, 244–245
Incentive salience, 264–265
Incentive theory
  conditioned incentives, 237–239
  CS as incentive, 236
  S–S reward *vs.* S–R reinforcement, 234–235
Incentive value
  cognitive expectation, 258–260, 269–271
  sensory preconditioning, 250
Incidence matrix, 106–108
Individual
  with group and society, 83–85
  in social context
    social cognition, 77–78
    sociocultural effects, 79–81
Individual recall condition, 93
Information
  movement through spatial diagram, 299–300
  requisite, storage, 127–128
Information-processing system
  group as, 81–83
  individual, group, and society as, 83–85
Insular cortex, 269–271
Intentionality, appreciation, causal status hypothesis, 35–36
Items, in spatial diagrams, 298–299

J

Judgments, in causal status effect, fast *vs.* slow, 48–49

K

Knowledge
  exemplar, in diagrammatic tool choice
    evidence for use, 285–287
    representational transfer studies, 287–288
    tool selection studies, 288–290
  explicit, in SRT analysis, 164–165
  spatial diagram, student
    computer science students, 312–315
    experimental methods, 301–303
    as function of expertise, 306–312
    math education students, 312–315
    structural analysis, 303–305
    typical undergraduate, 315–317
    unresolved issues, 317

L

Lag effects
  in adult aging, 157–162
  in learning performance, 175–178
Latent inhibition, taste aversion learning, 242
Law of Effect, 225–226
Learning
  cognitive incentive, brain substrates, 267
  lag effects, 175–178
  reward
    brain mechanism, 264
    mesolimbic dopamine substrates, 265–267
    mesolimbic dopamine systems, 265
  sequence, *see* Sequence learning
  situated, remembering as social process, 85–86
  taste aversion
    latent inhibition, 242
    remembered taste, 242–243

Levels-of-processing, spoken words, 136–137
Liking, Pavlovian incentive
  liking *vs.* wanting, 252–254
  liking without wanting, 254–255
  wanting without liking, 255–257
Links, in spatial diagrams, 298–299
Logical membership, LTM logical categories, 10–11
Long-term memory
  assumptions, 3
  commissurotomy, 3–5
  –hemispheric specialization
    biological logic, 17–18
    constraints, 16–17
  logical categories, 10–12
  natural superordinate categories, 5–10
  schemas of faces, 13–16
LTM, *see* Long-term memory

M

Map-reconstruction task, 100–101
Math education student, spatial diagram knowledge, 312–315
Mathematicians, prisoner's dilemma analysis, 328–329
Matrix
  payoff, in prisoner's dilemma playing card paradigm, 333
  in representational transfer, 286–287
  in social network analysis
    adjacency matrix, 107
    incidence matrix, 106–108
    relational matrix, 107–108
Matrix diagram, 311
Membership, logical, LTM logical categories, 10–11
Memory
  abstract tool schema
    experiments, 292–294
    predictions, 290–292
  collective
    fundamental questions, 91
    remembering as social process, 86–90
    social network analysis, 105–111
    story, 111–115
  declarative, neural implementation, 124–131
  group-level, transactive memory system, 88–90
  human, cognitive neuroscience, 138–140
  implicit tests, for declarative memory, 125
  long-term, *see* Long-term memory
  processes, ERPs, 131–132
  social nature
    Bartlett, Frederic C., 72–74
    Durkheim, Émile, 69–70
    Halbwachs, Maurice, 70–72
    Hutchins, Edwin, 75–76
    supporting arguments, 68–69
    Vygotsky, L. S., 74–75
  storage, requisite information, 127–128
  strength, ERPs, 134
  transactive system, in group-level memory, 88–90
Mesolimbic system, dopamine
  and cognitive act–outcome expectation, 267–269
  incentive salience, 265
  Pavlovian incentive salience, 265–267
Models
  computational, causal status hypothesis, 37–38
  fitting with Simplex estimation method, 191
  incentive
    motivation and reward, consequences, 239–244
    reward and motivation, drive, 244–245
Motivation
  conditioned incentive stimulus, 245–246
  human, drive reduction role, 231–232
  incentive model
    consequences, 239–243
    drive, 244–245
    drive reduction, 243–244
    by Pavlovian cues, 248–249
Movement, information through spatial diagram, 299–300

