Constructions of REMEMBERING and METACOGNITION

Essays in Honour of Bruce Whittlesea

Edited by

PHILIP A. HIGHAM and JASON P. LEBOE

Constructions of Remembering and Metacognition

Constructions of Remembering and Metacognition

Essays in Honour of Bruce Whittlesea

Edited by

Philip A. Higham
University of Southampton, UK

and

Jason P. Leboe
University of Manitoba, Canada

palgrave
macmillan

© Selection and editorial matter Philip A. Higham and Jason P. Leboe 2011

All rights reserved. No reproduction, copy or transmission of this publication may be made without written permission.

No portion of this publication may be reproduced, copied or transmitted save with written permission or in accordance with the provisions of the Copyright, Designs and Patents Act 1988, or under the terms of any licence permitting limited copying issued by the Copyright Licensing Agency, Saffron House, 6-10 Kirby Street, London EC1N 8TS.

Any person who does any unauthorized act in relation to this publication may be liable to criminal prosecution and civil claims for damages.

The authors have asserted their rights to be identified as the authors of this work in accordance with the Copyright, Designs and Patents Act 1988.

First published 2011 by
PALGRAVE MACMILLAN

Palgrave Macmillan in the UK is an imprint of Macmillan Publishers Limited, registered in England, company number 785998, of Houndmills, Basingstoke, Hampshire RG21 6XS.

Palgrave Macmillan in the US is a division of St Martin's Press LLC, 175 Fifth Avenue, New York, NY 10010.

Palgrave Macmillan is the global academic imprint of the above companies and has companies and representatives throughout the world.

Palgrave® and Macmillan® are registered trademarks in the United States, the United Kingdom, Europe and other countries.

ISBN 978–0–230–57941–5 hardback

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

10 9 8 7 6 5 4 3 2 1
20 19 18 17 16 15 14 13 12 11

Printed and bound in Great Britain by
CPI Antony Rowe, Chippenham and Eastbourne

Phil: *To my mum*

Jason: *To Launa and Delica, the key support beams for my own construction of things*

Contents

Tables and Figures

Tables

Figures

Acknowledgements

First and foremost, we would like to thank our contributors for providing us with essays for this volume. Without their hard work and dedication to science, this dedication to Bruce would not have been possible. We would also like to thank Tamara Ansons, Tim Redford, and Yunqiao Wang for their assistance with formatting the volume in preparation for publication. Finally, we'd like to thank Don Addis/*St Petersburg Times* for permission to use of the cover drawing and Life Science Associates, Bayport, NY for permission to use Figure 3.1 (Copyright 2010; lifesciassoc@pipeline.com).

Preface

> 'I have done that,' says my memory. 'I cannot have done that' – says my pride, and remains adamant. At last – memory yields.
>
> Friedrich Nietzsche

It is an honour and a privilege to dedicate this volume to our friend and esteemed colleague, Bruce Whittlesea. Bruce's extraordinary contribution to the field is unique in many ways. One notable aspect has been its sheer breadth. Throughout his career, Bruce treated no aspect of psychology as off-limits, and the result was important publications in a host of top-notch journals on perception, memory, metacognition, decision making, and emotion, to name but a few. To most, these topics likely appear quite disparate, with the literatures that they generate progressing pretty much independently. Indeed, most of us conducting research in more than one field might develop a variety of 'mini-theories' to account for subsets of data produced by each different paradigm. Instead, Bruce's position has been that most, if not all, psychological phenomena can and probably should be studied together because they are guided by similar fundamental processes. These processes involve what Bruce dubbed as production and evaluation, which form the basis of his comprehensive theory of the human mind: the *Selective Construction and Preservation of Experience* (SCAPE) model of memory. As with all good theories, SCAPE is surprisingly simple and yet widely applicable. And in line with Friedrich Nietzsche's observation above, a fundamental assumption of SCAPE is that experience of our personal past is inferential and reconstructive in nature.

What follows are 17 essays to celebrate Bruce's remarkable contribution to experimental psychology. The volume is divided into five parts, each containing one or more essays. We begin with an essay by Arnold who provides an account of the reasons that it is necessary to consider both objective and subjective measures of memory. The second part is devoted to essays on inferential processes and fluency/familiarity, an important topic that formed the basis of much of Bruce's early work (e.g., Whittlesea, 1993). This part contains four essays: Evans and Benjamin discuss the relationship between perceptual fluency and feelings of familiarity in recognition; Miller and Lloyd discuss developmental aspects of the fluency heuristic; Mantonakis, Bernstein, and Loftus discuss the role of fluency in producing feelings of familiarity, preference judgements, and senses; and Dienes, Scott, and Wan consider the role of fluency and familiarity in implicit learning tasks.

Although researchers have accepted for some time that feelings of familiarity in recognition have a heuristic basis (e.g., Jacoby & Dallas, 1981), the feeling of recollection has typically been viewed as being more veridical and direct (i.e., based on trace access) and less open to the impact of inferences. However, more recent research by Ansons and Leboe as well as Kurilla and Westerman, which is presented in Part III, demonstrates that recollection has a heuristic basis as well. An essay by Mazzoni and Hanczakowski, also presented in Part III, considers the role of metacognitive processes in constraining voluntary retrieval.

In Part IV, essays are presented that consider inferential processes in regulating or maximizing accuracy in a variety of different tasks. Higham demonstrates how signal detection theory can be used to model the regulation of accuracy in tasks both with and without an explicit report option. Goldsmith's commentary on Higham's essay provides an alternative account of accuracy regulation that relies on Goldsmith and Koriat's (2008) quantity-accuracy profile methodology. Next, Lindsay and Kantner as well as Hockley respectively discuss the roles of feedback and criterion setting in regulating recognition accuracy. The final two essays in Part IV explore the regulation and maximization of accuracy in tasks entailing the allocation of study time. Regarding such allocation, Moulin, Perfect, Akhtar, Williams, and Souchay consider the effect of memory impairment, whereas Dunlosky, Ariel, and Thiede detail the effect of personal agendas.

Finally, Part V is devoted to discussion of Bruce's SCAPE framework. Mantonakis and Hastie begin with a review of the data supporting SCAPE, whereas Papesh and Goldinger consider psychophysiological measures of production and evaluation, the two main components of SCAPE. Last but certainly not least, Whittlesea presents his own chapter describing experiments that involve recognition of words within sentences.

There is no doubt that we (and all our contributors) regret that Bruce is retiring from psychological research. It is rare that a mind so keen is combined with such cogency, making Bruce's articles, chapters, and presentations a pleasure to behold for any researcher in any area. Our hope is that this collection of essays, *Constructions of Remembering and Metacognition*, will go some way to recording his contribution for posterity.

PHILIP A. HIGHAM AND JASON P. LEBOE

Contributors

Shazia Akhtar is Lecturer of Psychology at Westminster-Kingsway College, London, UK. She was Chris Moulin's first PhD student and conducted her thesis on metacognition and memory rehabilitation in Mild Cognitive Impairment at the University of Leeds, UK.

Tamara L. Ansons is a doctoral candidate in the Brain & Cognitive Sciences area of the Department of Psychology at the University of Manitoba in Winnipeg, Manitoba, Canada. Broadly, her research examines the reconstructive nature of remembering, and how these reconstructive memory processes guide a plethora of cognitive and emotional responses. Currently, she is exploring how the inferential use of recollection guides consumer behaviour after exposure to a product placement that occurs within an entertainment context.

Robert Ariel is a graduate research assistant working with Dr John Dunlosky at Kent State University, Ohio, USA. His research interests involve self-regulated learning, with emphasis on monitoring and control of study time.

Michelle M. Arnold is Lecturer at the University of St Andrews, Scotland. Although Michelle has been working in the area of metacognitive judgements and decision making, her interests span a wide spectrum within the field of cognitive psychology. For example, her current research interests include the reconstructive nature of memory (e.g., source monitoring, illusory recollection), memory and identification for missing persons, judgements of remembering and forgetting, and the strategic regulation of accuracy.

Aaron S. Benjamin is Associate Professor of Psychology at the University of Illinois, Urbana-Champaign, USA. His research interests include human memory, metamemory, and decision making. Recent publications include *Successful Remembering and Successful Forgetting: A Festschrift in Honor of Robert A. Bjork* (in press), New York.

Daniel M. Bernstein is an instructor in the Department of Psychology at Kwantlen Polytechnic University in British Columbia, Canada. He obtained his PhD in Cognitive Psychology from Simon Fraser University British Columbia, Canada in 2001, under the supervision of Bruce Whittlesea, and then completed post-doctoral training at the University of Washington. His research interests include belief and memory, lifespan cognitive development, perspective-taking, and metacognition.

Zoltan Dienes is Professor of Psychology at the University of Sussex, UK. He is interested in the distinction between conscious and unconscious mental states, especially knowledge, and also in hypnosis and in Bayesian statistics. He is co-author of *Implicit Learning: Theoretical and Conceptual Issues* (with Dianne Berry) and author of *Understanding Psychology as a Science*.

John Dunlosky is Professor in the Psychology Department at Kent State University, Ohio, USA. His research interests involve self-regulated learning, including how aging in adulthood influences strategy use and memory performance and how individuals of all ages monitor ongoing learning as well as dynamically control study and retrieval. Recent publications include *Metacognition* (co-authored with Janet Metcalfe).

Karen M. Evans is a post-doctoral researcher at the Rochester Institute of Technology, USA. Her work explores the roles of stimulus characteristics, visual attention, and strategic processing on memory performance and skill acquisition across the lifespan. She also applies the findings of this research to education, design, and marketing.

Stephen D. Goldinger is Professor of Psychology at Arizona State University, USA. He has wide-ranging research interests, including spoken and printed word perception, basic memory, face perception, and individual differences in working-memory capacity. He is also interested in applying eye-tracking methods to the study of memory. With colleagues and students, he has authored 51 articles and book chapters.

Morris Goldsmith is Senior Lecturer of Psychology and researcher in Cognitive Psychology at the Institute of Information Processing and Decision Making, University of Haifa, Israel. His highly influential work with Asher Koriat on the metacognitive processes underlying the strategic regulation of memory accuracy has earned international acclaim and spawned a large number of studies and publications. His most recent accolade is the 2008 'Award for Outstanding Contribution to Research in Cognitive Psychology, presented by the Cognitive Section of the British Psychological Society'.