N

Name *vs.* conceptual centrality, 39–40
Natural kinds
  vs. artifacts, causal status hypothesis
    appreciation for intentionality, 35–36
    category use effect, 34–35

developmental studies, 36–37
related phenomena, 33–34
contrast with artifacts, 24
Network analysis, collective story memory, 111–115
Network cohesion, 110
Network diagram, 310–311
Neural implementation, declarative memory, 124–131
Neuronal ensembles, 128–129
Neuroscience
for cognition understanding, 122–124
cognitive, human memory, 138–140
NF, *see* Normal faces
Normal faces, schemas in LTM, 14–16
Nouns, in recall process, 204–205

P

Painting, distortion in commissurotomy, 4–5
Parity category membership, LTM categories, 11–12
Patterned trials
in prisoner's dilemma game board paradigm, 347–348
in prisoner's dilemma playing card paradigm, 337–339
Pavlovian incentive
and cognitive expectation
pulling apart, 259–260
semipermeable boundary, 261–263
cues in motivation, 248–249
and incentive expectations, 257–258
liking *vs.* wanting, 252–254
liking without wanting, 254–255
mesolimbic dopamine substrates, 265–267
wanting without liking, 255–257
Pavlov strategy
in prisoner's dilemma computer paradigm, 358–362
in prisoner's dilemma playing card paradigm, 340
Payoff matrix, in prisoner's dilemma playing card paradigm, 333
Perceptual priming
electrophysiology, 137–138
in implicit memory tests, 125
Performance, in sequence learning measurement, 148
PFR, *see* Contour frame was present
Photographs, face, ERPs, 136
Physiological psychology, 232–234
Plausibility, causal beliefs, 46–47
Player, in prisoner's dilemma computer paradigm, 354–356
Playing card paradigm, prisoner's dilemma
grouped trials
patterned trials, 337–339
self-control situation, 336–337
single trials
cooperation, 332–335
payoff matrix, 333
subject *vs.* experimenter, 332–333
subject *vs.* subject, 333
tit-for-tat strategy, 335–336
Pleasure
alliesthesia, 237–238
as triggered affective state, 256
PR, *see* Probability of reciprocation
Prefrontal cortex, 269–271
Primary incentive stimulus, 245–246
Priming
in declarative memory studies, 129–131
in implicit memory tests, 125
Prisoner's dilemma game
analysis
by cognitive psychologists, 330
by economists, 328–329
by mathematicians, 328–329
by social psychologists, 330
behavior, 328
communication, 330
computer paradigm
PR with other player, 354–356
PR of reinforcement and punishment, 356–358
PR with spinners, 348–353
PR without spinners, 353–354
tit-for-tat *vs.* Pavlov strategy, 358–362
contingencies, 330
cooperation, 327
defection, 327
game board paradigm
patterned trials, 347–348
reward for cooperation, 345–347
single trials
apparatus, 340–342
income maximization, 343–344
rules, 342

Prisoner's dilemma game (*continued*)
subject situation effect, 344–345
tit-for-tat play, 342–343
playing card paradigm
grouped trials
patterned trials, 337–339
self-control situation, 336–337
single trials
cooperation, 332–333, 334–335
payoff matrix, 333
subject *vs.* experimenter, 332–333
subject *vs.* subject, 333
tit-for-tat strategy, 335–336
tit-for-tat strategy, 330–331
Probability of reciprocation, prisoner's dilemma computer paradigm
other player, 354–356
reinforcement and punishment varying, 356–358
with spinners, 348–353
without spinners, 353–354
tit-for-tat *vs.* Pavlov strategy, 358–362
Problem solving
with exemplar knowledge, 285
toolbox for thought, 279–281
Process dissociation
extension by conjoint recognition, 202
in recognition memory, 184
Psychological disorders, 52–57
Psychology
cognitive, prisoner's dilemma analysis, 330
physiological, drive reduction theory role, 232–234
social
drive reduction role, 231–232
prisoner's dilemma analysis, 330
Punishment, PR, in prisoner's dilemma computer paradigm, 356–358