Maciej Hanczakowski is a PhD student at the University of Hull, UK. He graduated in Psychology from Jagellonian University, Krakow, and received a master's degree in financing. He is currently studying inhibitory processes in memory and holds a grant in Poland to study directed-forgetting mechanisms.

Reid Hastie is the Robert S. Hamada Professor of Behavioural Science at the Chicago Booth Graduate School of Business, USA. His primary research interests are in the areas of judgement and decision making (managerial, legal, medical, engineering, and personal), memory and cognition, and social psychology. He has published over 100 articles in scientific journals on these topics. He is best known for his research on legal decision making.

Philip A. Higham is Reader in the School of Psychology, University of Southampton, UK. His research is focused on both theoretical and applied aspects of memory and metamemory, primarily using signal detection theory as an analytic tool. He has obtained research funding from such bodies as the Natural Sciences and Engineering Research Council of Canada *(NSERC)*, the Economic and Social Sciences Research Council *(ESRC)*, the British Academy, and the Professional Association of Diving Instructors *(PADI)*. He is the current director of the Memory and Associative Learning Laboratory at the University of Southampton, and he has served on the editorial board of the *Canadian Journal of Experimental Psychology*.

William E. Hockley is Professor in the Department of Psychology at Wilfrid Laurier University, Ontario, Canada. His research largely addresses questions concerning the processes underlying recognition memory with the aim of specifying the roles of familiarity and recollection, distinguishing between memory-based and decision-based influences, and examining the effects of stimulus characteristics on recognition performance.

Justin Kantner is a PhD student in Cognition and Brain Sciences at the University of Victoria in British Columbia, Canada. His research examines the cognitive operations that underlie recognition memory judgements, focusing on the top-down, strategic, and heuristic processes used by the system to determine prior occurrence.

Brian P. Kurilla is a post-doctoral fellow at the University of Illinois, Urbana-Champaign, USA. His research focuses on the subjective experiences that accompany recognition, as well as the relationship between familiarity and recollection-based processes in memory.

Jason P. Leboe is Associate Professor in the Brain & Cognitive Sciences area of the Department of Psychology at the University of Manitoba in Winnipeg, Manitoba, Canada. His research focuses on memory and perception processes with a particular emphasis on sources of inefficiency and error in remembering, decision making, and performance. Dr Leboe's research has received funding from the Natural Science and Engineering Council of Canada, the Canadian Foundation for Innovation, and the Manitoba Research & Innovations Fund.

D. Stephen Lindsay is Professor of Psychology at the University of Victoria, British Columbia, Canada. Most of his research explores the 'source-monitoring' processes by which people identify the origins of their thoughts, images, and feelings, and the application of theories concerning source monitoring to everyday memory phenomena (e.g., eyewitness memory). He has published numerous articles and chapters, served as Editor of the *Journal of Experimental Psychology: General* from 2001 to 2007, and is a Fellow of the *Association for Psychological Science*.

Marianne E. Lloyd is Assistant Professor of Psychology at Seton Hall University, New Jersey, USA. Her research interests include memory development in childhood and recognition memory illusions.

Elizabeth F. Loftus is Distinguished Professor at the University of California, Irvine, USA. She holds faculty positions in three departments (Psychology & Social Behaviour; Criminology, Law & Society; and Cognitive Sciences), and in the School of Law, and is also a Fellow of the Centre for the Neurobiology of Learning and Memory. Loftus's research of the past 30 years has focused on the malleability of human memory. She has published 22 books (including the award winning *Eyewitness Testimony*) and over 450 scientific articles. She has been recognized for this research with six honorary doctorates and election to the *Royal Society of Edinburgh*, the *American Philosophical Society*, and the *National Academy of Sciences*. She is the former president of the Association for Psychological Science, the Western Psychological Association, and the American Psychology-Law Society.

Antonia Mantonakis is an Assistant Professor of Marketing at Brock University in Niagara, Canada. She obtained her PhD in Cognitive Psychology from Simon Fraser University in British Columbia, Canada in 2006 under the supervision of Bruce Whittlesea, and was then a Visiting Scholar at the University of Chicago Booth School of Business. Her research interests include memory and judgment and decision making.

Giuliana Mazzoni is Professor of Psychology at the University of Hull, UK. Her main area of investigation is memory, with a special interest in memory distortions and the effect of suggestion on memory. Her research also investigates autobiographical memory and the effect of suggestion on different types of health-related behaviours. She is the author of a metacognitive model on autobiographical memory distortions (Mazzoni & Kirsch, 2002), and of more than 100 papers, 8 books, and 31 book chapters.

Jeremy K. Miller is Assistant Professor of Psychology at Willamette University in Salem, Oregon, USA. His research is currently focused on understanding the various kinds of evidence that people use to make memory judgements as well as the strategies (both conscious and otherwise) that people use to help them accurately and efficiently make memory decisions.

Chris J. A. Moulin is a cognitive neuropsychologist at the Institute of Psychological Sciences at the University of Leeds, UK. His current research interests focus on neuropsychological impairments of memory. In particular, he is interested in the interaction of executive function and long-term memory. Research themes include metacognition (i.e., knowing what you know), inhibition (i.e., suppressing unwanted information), and the sensations of memory (e.g., déjà vu). Dr Moulin has been very involved in researching and making people aware of déjà vécu, a rare disorder of memory, similar to persistent déjà vu.

Megan H. Papesh is a graduate student at Arizona State University, anticipating the completion of her PhD in 2011. Her primary research focus is on the specificity of memory, which she studies using both behavioural and physiological indices. She also investigates the processes by which we perceive (and subsequently recognize) faces, sentences, and words.

Timothy J. Perfect is Professor of Experimental Psychology at the University of Plymouth, UK. His research interests span theoretical and applied aspects of memory and metamemory in younger and older adults. With colleagues, he has edited three books: *Models of Cognitive Aging, Applied Metacognition*, and *The Handbook of Applied Cognition, 2nd Edition*. He is author of 75 papers and book chapters, and he is Associate Editor of *Memory* and *Applied Cognitive Psychology*.

Ryan B. Scott is a research fellow at the University of Sussex, UK. He is interested in implicit learning, familiarity, subliminal perception, and detecting conscious mental states in vegetative state patients.

Celine Souchay is a cognitive neuropsychologist and Senior Lecturer at the University of Leeds, UK. Her current research interests focus on evaluation of the functioning of memory, metamemory and executive processes in normal aging, mild cognitive impairment in aging, Alzheimer's disease, Parkinson's disease, and Frontotemporal lobe dementia. She recently extended her research interests to the development of memory and metamemory in healthy children and children with learning disabilities.

Keith W. Thiede is Professor of Education at Boise State University, Idaho, USA. His research interests relate to metacognition and self-regulated learning, including how the accuracy of metacognitive monitoring affects regulation of study and learning.

Lulu Wan is affiliated with both the Chinese Academy of Sciences, China, and University of Sussex, UK. She is interested in implicit learning and familiarity. She has published in *Consciousness and Cognition*, *Acta Psychologica Sinica*, and *Psychological Science* (Chinese).

Deanne L. Westerman is Associate Professor of Psychology at Binghamton University (State University of New York), USA. Her research interests include recognition memory, the subjective experience of familiarity, and illusions of memory. She is currently an associate editor at *Memory & Cognition*.

Bruce W. A. Whittlesea. In his own words: After many years of dropping out of school to work in industry and various psychiatric institutions, I found my true vocation while a graduate student in Lee Brook's lab at McMaster University, Canada. On graduation in 1983, I took a post at

Carleton University, then moved to Mount Allison University, and finally to Simon Fraser University in 1989, retiring in 2009. I loved the freedom of academia, being at the frontier of human understanding, groping forward into the dark in collaboration with (and often in competition with) some of the finest minds in science. Having a lab granted me the joy (denied to most people in this world) of being able to ask Mother Nature serious questions about her most intimate workings; of course, she never gives a straight answer, but that was half the fun. Teaching was also a delight to me; seeing a stroke of sudden comprehension light up a student's face was a powerful and addicting reward. But all of that is in the past. I am now in my next and probably final life, living at ocean's edge on the Sunshine Coast of British Columbia. I now have the freedom to indulge in all the non-academic interests of my life: woodworking, house reconstruction, boating, involving myself in the affairs of the tiny community in which I live, and spending time with my beloved partner, Beverley. I've had a wonderful career, and I'm glad it's over: time to move on. *Dr Whittlesea recently retired from his position as a Full Professor in the Department of Psychology at Simon Fraser University, Canada.*

Helen L. Williams is currently a completing a post-doctoral position at the University of Richmond, Virginia, USA. Her research interests include higher order cognitive processes including both online metacognitive judgements and memory self-efficacy in younger and older adults, aging, and autobiographical memory.

Part I

Introduction

1
The Importance of Untangling Subjective Experience and Objective Measures for Understanding Memory Performance

Michelle M. Arnold

Introduction

Bruce Whittlesea has played a large role in shaping our understanding of memory processes, and due to his prolific research it is likely that various researchers would emphasize different aspects of his theories as his most significant contribution. It is hard to sum up concisely how he has influenced my own theoretical framework of memory, but if I had to focus on a specific topic it would be his hypotheses on the production and evaluation of performance. More specifically, throughout his research Whittlesea has argued that subjective phenomenology is not merely a 'by-product' of the production of past experiences; that is, that objective properties of performance (e.g., fluency) can lead both to predictable and unpredictable subjective experience. Further, Whittlesea has strived to demonstrate that remembering in general must be an inferential process because the 'mental content' found in veridical recollection can also be found in an imagined event (Leboe & Whittlesea, 2002; Whittlesea, 2002, 2004). This chapter will focus on how the uncoupling of objective and subjective performance, together with the characterization of memory as the end product of an attributional process, provides a more precise understanding of the underlying mechanisms of memory.

Objective memory performance refers to any measure of memory that is experimenter-defined, and therefore independently quantifiable. For example, in a recognition task it is the experimenter who has pre-determined which of the test items are 'old' (i.e., which items the participants studied), and therefore an objective measure of memory performance can be calculated simply by comparing participants' responses on each test trial to the known old/new status of the test item. Conversely, subjective memory performance refers to measures that are geared towards determining internal states of remembering – that is, how the act of remembering *felt* for individuals (e.g., how confident participants are in their old/new recognition judgements). It is important to distinguish between objective and subjective memory performance because researchers sometimes draw conclusions regarding the accompanying subjective phenomenology of an effect based solely on an objective measure (e.g., Sanna et al., 2002). However, the pattern of performance on an objective test of memory does not necessarily correspond to a specific accompanying subjective phenomenology: A high level of recall on a memory test does not automatically translate into a high level of confidence, or a strong feeling of recollection or familiarity for those items.

Acknowledging that we cannot make assumptions about subjective experience based on objective performance (or vice-versa) is essential, but in fact there are several important reasons for why untangling subjective and objective measures may help us better understand memory in general. The current chapter concentrates on two specific reasons regarding why it is valuable to consider the separate contributions of objective measures and subjective phenomenology to overall memory performance. The first section focuses on how differing combinations of objective and subjective performance help to inform our understanding of the underlying mechanisms of memory. Specifically, the main interest in this section is what the variations of phenomenology in a single experiment for the same level of objective memory indicates about basic memory processes. The second section explores the related issue of differences in subjective phenomenology across materials, experiments, etcetera, for a given objective measure; that is, whether these differences in subjective experience may be indicating that more than one phenomenon is being investigated under the same label.

The role of objective and subjective measures in uncovering the underlying mechanisms of memory

There are numerous ways to measure subjective phenomenology, but two of the more common techniques are confidence ratings and Remember/Know (R/K) judgements. In a typical R/K judgement, participants are instructed that for any items they claim are 'old' (e.g., were presented on a study list) they must indicate whether they can bring to mind specific details of studying the items (R), or whether the items simply feel old/familiar but no conscious details of previously encountering the items come to mind (K). Although confidence ratings are a more straightforward judgement for participants to complete (i.e., they require less instructions, are argued to be more intuitive for participants, etc.) many researchers use R/K judgements to attempt to tease apart the underlying mechanisms that contribute to memory performance.[1] Therefore, although confidence ratings also provide important information regarding the relationship between objective and subjective measures of memory (e.g., confidence-accuracy calibration; Weber & Brewer, 2003), this section focuses on research that has implemented R/K judgements as the measure of subjective phenomenology.

The history of research utilizing R/K judgements predominantly has been rooted in exploring two general types of memory theories: quantitative versus qualitative. The R/K models under the qualitative umbrella emphasize the idea that remembering is the result of two distinct processes that give rise to different types of subjective experience: namely, recollection and familiarity. However, qualitative approaches typically differ in how they define the nature of the underlying structures responsible for recollection and familiarity. For example, in a standard R/K paradigm, some researchers interpret the R option as a measure of recollection and the K option as an index of familiarity (e.g., Gardiner et al., 1996). Conversely, researchers such as Jacoby and colleagues (e.g., Jacoby et al., 1997) have argued that the K should not be taken as a straightforward measure of familiarity because R responses displace K responses when recollection and familiarity co-occur: An individual who believes that an event is old will only choose K if s/he is unable to recollect specific details of this prior event. Additionally, the equations for estimating recollection and familiarity in Jacoby's (1991) *dual-process model* rest upon the assumption that conscious (recollection) and unconscious (familiarity) processing are independent of one another; that is, conscious and unconscious processing can occur either in isolation or together.

Quantitative approaches to R/K data specify that the difference between remembering and knowing is dependent on the decisional processes; both judgements are based on the same memory traces (i.e., the same information), and they simply reflect differences such as trace strength (e.g., Donaldson, 1996; cf. Dunn, 2004). Similar to qualitative models of R/K judgements, quantitative approaches differ in how they define the decisional processes that lead to an R or K response. For example, a classic quantitative interpretation of R/K data is that K responses in a recognition task represent the divide between judging items to be 'old/new,' whereas the R responses correspond to the high confidence 'old' judgements (Donaldson, 1996). Conversely, Rotello, Macmillan, and Reeder (2004) argued that, although recollection and familiarity are not independent processes, two dimensions are required to model recognition data; one dimension is responsible for producing the overall 'old/new' recognition judgements, and the second dimension distinguishes between R and K experiences.

Not all theories of R/K data fit neatly under a quantitative versus qualitative approach distinction. For example, the *distinctiveness-fluency model* (Rajaram, 1996; Rajaram & Geraci, 2000) maintains that the key issue is not whether it is a single continuum or dual-processes that is responsible for R and K responses, but rather that non-distinctive fluent processing gives rise to familiarity (K) and distinctive processing (e.g., level-of-processing manipulations at encoding) produces recollection (R). Further, two recent proposed memory theories – the *functional account* and the *expectancy heuristic account* – combine aspects of both qualitative and quantitative approaches to R/K data. In particular, both theories argue that certain patterns of R and K performance demonstrate that recollection is inferential in nature, and therefore that an R experience is decisional- and context-dependent.

According to the functional account of recollection, whether an event is judged as an R or K experience depends on the context in which the decision is made: If the information that comes to mind about that particular event allows you to make the decision-at-hand then you will experience a feeling of conscious recollection (Bodner & Lindsay 2003; Gruppuso et al., 1997). Specifically, being able to bring details to mind regarding an event will not lead you automatically to judge that you are consciously recollecting that event because if the details are not source-specifying then the information is only strong enough to support a feeling of familiarity. An updated twist on Mandler's (1980) classic butcher-on-the-bus example can be used to illustrate this distinction: You see a woman walking towards you on the street who has a nose piercing and spiky purple hair and you want to determine if this is a person you have encountered before. As the woman walks past you, the combination of her nose piercing and hair brings to mind the thought 'oh, when I saw her before she reminded me of my younger sister because they have almost identical piercings and hair styles.' If your goal simply was to decide between whether this woman was a stranger or someone you have seen before, then likely you would claim that you recollect (R) encountering this person at least once in the past (i.e., that you remember seeing her before because you recalled that she had reminded you of your sister). However, if your goal was more stringent and instead you were trying to determine not just if, but *how* you know this person (i.e., in what capacity you have seen her before) then that same detail coming to mind at best would make her feel familiar (K) to you because it is not the type of detail that allows you to accomplish the task.

Although the above example is hypothetical, there is experimental evidence that supports the functional approach to memory (Bodner & Lindsay, 2003; Bodner & Richardson-Champion, 2007; Gruppuso et al., 1997; Kurilla & Westerman, 2008). For example, Bodner and Lindsay (2003) had participants study two lists of words; one list was always studied with a medium level-of-processing (LOP) task, and the other list was studied with either a shallow or a deep LOP task. The results showed that medium LOP words received significantly more R responses when they were studied and tested with shallow LOP words than when they were studied and tested with deep LOP words. The researchers argued that this effect was not due to differences between the conditions in how much list-related information could be brought to mind because participants were equally able to recall list source in the medium-with-shallow condition as in the medium-with-deep condition (Experiment 4). Instead, Bodner and Lindsay claimed that the context of the test list led the participants in the medium-with-shallow condition to use different attributes of the stimuli to define the properties of R and K than participants in the medium-with-deep condition. Evidence for the importance of the test list context was demonstrated further by the finding that the effect of higher R judgements in the medium-with-shallow condition disappeared when only medium LOP items were included on the test (Experiment 2).

In a similar vein to the functional approach, the expectancy heuristic account emphasizes that R/K judgements are inferential and the result of an attributional process. However, the expectancy heuristic approach goes a step further by explicitly claiming that it is an individual's *expectation* regarding the memorability of a situation that drives the decisional processes (McCabe & Balota, 2007). Specifically, according to this approach individuals have an expected level of memorability in any given situation (based on factors such as the context in which an item is studied/tested) and therefore items that pass this expected level will be labelled an R experience. Similar to Bodner and Lindsay's (2003) findings, McCabe and Balota (2007) showed that medium frequency words received more R responses when they were studied and tested with high frequency than low frequency words. To explain these results, McCabe and Balota argued that the average expected level of memorability was lower for participants in the medium-with-high-frequency condition than in the medium-with-low-frequency condition (i.e., because high frequency words are less memorable than low frequency words) and therefore more medium frequency items passed the expected threshold in this condition.

The functional and expectancy heuristic accounts may differ somewhat in how they explain R/K data, but as described in the preceding paragraphs, both approaches attempt to revise our understanding of the underlying mechanisms of memory by providing examples of situations where objective measures of memory remain constant, but subjective phenomenology changes according to the context.[2] Specifically, situations in which different conditions produce the same level of objective performance but very different subjective experience pose problems for many of the qualitative and quantitative memory theories. For example, it is not clear how independent recollection and familiarity processes (i.e., dual-process models) would be able to explain the differing levels of R responses for the same level of objective memory performance. Specifically, if the same type/amount of information for medium LOP is available at test for both conditions, then if recollection is independent of familiarity why would the presence of shallow versus deep LOP items lead to significantly different subjective phenomenology for the medium items? At a minimum, differing levels of subjective phenomenology for the same levels of memory performance indicates that recollection and familiarity are not static and rigid states that map directly to separate underlying components (for further discussion see Bodner & Lindsay, 2003; McCabe & Balota, 2007). Indeed, the data are a strong reminder that objective and subjective measures are not necessarily bound together, but rather they are more fluid and sometimes may combine in unexpected ways, depending on the context under investigation (although this is not meant to imply that they cannot be influenced in a similar manner under certain conditions).

The results available from the studies designed to test the functional and expectancy heuristic approaches to memory also lend support for the more general theory that memory is the result of an attributional process (e.g., Arnold & Lindsay, 2002; Whittlesea, 2004). For example, the *discrepancy-attribution hypothesis* argues that the feeling of familiarity is a consequence of the perception of discrepancy; that is, individuals can detect differences between how they expect to perform on a stimulus and how they actually perform and if their actual performance is more fluent than expected (i.e., 'surprising fluency') they may attribute this fluency to some source in the past. However, the attribution of the perception of discrepancy to either a source in the past or present depends on a multitude of factors (e.g., prior knowledge of the stimuli, the present conditions/context), and therefore is heavily influenced by principles such as encoding specificity and transfer-appropriate processing (Whittlesea, 2002). A key concept of the discrepancy-attribution hypothesis is that the feeling of familiarity is not an automatic result of discrepancy, but rather it is the result of an inferential process that is triggered by the detection of discrepancy.