R

Radical behaviorism, 224–225
Rape, The, 4–5
Reasoning
with exemplar knowledge, 285
in spatial diagrams, 318–319
Recall
cued, collaborative inhibition, 104–105
free, collaborative inhibition, 104
individual condition, 93
in recognition, 184–186
affecting factors, 193–194
conditions, 205–206
recall-to-reject processing, 212
related condition, 210–211
stimuli, 204–205
subjects, 204
target condition, 206–208
target + related condition, 212
specific nature, 186–192
Recall-to-accept process, 186
Recall-to-reject process
conjoint recognition estimation, 212–216
measurement, 194–195
in recall, 187
ROC curves, 198–201, 212
situational usage, 192
Receiver operating characteristic curve
conjoint recognition, 201–202
familiarity-based explanations, 216–217
familiarity and recollection, 198
familiarity values, 197
recall
recall-to-reject processing, 212
related condition, 210–211
target condition, 206–208
target + related condition, 212
recall in recognition, 185
recall-to-reject processing, 198–201
recognition experiments, 195–197
recognition judgments, 197–198
Recognition
associative, theory and experiment, 192
conjoint, *see* Conjoint recognition
ERPs, 133
familiarity and recognition, 198
recall process, 184–186
affecting factors, 193–194
conditions, 205–206
recall-to-reject processing, 212
related condition, 210–211
stimuli, 204–205
subjects, 204
target condition, 206–208
target + related condition, 212

recall-to-reject processing, 187, 198–201
ROC curves, 195–197
strength, ERPs, 134
Recognition judgment
ROC curves, 197–198
word lists, 189–191
Recollection
in amnesia, 131
conscious, in implicit memory tests, 125
electrophysiology, 131–137
ERPs, 135
facial photographs, ERPs, 136
ROC curves, 198
Reinforcement theory
drive reduction theory
in human motivation, 231–232
in physiological psychology, 232–234
in social psychology, 231–232
and S–R reinforcement, 228–231
Law of Effect, 225–226
PR, in prisoner's dilemma computer paradigm, 356–358
Skinner's radical behaviorism, 224–225
S–R association, 226–228
S–R reinforcement, 228–231
Related condition, for recall, 205, 210–211
Relational data, social network analysis, 105–111
Relational matrix, in social network analysis, 107–108
Relative familiarity, ERPs, 134–135
Remembered taste, 242–243
Remembering
collaborative
fundamental questions, 91
inhibition, 92–99
inhibition in other tasks, 99–105
as social process
activity theory, 85–86
collective memory, 86–90
information-processing system group, 81–83
individual, group, and society, 83–85
situated learning, 85–86
social cognition, 77–78
social development theory, 85–86
sociocultural effects, 79–81
Remember–know, in recognition memory, 184
Representational tool, appropriate selection, 281–282
Representational transfer
in exemplar knowledge studies, 287–288
with matrix representation, 286–287
Representativeness effect, for natural category exemplars, 9–10
Residual sum of squares, 191
Response signal, in recognition memory, 184
Retrieval interference, in collaborative remembering, 95–96
Reversals, frequency, effect on sequence learning
experimental methodology differences, 164
explicit knowledge, 164–165
general method, 162–163
sequence position effects, 168–172
SRT performance, 165–168
Reward
brain substrates of cognitive incentive learning, 267
for cooperation, in prisoner's dilemma game board paradigm, 345–347
incentive model
consequences, 239–243
drive reduction, 243–244
mesolimbic dopamine, 267–269
prefrontal and insular cortex, 269–271
Reward learning
brain mechanism, 263–264
mesolimbic dopamine substrates, 265–267
mesolimbic dopamine systems, 265
Rey–Osterreith figure, 101
Rigid object, shape, feature centrality, 23–24
ROC curve, *see* Receiver operating characteristic curve
RSS, *see* Residual sum of squares

S

Satisfier, in Law of Effect, 225–226
Schedule-induced polydipsia, 235
Schemas
abstract tool, *see* Abstract tool schema
commissurotomy, 3–5
faces, highly familiar visual object, 13–16