Although not all researchers concur with the discrepancy-attribution hypothesis, many have proposed theories that do paint familiarity as the end result of some type of attributional process. In contrast, it is important to highlight that conscious recollection has not typically been defined as inferential in nature (e.g., Jacoby et al., 1997; Gardiner et al., 2002). However, the differing levels of subjective experience accompanying the same objective memory performance that has been found by researchers such as Bodner and Lindsay (2003) and McCabe and Balota (2007) indicate that, at least under some manipulations, decisions regarding the presence or absence of conscious recollection are also the end product of an attributional process. Although this idea is not novel – researchers such as Leboe and Whittlesea (2002; Whittlesea, 2002, 2004) previously have argued that familiarity and recollection arise from the same attributional processes – the predominant view within cognitive psychology continues to be that different process (i.e., either qualitatively different, or different in strength/ number of dimensions) are responsible for producing the subjective states of recollection and familiarity. Nonetheless, the data presented to support the functional and expectancy heuristic approaches are a strong reminder that the conceptualization of recollection as an inferential process warrants more attention than it currently receives in the memory literature.

Uncovering that similar levels of memory performance within the same experiment can have different accompanying subjective phenomenology would not have been possible if only objective memory had been measured (Bodner & Richardson-Champion, 2007). However, all of the data supporting the functional and expectancy heuristic approaches to memory come from recognition tasks; consequently, we need to consider whether we currently are missing important effects when it comes to recall tasks. More specifically, there is a large body of research showing that free recall involves an output monitoring process that determines what items should be reported (e.g., determines whether an item is 'old' or 'new,' whether the item has already been reported, etc.; Koriat et al., 1988, see also Higham, this volume, Chapter 9, for detailed discussion). Although Bodner and Lindsay (2003) found that participants equally were able to recognize studied medium LOP items in the medium-with-shallow and medium-with-deep conditions, would the same conclusions have been drawn if the task had been free recall? For example, if participants in a free recall task had been instructed only to report items they *remembered* from the study phase, then the results may have shown that their memory appeared better (i.e., higher output of studied words) for medium words in the medium-with-shallow condition.

It is not necessarily the case that results from a recognition task (e.g., more R responses for medium words in the medium-with-shallow condition) would directly map onto a free recall task (e.g., more output of medium words in the medium-with-shallow condition). However, if both recollection and familiarity are the result of inferential processes then one hypothesis is that a manipulation that leads to differences in subjective experience on a recognition task may also lead to differences in the level of output on a free recall task (i.e., due to output monitoring processes). Further, because Bodner and Lindsay (2003) were able to show that it was not a difference in the amount of available information that led to the different subjective phenomenology in their recognition tasks (but rather how this information was used by individuals), it is important to look at the patterns of performance that would be produced if the experiments were replicated with a free recall task.[3] Specifically, if recollection is the result of an inferential decision process then in a free recall task where the amount of available information between different contexts is constant, the level of output should depend at least to some extent on how subjective phenomenology impacts the output monitoring processes. Therefore, further research is necessary to examine the cause-and-effect relationship between phenomenology, monitoring, and memory performance (see also Koriat et al., 2006).

Using subjective phenomenology to help distinguish between effects

The supporting evidence for the functional and expectancy heuristic approaches to memory comes from comparing patterns of objective and subjective measures of memory within-experiments. However, in many situations researchers are required to draw conclusions about the underlying mechanisms of a given effect by comparing data across different experiments; that is, a compelling theory must be able to account for patterns of data that have been collected in different laboratories, with varying stimuli, different methodology, etcetera. Comparing across experiments can lead to difficulties for producing an integrated theory because it is not always clear why some effects are not consistently found or, perhaps even more challenging, why the same effect seems to occur under very diverse situations. In some instances producing a theory that encompasses the majority of experimental findings may not be possible because researchers are trying to explain what they believe is a single phenomenon, when in fact there are two (or more) similar effects under investigation. Unfortunately, it can be difficult to separate out related effects when they historically have been studied as a single phenomenon, but comparing combinations of objective and subjective measures across experimental settings can provide useful clues for establishing boundaries between related phenomena.

Hindsight bias (also commonly referred to as the knew-it-all-along effect) is just one example of a well-known effect that has been difficult to explain with a single unified theory. A hindsight bias occurs when individuals report that they had previously known something that they in fact learned only recently (i.e., after being exposed to correct feedback). In general, most studies of hindsight bias use one of two paradigms; a memory versus hypothetical design. In a memory design the effects of feedback are determined within-subject by having participants complete the same set of judgements twice – once before and once after exposure to correct feedback (e.g., Fischhoff & Beyth, 1975). Conversely, in a hypothetical design the judgements made in the presence of feedback for one group (hindsight group) are compared to judgements of the same stimuli made by a second group who have not been exposed to the answers (foresight group; Fischhoff and Beyth, 1975). Various explanations have been proposed to explain hindsight bias (e.g., automatic assimilation, availability and anchoring heuristics, attributional processes), but no single theory has been able to encompass the numerous and diverse research findings (Arnold & Lindsay, 2007; Blank et al., 2008).

Blank et al. (2008) presented several experiments to support the claim that a major reason it has been difficult to develop a cohesive theory of hindsight bias is that there actually are three hindsight effects: two effects involve the foreseeability and necessity of an answer/outcome, and the third component involves memory distortions. Due to the major differences between hypothetical and memory hindsight paradigms, it likely is essential to parcel out memory issues from other contributing components; for example, it is reasonable to presume that memory distortions play a larger role in memory designs than hypothetical designs (although Blank et al., 2008, contend that the three components are not necessarily found in isolation). However, the separation of the hindsight bias into three distinct components may not go far enough, in that the components themselves may need to be further divided. For instance, Arnold and Lindsay (2007) demonstrated that, even across similar memory design paradigms, measures of subjective experience varied significantly depending on the class of stimuli used to measure the effect.

Because hindsight bias commonly has been referred to as a *feeling* of having known some piece of information in foresight, Arnold and Lindsay (2007) set out to measure the subjective experience of the effect by adding a Remember/Just Know/Guess (R/JK/G) judgement to a memory design. Their first set of experiments focused on the standard memory design stimuli of difficult general knowledge questions, and the R/JK/G judgement was inserted into both a traditional (number scale) and a modified-traditional (2-alternative-forced-choice) paradigm (see Figure 1.1A). A hindsight bias was found in both experiments, but there was no evidence that participants had a feeling of knowing the feedback information in foresight; that is, participants overwhelmingly claimed that they simply were guessing they had previously given the correct answers to the feedback questions. In a series of follow-up experiments the general knowledge questions were replaced with word puzzles, which could be rearranged or solved to form common words, phrases, or clichés (e.g., 'once' appearing above '4:56 pm' can be solved for 'once upon a time'; see Figure 1.1B).

Again, a typical hindsight bias was found in the experiments but, unlike with the general knowledge stimuli, the word puzzles produced an accompanying subjective experience. Specifically, for the puzzles that showed a hindsight bias (i.e., puzzles that participants had switched to the correct solutions after receiving feedback) participants were significantly more likely to claim they remembered giving the correct solutions prior to feedback, rather than just knowing or guessing they had previously provided the correct solutions.

Arnold and Lindsay (2007) argued that, due to the inherent qualities of the general knowledge questions and word puzzles, participants interacted with the two types of stimuli in qualitatively different ways. A key difference between the questions and word puzzles is that the puzzles allowed for a more rich interaction: Even when participants did not immediately know the solution to a word puzzle they could attempt to use various strategies and techniques to arrive at the correct solution. However, difficult general knowledge questions tend to be 'either-or,' in that participants either already know the correct answer, or they do not know the answer and have no avenue within the experiment for working out the correct response. Further,

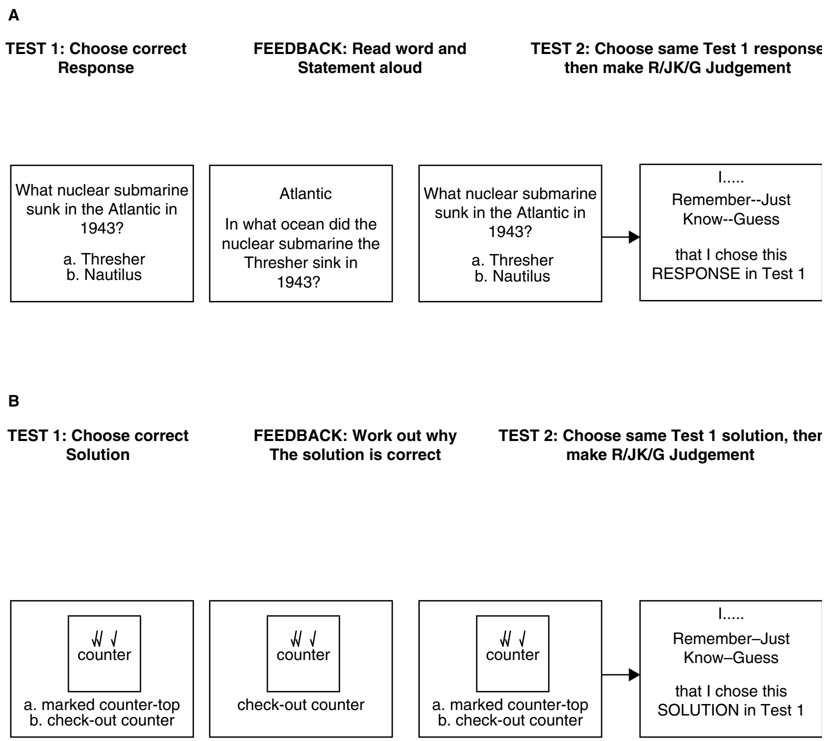

Figure 1.1 An example of a 2AFC hindsight bias trial across the three experimental phases for both the general knowledge questions (1.1A) and word puzzles (1.1B; see Arnold & Lindsay, 2007, for more detailed methodology).

the feedback phase of the experiments also allowed for a deeper interaction with the word puzzles and their solutions. That is, seeing the solution to a word puzzle affords the opportunity to work out why that solution is in fact the correct answer (e.g., 'oh I get it, "4:56 pm" represents the general concept of time, so "once" is on top of "a time"'), whereas the answers to difficult general knowledge questions do not lend themselves to the same processes (and thus likely feel somewhat arbitrary). Therefore, although both types of stimuli showed similar levels of hindsight bias, the differences in subjective phenomenology may be hinting that different mechanisms are responsible for the overall effects. Specifically, the hindsight bias found for puzzles may be the result of a memory distortion that is due to factors such as source-monitoring errors; for example, you may be able to bring to mind how you worked out the steps to the solution of a puzzle when you were shown the correct solution during feedback, but you misattribute this to when you originally were asked to solve the puzzle (see Lindsay, 2008, for a more detailed discussion on source-monitoring errors). Conversely, the hindsight bias produced with general knowledge questions appears to be more similar to a general response bias: When participants are unable to remember how they originally responded to a question in foresight they are influenced by having seen the correct answer during feedback (but with no accompanying feeling of having known the answer in foresight).