Selection task
  performance assessment, 290–291
  spatial diagrams, 319–320
Self-control situation
  laboratory studies, 337
  in prisoner's dilemma playing card paradigm, 336–337
  in real life, 337
Sensory preconditioning
  hedonic value, 250–252
  incentive value, 250
Sequence learning
  in amnesic patients, 149–154
  implicit and explicit, reversal frequency effects
    experimental methodology differences, 164
    explicit knowledge, 164–165
    general method, 162–163
    sequence position effects, 168–172
    SRT performance, 165–168
  measurement, 148
  SRT task, 147–149
Serial reaction time task
  aging and amnesia, 173–175
  in sequence learning studies, 147–149
    experimental methodology differences, 164
    explicit knowledge, 164–165
    general method, 162–163
    SRT performance, 165–168
Shape, rigid object, feature centrality, 23–24
Simplex estimation method, 191
Situated learning, 85–86
Situation, in prisoner's dilemma game board paradigm, 344–345
Social cognition, individual in social context, 77–78
Social context, individual
  social cognition, 77–78
  sociocultural effects, 79–81
Social development theory, 85–86
Social network, 105–111
Social process
  memory
    Bartlett, Frederic C., 72–74
    Durkheim, Émile, 69–70
    Halbwachs, Maurice, 70–72
    Hutchins, Edwin, 75–76
    supporting arguments, 68–69
    Vygotsky, L. S., 74–75
  remembering
    activity theory, 85–86
    collective memory, 86–90
    as information-processing system
      group, 81–83
      individual, group, and society, 83–85
    situated learning, 85–86
    social cognition, 77–78
    social development theory, 85–86
    sociocultural effects, 79–81
Social psychology
  drive reduction role, 231–232
  prisoner's dilemma analysis, 330
Society, with individual and group, 83–85
Sociocultural effects, individual in social context, 79–81
Spatial diagram tools
  abstract schema, 292
  basic structure properties, 297–298
  choice with exemplar knowledge
    evidence for use, 285–287
    representational transfer studies, 287–288
    tool selection studies, 288–290
  construction, 317–318
  for highlighting structural similarities, 282
  information movement, 299–300
  items and links, 298–299
  reasoning issues, 318–319
  selection, 319–320
  to story problems, 294
  structural analysis, 296–297
  student knowledge
    computer science students, 312–315
    experimental methods, 301–303
    as function of expertise, 306–312
    math education students, 312–315
    structural analysis, 303–305
    typical undergraduate, 315–317
    unresolved issues, 317
Spinner, in prisoner's dilemma computer paradigm, 348–354
Split-brain, *see* Commissurotomy
Spoken words, ERPs, 136–137
S–R habit, *see* Stimulus–response habit
SRT task, *see* Serial reaction time task
S–S reward, *see* Stimulus–stimulus reward
Standard deviation, density, 113–114

Statistical approach, to feature centrality, 26–27
Stimulus
  conditioned
    auditory, for sucrose, 240
    consumption, 248
    as incentive, 236
  presentation, in natural categories, 5–6
  primary incentive, motivational properties, 245–246
Stimulus–response habit, 226–228
Stimulus–response reinforcement
  as reinforcement theory, 228–231
  vs. S–S reward, 234–235
Stimulus–stimulus reward, 234–235
Story
  collective memory, 111–115
  problems, spatial diagrams, 294
Students, spatial diagram knowledge
  computer science students, 312–315
  experimental methods, 301–303
  as function of expertise, 306–312
  math education students, 312–315
  structural analysis, 303–305
  typical undergraduate, 315–317
  unresolved issues, 317
Sucrose, auditory CS, 240

T

Target condition, for recall, 205–208
Target + related condition, for recall, 205–206, 212
Tasks
  collaborative inhibition, 99–105
  selection
    performance assessment, 290–291
    spatial diagrams, 319–320
  serial reaction time, *see* Serial reaction time task
Taste aversion learning
  latent inhibition, 242
  remembered taste, 242–243
Theory-based approach, to feature centrality, 27
Tit-for-tat strategy
  in prisoner's dilemma, 330–331
  in prisoner's dilemma computer paradigm, 358–362
  in prisoner's dilemma game board paradigm, 342–343
  in prisoner's dilemma playing card paradigm, 335–336, 340
Toolbox for thought, in problem solving, 279–281
Tools
  abstract schema
    in memory, 290–294
    for spatial diagrams, 291
  representational, appropriate selection, 281–282
  selection
    abstract tools, 293
    in exemplar knowledge studies, 288–290
Transactive memory system, 88–90
Transfer–appopriate processing, 293–294
Transfer paradigm, for exemplar knowledge, 285
Traversal, in spatial diagrams, 300
Triggered affective state, pleasure as, 257

U

Undergraduate, spatial diagram knowledge, 315–317

V

Variance of degree
  network, 113
  in social network analysis, 109
Visual object, in face schemas, 13–16
Visual words, ERPs, 132–133
Vygotsky, L. S., 74–75

W

Wanting, Pavlovian incentive
  liking *vs.* wanting, 252–254
  liking without wanting, 254–255
  wanting without liking, 255–257
War of the Ghosts
  collaborative remembering inhibition, 94–95
  collective memory, 111–112
Words, list, recall process, 189–191

This Page Intentionally Left Blank

CONTENTS OF RECENT VOLUMES

Volume 29

Introduction: A Coupling of Disciplines in Categorization Research
Roman Taraban
Models of Categorization and Category Learning
W. K. Estes
Three Principles for Models of Category Learning
John K. Kruschke
Exemplar Models and Weighted Cue Models in Category Learning
Roman Taraban and Joaquin Marcos Palacios
The Acquisition of Categories Marked by Multiple Probabilistic Cues
Janet L. McDonald
The Evolution of a Case-Based Computational Approach to Knowledge Representation, Classification, and Learning
Ray Bareiss and Brian M. Slator
Integrating Theory and Data in Category Learning
Raymond J. Mooney
Categorization, Concept Learning, and Problem-Solving: A Unifying View
Douglas Fisher and Jungsoon Park Yoo
Processing Biases, Knowledge, and Context in Category Formation
Thomas B. Ward
Categorization and Rule Induction in Clinical Diagnosis and Assessment
Gregory H. Mumma
A Rational Theory of Concepts
Gregory L. Murphy
Concept Structure and Category Boundaries
Barbara C. Malt
Non-Predicating Conceptual Combinations
Edward J. Shoben
Exploring Information about Concepts by Asking Questions
Arthur C. Graesser, Mark C. Langston, and William B. Baggett
Hidden Kind Classifications
Edward Wilson Averill
Is Cognition Categorization?
Timothy J. van Gelder
What Are Concepts? Issues of Representation and Ontology
William F. Brewer
Index

Volume 30

Perceptual Learning
Felice Bedford
A Rational-Constructivist Account of Early Learning about Numbers and Objects
Rochel Gelman
Remembering, Knowing, and Reconstructing the Past
Henry L. Roediger III, Mark A. Wheeler, and Suparna Rajaram
The Long-Term Retention of Knowledge and Skills
Alice F. Healy, Deborah M. Clawson, Danielle S. McNamara, William R. Marmie, Vivian I. Schneider, Timothy C. Rickard, Robert J. Crutcher, Cheri L.

King, K. Anders Ericsson, and Lyle E. Bourne, Jr.
A Comprehension-Based Approach to Learning and Understanding
Walter Kintsch, Bruce K. Britton, Charles R. Fletcher, Eileen Kintsch, Suzanne M. Mannes, and Mitchell J. Nathan
Separating Causal Laws from Causal Facts: Pressing the Limits of Statistical Relevance
Patricia W. Cheng
Categories, Hierarchies, and Induction
Elizabeth F. Shipley
Index