Hindsight bias is only one situation where measuring the accompanying subjective experience may help delineate multiple effects. Another less prominent example is the revelation effect, which is the finding that participants are more likely to judge a recognition test item is old when the item is preceded by some task (e.g., unscramble *nescirof* before the test item *forensic*) than if the recognition judgement occurs in the absence of a preceding task (Watkins & Peynircioglu, 1990). The difficulty in explaining the revelation effect has come from the fact that it has been found both when the preceding task is related to the test item and when the task is unrelated (e.g., solve *giaetvan* for *forensic*; Westerman & Greene, 1996). Verde and Rotello (2004) implemented a confidence judgement in their recognition task and constructed ROC curves to demonstrate that the revelation effect is not in fact a single phenomenon. That is, the same-task revelation effect is caused by a decrease in memory sensitivity (i.e., a decrease in familiarity), whereas the different-task revelation effect is due to a change in response bias (i.e., more liberal responding; see also Hicks & Marsh, 1998). The results from other laboratories have supported Verde and Rotello's argument that the revelation effect is not a single phenomenon, and ensuing research has focused on fleshing out the mechanisms responsible for the two different components (e.g., Leynes et al., 2005; Major & Hockley, 2007).

Although comparing different combinations of subjective and objective performance may help delineate related memory effects, it is important to stress that finding differences in subjective experience for a given phenomenon is not necessarily an indication that more than one effect has inadvertently been lumped together under the same label. Indeed, as described in the first section of this chapter, many attributional approaches to memory explicitly argue that behaviour and subjective phenomenology are not fixed; the same objective performance may be experienced differently, depending on the current context (e.g., Arnold & Lindsay, 2002; Whittlesea, 2004). Rather, the main point of this section simply is that comparing the accompanying subjective phenomenology of an effect across different classes of stimuli, instructions, etcetera, may be a helpful tool in uncovering whether there are different underlying mechanisms leading to the observed objective performance.

Summary

Measuring both the subjective and objective components of an effect is important not only because we cannot make assumptions about one based on the other, but also because separate measurements lead to a clearer understanding in general about memory processes. One major problem with binding objective and subjective performance together (i.e., assuming that subjective experience will mirror objective performance) is that this mistaken assumption begins to compound as more and more research is conducted. For example, if hindsight bias researchers assume that participants really feel that they had known the feedback information in foresight, then any manipulation they find to modulate the hindsight bias effect likely also will be assumed to modulate subjective experience in the same manner. However, as researchers such as Whittlesea (2002, 2004) have emphasized, observed behaviour and subjective phenomenology can be found in both expected and surprising combinations, and therefore one important route to understanding memory is to uncover these various combinations. Further, as discussed in the first section of the chapter, some experimental manipulations appear to have little or no effect on objective memory measures, yet they have a significant impact on phenomenology. Therefore, subtle but important effects are easy to miss when the quality of the accompanying subjective experience is not evaluated (Bodner & Richardson-Champion, 2007). Although the various possible combinations of objective and subjective memory performance are not necessarily straightforward to interpret, these patterns help inform us about the nature of memory mechanisms.

Notes

1. There is a large body of work devoted specifically to examining whether confidence ratings and R/K judgements reflect qualitatively different measurements, or whether R/K judgements are just another (more complicated) measure of confidence (e.g., that R responses simply reflect high confidence; cf., Dunn, 2004; Rotello et al., 2004). Although this is an important issue, it is beyond the scope of the present chapter.
2. It is important to point out that not all of the experiments used to support the functional and expectancy heuristic approaches have demonstrated equal hit rates across conditions. However, differences in false alarm rates also were found across the conditions, and closer inspection of the data have shown that objective measures of discrimination (e.g., d') do not differ between the conditions (see Bodner & Richardson-Champion, 2007).
3. One obvious issue with using free recall is that it inherently makes it more difficult to manipulate test list context. However, as long as participants are able to recall at least some items from the manipulated categories (e.g., at least some items from both medium and shallow studied words) then at least a subtle context is present at test.

Part II

Inferential Processes and Fluency/Familiarity

2

Fluency and Familiarity: How Memory for Perceptual Detail Influences the Remembering of Events

Karen M. Evans and Aaron S. Benjamin

> The more I think about that seam between the familiar and the unfamiliar – and how it feels to pass from one to the other – the clearer it becomes that humans instinctively generate a sense of familiarity. You can sense it for yourself the next time you drive someplace you've never been before. Somehow, it always feels as though it takes longer to get there than it does to get back home again. It's as if there's a principle of relativity, a bending of time, in the very concept of familiarity. The road we know is always shorter than the road we don't know – even if the distances are the same (Klinkenborg, 2009).

Introduction

Recognizing events, objects, and persons from our past is a task fraught with significance. It is embarrassing to not remember someone's name, but the more socially adept among us can navigate such a situation delicately and perhaps even slyly elicit the sought-after name. Not recognizing a face as a familiar one, or misattributing that face to an incorrect prior encounter, is a failure from which we can not recover quite so inconsequentially.

There is a large and increasingly unwieldy literature on mnemonic sources of information in recognition (e.g., Wixted, 2007; Parks & Yonelinas, 2007) and on the decision processes underlying recognition judgements (e.g., Benjamin & Bawa, 2004; Benjamin et al., 2009; Stretch & Wixted, 1998). This chapter takes as a starting point the view that (at least) one mnemonic source of information can be characterized as the *familiarity* of a stimulus and that that familiarity is at least in part due to memory for prior perceptual experiences and the overlap of that memory with the current perceptual experience (e.g., Benjamin et al., 1998; Jacoby, 1983a; Whittlesea & Williams, 1998). It is this latter point that is the focus of the current chapter, in which we review evidence on the relationship between perceptual memory and recognition judgements. How does our notably poor memory for exact perceptual detail support feelings of familiarity and judgements of recognition (cf. Matzen & Benjamin, 2009; Matzen et al., in press)? If we can't remember the route, why would it seem to take longer to go somewhere than to get back home?

Memory for perceptual detail

A general and quite revealing finding in the memory literature is that items are processed more easily (i.e., with greater fluency) upon repetition (Feustel et al., 1983; Jacoby & Dallas, 1981; Scarborough et a., 1977). This rather ubiquitous effect underlies many indirect measures of memory, such as reductions in the time it takes to name a perceptually degraded word or to identify it at all, and may also contribute to judgements that are made during direct tests of memory. The facilitated processing of repeated items (i.e., repetition effects) may be rooted in different sources, including conceptual priming, but the importance of *perceptual* priming is demonstrated by the fact that changes in physical form across repetitions either dampen (Feustel et al., 1983; Roediger & Blaxton, 1987) or obliterate (Jacoby & Dallas, 1981) repetition effects, and that non-words, which cannot easily engender conceptual processing, nonetheless elicit robust facilitation effects (Feustel et al., 1983; Johnston et al., 1985; Whittlesea & Williams, 2000).

The claim that memory for perceptual detail supports recognition judgements violates the widely held assumption that our memory for perceptual details fades rapidly. Indeed, we seem to encounter numerous confirmations of this intuition (e.g., an inability to recall the exact wording of a recent email or to retrieve what the stranger in the elevator this morning looked like), and may even have the sense that there is little need to remember this information. Still, even when unable to reconstruct the details of a prior experience, we are often confronted with a strong sensation of familiarity when we encounter that same item again. In fact, the inability to readily retrieve information about a prior encounter may strengthen the role of perceptual overlap, as the surprise of fluency in such situations demands an explanation (Whittlesea & Williams, 1998; 2000). The facile processing of a repeated item can provide a 'fluency heuristic' to influence judgements of recognition memory. Research addressing the relationship between subjective senses and judgements about objective states of the world owes a great debt to the always innovative and pioneering work of Bruce Whittlesea, and we are pleased to present this brief review in the context of a volume dedicated to his career.

Perceptual fluency and recognition judgements

Before reviewing this literature, it is of use to highlight two maxims of the fluency heuristic that provide a framework for interpreting the following data, especially where null effects are observed. (1) The application of a fluency heuristic to recognition judgements is often a last resort relied upon when other sources of information (e.g., recollection) are not available. Thus, even if fluency cues from perceptual priming are available, they are only sometimes used to inform recognition judgements. (2) Use of a fluency heuristic assumes an attribution process by which facilitated perception is attributed to a task-relevant goal, such as prior exposure in a recognition task; this process is fallible, however, as fluency can be misattributed when the true source of fluency does not match the observer's goal. Note that we are not the first to point out these themes, as the following review will clarify.

Relationships between measures of fluency and recognition

Jacoby and colleagues have argued that perceptual priming and recognition memory are both classes of episodic memory, and that the degree to which performance on these two test types parallel one another is determined primarily by the specific retrieval demands of each task (Jacoby, 1983a, 1983b; Jacoby & Witherspoon, 1982). This is primarily based on early evidence that performance on perceptual tasks (usually a perceptual identification test in which degraded visual words are gradually clarified, and the time at which participants are able to identify the word is recorded) and recognition tests (old/new judgements to repeated and novel test words) alike is sensitive to manipulations that obscure or enhance access to the initial episodic trace. In particular, the magnitude of perceptual priming and hit rate associated with recognition are enhanced when study items are presented multiple times (Jacoby & Dallas, 1981), repetitions during study are spaced rather than massed (Jacoby & Dallas, 1981), higher old–new ratios are employed during test (Jacoby, 1983a), and the length of the retention interval is shorter (Jacoby, 1983a). That measures of perceptual priming and recognition often correlate has been taken as evidence that performance on both perceptual and recognition tests reflects the operation of a common episodic memory system, and that people can heuristically use the fluency of perceptual processing as evidence that an item is repeated (e.g., Jacoby & Dallas, 1981).