Volume 31

Associative Representations of Instrumental Contingencies
Ruth M. Colwill
A Behavioral Analysis of Concepts: Its Application to Pigeons and Children
Edward A. Wasserman and Suzette L. Astley
The Child's Representation of Human Groups
Lawrence A. Hirschfeld
Diagnostic Reasoning and Medical Expertise
Vimla L. Patel, José F. Arocha, and David R. Kaufman
Object Shape, Object Name, and Object Kind: Representation and Development
Barbara Landau
The Ontogeny of Part Representation in Object Concepts
Philippe G. Schyns and Gregory L. Murphy
Index

Volume 32

Cognitive Approaches to Judgment and Decision Making
Reid Hastie and Nancy Pennington
And Let Us Not Forget Memory: The Role of Memory Processes and Techniques in the Study of Judgment and Choice
Elke U. Weber, Wiliam M. Goldstein, and Sema Barlas
Content and Discontent: Indications and Implications of Domain Specificity in Preferential Decision Making
William M. Goldstein and Elke U. Weber
An Information Processing Perspective on Choice
John W. Payne, James R. Bettman, Eric J. Johnson, and Mary Frances Luce
Algebra and Process in the Modeling of Risky Choice
Lola L. Lopes
Utility Invariance Despite Labile Preferences
Barbara A. Mellers, Elke U. Weber, Lisa D. Ordóñez, and Alan D. J. Cooke
Compatibility in Cognition and Decision
Eldar Shafir
Processing Linguistic Probabilities: General Principles and Empirical Evidence
David V. Budescu and Thomas S. Wallsten
Compositional Anomalies in the Semantics of Evidence
John M. Miyamoto, Richard Gonzalez, and Shihfen Tu
Varieties of Confirmation Bias
Joshua Klayman
Index

Volume 33

Landmark-Based Spatial Memory in the Pigeon
Ken Cheng
The Acquisition and Structure of Emotional Response Categories
Paula M. Niedenthal and Jamin B. Halberstadt
Early Symbol Understanding and Use
Judy S. DeLoache
Mechanisms of Transition: Learning with a Helping Hand
Susan Goldin-Meadow and Martha Wagner Alibali
The Universal Word Identification Reflex
Charles A. Perfetti and Sulan Zhang

Prospective Memory: Progress and Processes
Mark A. McDaniel
Looking for Transfer and Interference
Nancy Pennington and Bob Rehder
Index

Volume 34

Associative and Normative Models of Causal Induction: Reacting to versus Understanding Cause
A. G. Baker, Robin A. Murphy, and Frédéric Vallée-Tourangeau
Knowledge-Based Causal Induction
Michael R. Waldmann
A Comparative Analysis of Negative Contingency Learning in Humans and Nonhumans
Douglas A. Williams
Animal Analogues of Causal Judgment
Ralph R. Miller and Helena Matute
Conditionalizing Causality
Barbara A. Spellman
Causation and Association
Edward A. Wasserman, Shu-Fang Kao, Linda J. Van Hamme, Masayoshi Katagiri, and Michael E. Young
Distinguishing Associative and Probabilistic Contrast Theories of Human Contingency Judgment
David R. Shanks, Francisco J. Lopez, Richard J. Darby, and Anthony Dickinson
A Causal-Power Theory of Focal Sets
Patricia W. Cheng, Jooyong Park, Aaron S. Yarlas, and Keith J. Holyoak
The Use of Intervening Variables in Causal Learning
Jerome R. Busemeyer, Mark A. McDaniel, and Eunhee Byun
Structural and Probabilistic Causality
Judea Pearl
Index

Volume 35

Distance and Location Processes in Memory for the Times of Past Events
William J. Friedman
Verbal and Spatial Working Memory in Humans
John Jonides, Patricia A. Reuter-Lorenz, Edward E. Smith, Edward Awh, Lisa L. Barnes, Maxwell Drain, Jennifer Glass, Erick J. Lauber, Andrea L. Patalano, and Eric H. Schumacher
Memory for Asymmetric Events
John T. Wixted and Deirdra H. Dougherty
The Maintenance of a Complex Knowledge Base After Seventeen Years
Marigold Linton
Category Learning As Problem Solving
Brian H. Ross
Building A Coherent Conception of HIV Transmission: A New Approach to Aids Educations
Terry Kit-fong Au and Laura F. Romo
Spatial Effects in the Partial Report Paradigm: A Challenge for Theories of Visual Spatial Attention
Gordon D. Logan and Claus Bundesen
Structural Biases in Concept Learning: Influences from Multiple Functions
Dorrit Billman
Index