Importantly, however, these correlations are not always observed, and such dissociations have been leveraged in support of alternate accounts that priming and recognition operate through separate mechanisms (semantic and episodic memory, respectively) and cannot influence one another (e.g., Wagner & Gabrieli, 1998). Specifically, the amount of observed perceptual facilitation is not necessarily dependent on recognition (i.e., it is sometimes equal for repeated words that are remembered and for those that are forgotten: Jacoby & Dallas, 1981; Jacoby & Witherspoon, 1982), nor is successful recognition contingent on perceptual facilitation (Jacoby & Witherspoon, 1982). In general, dissociations between performance on these two measures are observed when encoding conditions promote deeper conceptual processing or semantic elaboration, through the use of generation tasks (Jacoby, 1983b; Jacoby & Dallas, 1981), deep encoding (Jacoby & Dallas, 1981), or increased study time (which is arguably used to enhance elaborative encoding: Jacoby & Dallas, 1981). Levels of processing manipulations have even revealed that encoding depth has opposing effects on perception and recognition, with facilitation on perceptual tasks being greater for more shallow, data-driven encoding (e.g., a word presented in isolation) and recognition rates being higher under deeper encoding conditions (e.g., words generated in an antonym task: Jacoby, 1983b). This pattern of sensitivity to episodic details such as encoding conditions requires that explanations of perceptual facilitation admit properties of episodic memory, because mere activation of decontextualized lexical representations (i.e., semantic memory) cannot account for such effects.

Although it is informative to investigate correlations between perceptual facilitation and recognition memory, later designs sought more direct evidence of the use of fluency heuristics. Rather than measuring the relationship between separate blocks of perceptual identification and recognition, Johnston et al. (1985) followed each perceptual identification trial with an immediate recognition judgement for the same word (after a separate block of study words). This sequential judgement paradigm provides participants with a readily accessible fluency cue (i.e., the ease of the preceding identification) at the time of the recognition judgement. It also provides experimenters with a measure of item fluency (identification time) for both repeated and unstudied test words. In this design, use of a fluency heuristic is inferred if items that are rapidly identified are more likely to be judged as old; of particular interest is an examination of error trials (misses and false alarms), as attributing perceptual fluency to repetition status may cause these incorrect classifications. Critically, Johnston et al. found not only that repeated words were identified more rapidly (thus observing typical perceptual fluency effects), but that words that were *judged* as old were identified faster than those judged as new (i.e., hits were faster than misses, and false alarms were faster than correct rejections). Johnston et al. additionally found that pronounceable non-words that were rapidly identified were more likely to be called old *regardless* of their actual status (i.e., hits and false alarms were identified faster than misses and correct rejections). The authors attributed the greater role of fluency cues in judging non-words to the reduced availability of elaborative encoding for study stimuli with no semantic meaning. These results provide support for the first maxim, demonstrating that fluency cues appear to be more important under conditions in which other bases for the recognition decision, such as recollection, were reduced.

Johnston, Hawley, and Elliott (1991) further established the inverse roles of fluency cues and recollection. Across several experiments, the degree of elaborative encoding was manipulated by having subjects name study words, count vowels in study words, or view a sham study phase in which no words were actually presented but participants were told that words were being presented subliminally. During the test phase, participants completed sequential perceptual identification (naming a word as rapidly as possible as a mask slowly disappeared) and recognition judgements (as in the Johnston et al., 1985, design). When encoding conditions provided the least support for test-phase recollection (by presenting no study words to be recollected), the likelihood of calling an item old increased as the speed of identification increased, suggesting a strong reliance on fluency cues. When encoding conditions provided the strongest support for recollection (verbal production), there was no relationship between perceptual fluency and recognition, suggesting that recognition judgements were primarily informed by explicit retrieval mechanisms. That evidence for applying a fluency heuristic was absent for words studied in the production task is particularly striking given that, across the three encoding conditions, repetition effects in the identification task were actually greatest for the production group! Thus, despite the fact that repetition strongly affected identification speed, participants did not employ this heuristic to any observable degree. This contrast highlights the important fact that the cue validity of a fluency heuristic is less important in determining its contribution to recognition than is the presence of alternative sources of information (first maxim). Accordingly, the mere presence of perceptual facilitation does not ensure that this information is used to bias recognition judgements; thus, although the studies discussed earlier found correlations between perceptual identification tasks and recognition tasks, item-level analyses of sequential judgements are necessary to examine the use of fluency information during recognition judgements.

An important finding in Johnston et al.'s (1991) was revealed when they compared the use of fluency heuristics in sequential judgement conditions (as described above) to that in blocks judgements (i.e., a perceptual identification block, and then a recognition block). When recognition judgements were performed in a separate block from the perceptual identification of the same words, there was no relationship between fluency and recognition in any encoding condition. Such a finding is important in validating the use of a fluency heuristic. An alternative explanation is that the fluently processed words might be more easily recognizable due to some other stimulus characteristic (e.g., perhaps the shortest words are both easy to read and easy to recognize), but an item-selection account (Watkins & Gibson, 1988) would predict parallel effects across the mixed and blocked conditions (see also Higham & Vokey, 2000, for counterevidence to item selection).

Whittlesea and Leboe (2003) also examined recognition responses based on the fluency with which test words were processed, by performing a median-split on fluency measures (naming latencies to test words). If more fluent processing of test words biases subsequent judgements, then faster named words should be associated with higher claims of recognition than those that are named slowly. Whittlesea and Leboe found that this was true for pronounceable non-words but not for meaningful stimuli, suggesting that additional sources of evidence were available when an item was familiar. Although Whittlesea and Leboe did not assume that this other mnemonic source was recollection (but rather, a different form of fluency, as discussed at the end of this chapter), these results echo the first maxim in finding that reliance on fluency heuristic is sensitive (and generally, inversely related) to the availability of other cues.

Kelley, Jacoby, and Hollingshead (1989) discovered that perceptual fluency can also bias judgements of source recognition. After studying a mixture of visual and auditory words, participants completed a test phase in which words were presented for perceptual identification (shown at a single brief duration between visual noise masks), and then presented in full view and tested for both source and oldness simultaneously (i.e., participants classified a word as read, heard, or new). In the perception task, studied words were more likely to be identified than new words, and seen words were more likely to be identified than heard words. Seen words were also more likely to be remembered (i.e., not called 'new') than heard words. Hence, modality effects were present in both the identification and recognition tasks. The source judgements made to false alarms (which, in actuality, had no study-phase source) were particularly revealing with respect to the use of a fluency heuristic. For new words that were incorrectly judged as old, participants were more likely to call the word 'seen' if it had been identified successfully in the preceding perceptual task, and more likely to call it 'heard' if it had not been identified. The authors interpreted this effect as resulting from the application of a fluency heuristic: when words are easily processed, participants attribute this fluency to having encountered the item in the same source. In a second experiment, participants were provided with a mnemonic strategy to help them remember modality (they were told to think of positive associations for seen words and negative associations for heard words), and this greatly reduced the bias to label false alarms as 'seen.' As in the case of old/new recognition, therefore, the first maxim extends to source recognition, as source judgements are more likely to rely on fluency heuristics in the absence of alternative sources of information.

Experimental manipulations of perceptual fluency

Perhaps the most compelling evidence that perceptual fluency can be recruited heuristically during recognition is the ability to induce a sense of familiarity by experimentally manipulating the perceptual clarity of test items. In both visual and auditory modalities, subtle manipulations of perceptual noise levels at test have been shown to promote higher rates of 'old' judgements for words presented in low noise backgrounds, relative to those in high noise backgrounds (Goldinger et al., 1999; Whittlesea et al., 1990). In both of these studies, a single degraded test probe followed a short series of study words (seven words seen for 60 ms each in Whittlesea et al., 1990; eight words spoken at a normal rate in Goldinger et al., 1999), and the test probe was presented in light or heavier noise (though the difference was intended to be unnoticeable). For both repeated and unstudied words, 'old' judgements were higher in the light noise condition. Thus, even though the level of clarity was manipulated by the experimenters, independent of old–new status, participants appeared to use this fluency information in forming their recognition responses. By demonstrating that participants will attribute fluency not necessarily to the correct source (which in this case is perceptual noise levels) but to the source that the task renders most likely or salient, these results emphasize the goal sensitivity of fluency attributions stressed in the second maxim.

Several experimenters have attempted to elicit similar effects, but failed. For example, Johnston, Hawley, and Elliot (1991) followed a study list with a test phase of sequential identification and recognition tasks, in which the critical manipulation was the rate at which the visual mask disappeared during the identification task (rapid or slow). Study trials either involved naming the study word, counting vowels, or studying non-words. Across this range of encoding depths, there was no evidence that the mask removal rate biased recognition judgements. In this design, however, there was no attempt to conceal the manipulation, allowing the possibility that participants were aware of the rate changes and thus attributed the faster identification to faster mask removal. This highlights the importance of the second maxim: fluency effects are not always attributed to prior exposure, but can be attributed to other sources when they seem more likely.

In another experiment that failed to induce fluency attributions on recognition judgements, Watkins and Gibson (1988) had participants study a list of words and then complete a test phase in which identification judgements were followed by recognition judgements. The key manipulation was that during the identification task, some words were presented for longer durations than others. Neither with visual nor auditory presentation was this manipulation successful in soliciting a greater proportion of 'old' responses to long presentation items, despite strong priming effects of prior exposure on the identification task. Although the authors were careful in the visual presentation experiment to reduce the possibility that participants were aware that the duration of presentation was manipulated, they may have overlooked the first maxim of the fluency heuristic: namely, participants will only rely on fluency if they need to, and when other sources of information are available, those will likely be used instead. Therefore, it is possible that deeper processing reduced the contributions of fluency cues (as suggested by Whittlesea et al., 1990). Additionally, as noted by Higham and Vokey (2000), Watkins and Gibson's conclusion is based on null results obtained through a manipulation that may have been too weak to pose an adequate test (i.e., prime durations may not have differed enough across conditions).