Volume 36

Learning to Bridge Between Perception and Cognition
Robert L. Goldstone, Philippe G. Schyns, and Douglas L. Medin
The Affordances of Perceptual Inquiry: Pictures Are Learned From the World, and What That Fact Might Mean About Perception Quite Generally
Julian Hochberg
Perceptual Learning of Alphanumeric-Like Characters
Richard M. Shiffrin and Nancy Lightfoot
Expertise in Object and Face Recognition
James Tanaka and Isabel Gauthier
Infant Speech Perception: Processing Characteristics, Representational Units, and the Learning of Words
Peter D. Eimas
Constraints on the Learning of Spatial Terms: A Computational Investigation
Terry Regier

Learning to Talk About the Properties of Objects: A Network Model of the Development of Dimensions
Linda B. Smith, Michael Gasser, and Catherine M. Sandhofer
Self-Organization, Plasticity, and Low-Level Visual Phenomena in a Laterally Connected Map Model of the Primary Visual Cortex
Risto Mikkulainen, James A. Bednar, Yoonsuck Choe, and Joseph Sirosh
Perceptual Learning From Cross-Modal Feedback
Virginia R. de Sa and Dana H. Ballard
Learning As Extraction of Low-Dimensional Representations
Shimon Edelman and Nathan Intrator
Index

Volume 37

Object-Based Reasoning
Miriam Bassok
Encoding Spatial Representation Through Nonvisually Guided Locomotion: Tests of Human Path Integration
Roberta L. Klatzky, Jack M. Loomis, and Reginald G. Golledge
Production, Evaluation, and Preservation of Experiences: Constructive Processing in Remembering and Performance Tasks
Bruce W. A. Whittlesea
Goals, Representations, and Strategies in a Concept Attainment Task: The EPAM Model
Fernand Gobet, Howard Richman, Jim Staszewski, and Herbert A. Simon
Attenuating Interference During Comprehension: The Role of Suppression
Morton Ann Gernsbacher
Cognitive Processes in Counterfactual Thinking About What Might Have Been
Ruth M. J. Byrne
Episodic Enhancement of Processing Fluency
Michael E. J. Masson and Colin M. MacLeod
At a Loss *From* Words: Verbal Overshadowing of Perceptual Memories
Jonathan W. Schooler, Stephen M. Fiore, and Maria A. Brandimonte
Index

Volume 38

Transfer-Inappropriate Processing: Negative Priming and Related Phenomena
W. Trammell Neil and Katherine M. Mathis
Cue Competition in the Absence of Compound Training: Its Relation to Paradigms of Interference Between Outcomes
Helena Matute and Oskar Pineño
Sooner or Later: The Psychology of Intertemporal Choice
Gretchen B. Chapman
Strategy Adaptivity and Individual Differences
Christian D. Schunn and Lynne M. Reder
Going Wild in the Laboratory: Learning About Species Typical Cues
Michael Domjan
Emotional Memory: The Effects of Stress on ''Cool'' and ''Hot'' Memory Systems
Janet Metcalfe and W. Jake Jacobs
Metacomprehension of Text: Influence of Absolute Confidence Level on Bias and Accuracy
Ruth H. Maki
Linking Object Categorization and Naming: Early Expectations and the Shaping Role of Language
Sandra R. Waxman
Index

Volume 39

Infant Memory: Cues, Contexts, Categories, and Lists
Carolyn Rovee-Collier and Michelle Gulya
The Cognitive-Initiative Account of Depression-Related Impairments in Memory
Paula T. Hertel

Relational Timing: A Theromorphic Perspective
J. Gregor Fetterman
The Influence of Goals on Value and Choice
Arthur B. Markham and C. Miguel Brendl
The Copying Machine Metaphor
Edward J. Wisniewski
Knowledge Selection in Category Learning
Evan Heit and Lewis Bott

This Page Intentionally Left Blank