Effects of preexposure to test words

One of the most revealing and well replicated manipulations of perceptual fluency is Jacoby and Whitehouse's (1989) use of subliminal 'context words' to facilitate test word processing. In this paradigm, participants view a long list of study words and then make recognition judgements to test words preceded by masked primes that match the following test word, mismatch the following test word, or are meaningless strings (e.g., xoxoxox). In Jacoby and Whitehouse's original test, participants were told either that primes sometimes matched the test words and should therefore be read in order to assist performance, or that the mask was simply a meaningless attention cue to signal the test word. For participants who were told to ignore the cues, new words were more likely to be judged as old when they were preceded by a matching prime than when preceded by a meaningless prime. Participants in this group were also less likely to judge new words preceded by a mismatching word as old, relative to the meaningless primes. For subjects who were aware that context words sometimes matched the target, the opposite pattern occurred, such that they were less likely to call new items old when they were preceded by a matching prime. Both groups experienced more fluent processing of test words that were preceded by a matching prime, but whereas participants who knew about this manipulation correctly attributed fluency to the prime word, those who were unaware of this manipulation used task goals to attribute fluency to prior exposure (demonstrating the second maxim). A similar pattern of results was found when the presentation duration of the prime was increased, suggesting that the supraliminal exposure caused subjects to be aware of the prime's presence and to discount it accordingly.

An alternative interpretation of fluency effects on recognition is provided by Huber, Clark, Curran, and Winkielman (2008). They generalized a model of perceptual identification (Huber & O'Reilly, 2003) to the recognition task of Jacoby and Whitehouse; the critical mechanism in that model is that priming first enhances fluency (by aiding perceptual mechanisms in a top-down manner) and, after longer exposure durations, decreases fluency (because of habituation). In this explanation, no attribution is necessary to explain the reversal of priming effects when the prime is presented for a longer duration. However, it is not clear that this explanation can easily accommodate the result that the effects of the prime vary with instructions to the subject, as reviewed above.

Subsequent work using this paradigm has found that the lack of awareness of primes is not necessary to the success of the manipulation (Joordens & Merikle, 19992; Gellatly et al., 1995), and in some cases awareness can strengthen the illusion (Higham & Vokey, 2000). Joordens and Merikle (1992) compared recognition following primes presented above perceptual identification threshold to those presented subliminally, and found that prime duration was sufficient to produce the Jacoby and Whitehouse (1989) illusion, independent of whether participants were told about the matching prime words, as predicted by Huber et al. (2008). Gellatly et al. (1995) found that when prime duration (for a stream of rapidly presented prime words) was held constant, the illusion could be selectively produced under instructions directing participants to encode the words, versus instructions directing them to monitor the stream for a word matching the subsequent recognition probe. Gellatly et al. concluded that the matching instructions did not produce the illusion because they rendered the match between prime and test probes salient, thus making the prime a stronger candidate for explaining the fluency (similarly, salience was manipulated by prime duration in Joordens and Merikle's study and by awareness in Jacoby and Whitehouse's study). Higham and Vokey (2000) proposed that the illusion itself is due to an identification heuristic in which participants attribute their ability to read a rapidly presented prime to prior exposure; awareness of the prime's relationship to the target thus motivates use of this heuristic. Long durations fail to produce this illusion because the prime identification is too easy to be influenced by prior exposure, thus making the identification heuristic less viable.

A recent series of experiments by Westerman and colleagues has extended Jacoby and Whitehouse's (1989) paradigm to a variety of form manipulations, in order to assess the role of expectancy in fluency heuristics. These studies have revealed that enhanced false alarm rates to fluently processed (primed) words can be prevented not only by providing a more likely explanation for the fluency effects (as in Jacoby & Whitehouse, 1989), but also by making perceptual fluency an unlikely explanation (i.e., reducing cue validity without presenting a better alternative). For example, Westerman, Lloyd, and Miller (2002) reduced the validity of fluency cues by having participants complete an auditory study list followed by a visual test list containing context words. In this paradigm, prior auditory processing would not be expected to facilitate visual processing at test, and so the sense of fluency produced by the matching primes should not be attributed to prior exposure. Supporting this hypothesis, matching primes did not enhance false alarm rates in the presence of a modality change. Westerman et al. also found that words studied aurally were more likely to be judged as 'old' when the study list also contained visual words (that match the test modality). Thus, participants' willingness to attribute enhanced fluency to prior exposure was sensitive to their expectations that the test words should be processed more fluently; when modality always changed between study and test, participants had no reason to expect more fluent processing, and thus did not attribute fluency to repetition. (A very similar pattern of results also obtained for more subtle, within-modality changes of words to pictures, and changes in font style: Westerman et al., 2003.) Additionally, Westerman et al. found that when given a sham 'subliminal' study list (as in Johnston et al., 1991), participants who viewed visual noise (that allegedly contained subliminal study words) exhibited greater fluency attributions to visually presented (and primed) test words than participants who heard auditory study noise. As in Johnston et al.'s study, these effects, in the absence of any memory signal to counter them, were greater than when there was a true study list.

Experimental manipulations of recognition

Demonstrations of the fallibility of fluency heuristics have not only examined the extent to which enhanced perceptual ease due to stimulus characteristics can be falsely attributed to prior exposure, but also the extent to which fluency resulting from prior exposure can be falsely attributed to perceptual characteristics. Witherspoon and Allan (1985) had participants view a list of words, and then (in a superficially unrelated task) evaluate the duration for which briefly presented words remained onscreen. Words that had been seen before were evaluated as remaining onscreen longer than new words, and this effect obtained whether participants were asked to name the words or not. Jacoby, Allan, Collins, and Larwill (1988) had participants listen to a series of sentences, and then rate the noise levels of a set of purportedly unrelated sentences. Participants rated repeated sentences as occurring in less auditory noise than new sentences, even though the noise levels were matched across stimulus classes. Similar effects have been found for single words presented aurally (Goldinger et al., 1999) and visually (Whittlesea et al., 1990). These studies underscore the importance of the second maxim: the use of a fluency heuristic in recognition memory and the presence of repetition-based perceptual fluency effects are not the same thing. Perceptual fluency can arise from a variety of sources, and can be attributed to a variety of sources, sometimes leading to an imperfect mapping (reviewed in Benjamin & Bjork, 1996; Jacoby et al., 1989).

The fluency criterion

Until this point, we have presented evidence that certain stimuli are perceived as 'more fluent' without providing the necessary qualifier: fluent relative to what? Jacoby proposed that fluency is evaluated relative to the difficulty of the current task (Jacoby & Dallas, 1981). By this rationale, performing certain operations (e.g., reading words) is associated with a general level of difficulty, and when an individual item is further processed more fluently than expected (Benjamin et al., 1998; Whittlesea & Williams, 1998), this deviation is attributed to repetition (or alternative sources that the experimental context renders plausible: Goldinger et al., 1999; Whittlesea et al., 1990). In contrast, Whittlesea and Leboe (2003) suggested that fluency can be judged in two ways. First, item fluency can be assessed relative to other items from the same stimulus class. This is similar to Jacoby's relative fluency, though Whittlesea and Leboe emphasize the comparison to items in the stimulus class (rather than items in the current task); because this class-wide fluency is a contextually invariant property, Whittlesea and Leboe label it 'absolute fluency'. The second type of fluency judgement they propose is assessed relative to the expected fluency for that particular item. This item-level expectation of fluency requires a history of experience processing that item and therefore is only applicable to familiar stimuli (i.e., non-words do not give rise to this type of fluency). Finally, Westerman (2008) proposed that fluency is compared not to the fluency of the task, the stimulus class, or the particular item, but to the fluency of all other items in the current context (i.e., other test probes). This is supported by evidence that illusions of familiarity (obtained through Jacoby and Whitehouse's, 1989, manipulation of perceptual priming and through Whittlesea's, 1993, manipulation of conceptual priming) weaken as the proportion of test items that are primed increases (Westerman, 2008), and that this effect holds for within- but not between-subject manipulations. It is unclear whether one, none, or all of these theories are correct, but the recent revival of interest in identifying the basis of fluency judgements holds promise for continued progression.

Summary

Human memory systems are highly fallible, and a premium is placed on the ability to adaptively respond to the particular demands of infinitely varying situations in which remembering is required and yet details are sparse (Benjamin, 2008). One important tool used to confront imperfect memory is the monitoring and interpretation of ongoing perceptual events. When we see ourselves engaging in more rapid perception than we expect, we ask: does this enhanced perception owe perhaps to a recent prior encounter with this stimulus? This chapter reviewed evidence that this process takes place and that the answer is often in the affirmative, particularly when the situation lends that attribution plausibility and no superior basis for a memory judgement is available. It is true that memory affects perception, as noted so poetically at the outset of this chapter. But it is no less true that perception affects memory, and that sometimes the road seems short because it is short, not because we have travelled it previously.

3

The Development of the Fluency Heuristic in Childhood: More Questions than Answers

Jeremy K. Miller and Marianne E. Lloyd

Introduction

Imagine the challenges that a 3-year-old child's memory faces compared to that of an adult: The world is less predictable and the vocabulary is less familiar. Typically developing 3-year-olds are just beginning to harness the vast power of human language, and 3-year-old's memory skills are only beginning to develop into their adult forms. For instance, children's working-memory capabilities have been demonstrated to increase as they mature (Siegel & Ryan, 1989). When compared with adult memory performance, young children demonstrate greater susceptibility to false memory formation in some circumstances (Ceci et al., 2007) and less susceptibility in others (Brainerd et al., 2008). Metacognitively, children are often less effective at generating and implementing helpful retrieval and encoding strategies than adults (Chi, 1978). In many ways, young children's memories are quite different from adult memories.

It is unsurprising then, that the available research examining competencies in memorial attribution and memory decision making demonstrate that these skills do not appear until later in childhood (for a review see Bjorklund et al., 2009). Currently, the available data examining this question is rather sparse. Understanding how children develop the skills required in order to use these complex response strategies is an important topic of study for memory researchers for several reasons. First, research examining the development of children's memory skills helps to address classic developmental questions such as the extent to which these memory response strategies may develop through the child's examination of their own memory's successes and failures or the extent to which retrieval is constrained by biology. Second, understanding how children develop advanced memory skills could help in the development of techniques and interventions designed to assist developmentally delayed children in generating more effective memory strategies. Finally, understanding the developmental trajectory of children's memory response strategies may yield important insights into the question of how, when, and where both children and adult participants use metamemory to guide their response strategies – an issue that has recently generated some theoretical debate (cf. Karpicke et al., 2008; Whittlesea et al., 2005). The goal of the present chapter is to review what is known about the development of one such process: the use of the fluency heuristic in recognition memory.

What does it take to use the fluency heuristic?

Fluency may be described as the speed or ease with which an item, person, event, or experience is processed. Two types of fluency have been discussed most frequently: perceptual and conceptual. Perceptual fluency refers to enhanced processing due to the physical characteristics of the item (e.g., easy-to-read font) whereas conceptual fluency is a product of semantic overlap (e.g., faster access to 'cat' after seeing 'dog'). There is quite a bit of evidence suggesting that people are more likely to claim to remember fluently processed test items, suggesting that participants are at least implicitly aware of the correlation between speedy mental processing and familiarity (e.g., Jacoby & Dallas, 1981; Whittlesea, 1993; for an alternative interpretation see Winkielman et al., 2003). Environmental stimuli that have been previously encoded and processed are easier to subsequently process than novel stimuli. When rememberers become aware that previously experienced stimuli tend to be quickly processed they use this knowledge to guide future memory decisions. Specifically, they develop a heuristic in which they are more likely to identify a fluently processed stimulus as a target relative to stimuli that are less fluently processed. In addition to helping recognize familiar items, over-reliance on this heuristic has been shown to result in memory illusions. During laboratory memory tests, participants who are exposed to stimuli that have been made artificially fluent are more likely to call these stimuli 'old,' regardless of whether they have been previously experienced. This artificial fluency can be perceptual (e.g., enhanced through the presentation of a masked word that matches the test item) or conceptual (e.g., preceding the test word with a predictive sentence stem).

The manner in which participants decide whether or not to attribute fluency as evidence of previous occurrence is not entirely straightforward. That is, high levels of fluency do not always lead to high levels of 'old' recognition decisions. Whittlesea and Williams (1998) addressed this point nicely by asking the question: 'Why do strangers feel familiar, but friends don't?' Despite the fact that the faces of our friends are no doubt processed quite fluently, we are not generally overcome by feelings of familiarity when we see them. However, unexpectedly encountering a long forgotten high school classmate in the supermarket may cause a tidal wave of familiar feelings as we attempt to 'put a name with the face.' Much recent research indicates that the explanation for this apparent paradox lies in the role of expectations (Westerman, 2008; Westerman et al., 2002; Whittlesea & Williams, 1998, 2001a, 2001b). When we arrive at a friends' house, we expect to be surrounded by familiar faces and we adjust our evaluation of the incoming fluency of our processing accordingly. Conversely, we do not expect to run into high school friends in the supermarket and the discrepancy between the observed levels of fluent processing and our low expectations result in a powerful feeling of familiarity.

As is clear from the above discussion, the use of perceptual fluency as a cue to memory is a complex process. Consequently, any analysis of the development of the fluency heuristic needs to examine the functional capabilities of children at various age levels to accomplish the various tasks necessary in order for the fluency heuristic to function in a mature manner. Therefore, before discussing the limited research on fluency use for recognition memory in childhood, we briefly consider each of the capabilities that would need to be in place in order for a child to demonstrate mature fluency attribution. Specifically, we propose that there are four basic elements of fluency attribution that children need to master in order to perform in the manner an adult typically does. These elements are presented below roughly in order of the assumed complexity of the cognitive operations necessary in order for a child to engage in the necessary activity associated with each element of the mature fluency attribution process. After briefly describing the elements, we review the extant data in the memory development literature. In addition, the limited available data describing the emergence of familiarity attribution strategies seems to indicate that children develop these abilities roughly in the presented order (although there is little data looking at the development of the final two elements).

Element 1: Facilitated perceptual processing of previously presented information. The foundation of adult fluency use is the speed and ease with which previously experienced stimuli are processed. In order to demonstrate mature use of the fluency heuristic in memory processing, children would need to demonstrate fluency effects, that is, the speeded processing of an item due to prior exposure.

Element 2: Understanding that fluency is a cue to memory. A child who does experience an item more fluently must still understand that this information can be used as a cue to memory. That is, the child must realize that fluency is positively correlated with oldness. This element is what makes attribution possible: once children 'realize' that fast and easy mental processing is often associated with the presence of previously experienced stimuli, they may begin to attribute information regarding ease of processing into their memory decisions. Critically, this 'realization' may not come in the form of a conscious strategy. Rather, children may simply subconsciously note the frequent co-occurrence between fluency and oldness and allow this relationship to inform their memory decisions.

Element 3: Learning to discount fluency in inappropriate contexts. A third component in developing a mature fluency heuristic would be learning not to *over* apply the fluency heuristic. Over-reliance on the fluency heuristic can decrease memory accuracy. Consequently, mature fluency heuristic users are sensitive to a number of factors when deciding if fluency is a diagnostic cue to memory in a given context. Attribution at this level is heuristic in nature and may happen at a non-conscious level. For example, Westerman et al. (2002) have demonstrated that when making recognition decisions, adults are sensitive to the modality in which the information was originally presented. That is, if a stimulus was originally presented in an auditory modality and is processed fluently in a visual modality at the time of the recognition test, participants discount the visual fluency and do not factor this into their recognition decisions. This is generally consistent with the actual effects of modality on memory priming: priming is greatest when the study and test items are presented in the same form (e.g., Rajaram & Roediger, 1993).

However, participants do not show sensitivity to modality when conceptual fluency is being manipulated. Participants will readily attribute this form of fluency as evidence of previous occurrence, even after a change in modality between study and test (Miller et al., 2008; Thapar & Westerman, 2009). Again, this is consistent with the effect of modality on conceptual fluency – it is not affected by changes in modality because it is based on meaning rather than perceptual characteristics.

Element 4: Purposefully strategizing about fluency as a cue to memory. Finally, advanced heuristic users may use metacognitive skills to actively strategize about the appropriateness of reliance on fluency as a source of evidence when making a memory decision. For instance, Jacoby and Whitehouse (1989) have demonstrated that a 50 ms matching prime word presented just before the onset of a test word can lead to an illusion of memory. Participants attribute the fluency generated by the prime as evidence that they have previously encountered the test word and are more likely to claim to remember the word, regardless of whether or not the word appeared on a study list. Critically though, when participants are made aware of the influence of the prime by extending the duration of the prime from 50 ms to 200 ms, participants actively and strategically discount the fluency generated by the prime and do not display a bias towards saying 'yes' to a matching prime. Some explanations of this finding posit that participants are consciously aware that fluency may be biasing their recognition decisions, and strategically taking steps to correct for its influence. However, it is important to note that Huber and colleagues have recently developed a counterhypothesis suggesting that the effect of the long prime may be due to negative priming rather than attributional factors (see Huber et al., 2008, for a review). Despite this new evidence, conscious strategic decision making regarding fluency continues to be a core element of some conceptualizations of the fluency attribution process.

For example, conscious strategizing about the role of fluency is particularly critical for theories such as Whittlesea and Williams' (1998, 2001a, 2001b) Discrepancy-Attribution Hypothesis. In the Discrepancy-Attribution Hypothesis, memory illusions arise when observed levels of fluency are significantly higher than expectations generated from the context, resulting in a conscious feeling of surprise. When participants experience this feeling of surprise they attribute the feeling as evidence of previous exposure to a stimulus. For example, Whittlesea and Williams (2001a) presented participants with a study list followed by a recognition test. Test words were preceded by either a predictive sentence stem or a non-predictive sentence stem. For example, the sentence stem 'She cleaned the floor with the ________' is predictive in that it allows the subject to generate a reasonable prediction regarding what the test word would be: the word 'broom' is an appropriate completion; the word 'guitar' is not. On the other hand, a sentence stem like 'She couldn't find a place to put the ________' is not predictive. It can sensibly be completed with any of a large number of possible answers. Critically, Whittlesea and Williams found that false alarms occurred more frequently when the words occurred in predictive sentences than when they occurred in non-predictive sentences. This effect was observed only when there was a pause between the sentence stem and the target word. The authors argued that predictive sentences led subjects to generate a set of general, indefinite expectations regarding the appropriate potential test words. When a non-studied word appeared after a predictive sentence, participants experienced a surprising level of fluency for a new test item due to the satisfying match between expected outcomes and reality. The discrepancy in processing fluency lead subjects to experience surprise, and subjects attributed their surprise to the word being old. This complex and (at least at times) conscious metacognitive attribution strategy exemplifies the nuanced strategies that may be employed by expert fluency users.

A note of caution is necessary when discussing the role of conscious processing in memory attribution. A fairly straightforward prediction of theories of fluency attribution that propose a conscious attribution process is that participants should be able to adjust their attribution strategies. If participants are consciously aware that they use a specific strategy when making decisions about the fluency of a stimulus, they should be able to adjust these strategies if experimental conditions suggest that a strategy shift would be advantageous. Experiments that have tried to create such experimental conditions have not always demonstrated the strategy shifts predicted by conscious models. For example, Miller et al. (2008) showed that participants are disinclined to attribute fluency as evidence of previous occurrence on a visual recognition test when the target had previously been presented in an auditory modality. Critically, this was true even when the experimental procedures were designed to make participants believe that perceptual fluency would be a relevant cue to memory. These results suggest that participants' response strategies may be less malleable than predicted by a conscious and volitional attribution system. Further examination of the developmental trajectory of metacognition and attribution strategies may afford a unique opportunity to shed light on this unresolved question: if children's attribution strategies change as their metacognitive skills improve, this would constitute support for metacognitive theories of fluency attribution.

Having laid out a basic framework for the cognitive skills necessary in order to attribute fluency in an adult fashion, we now turn to the literature in order to assess what is known about the development of these skills across childhood. To preview, there is good evidence of the developmental timeframe for element 1, limited evidence for element 2, and the skills necessary for children to complete elements 3 and 4 have not yet been fully examined in the developmental literature.

When do children show priming effects?

Evidence for the first step in fluency attribution, priming, has been well established (for reviews, see Lloyd & Newcombe, 2008; Parkin, 1998). In the adult literature, priming effects are demonstrated for words by measures of naming latency or perceptual identification (Murrell & Morton, 1974; Neisser, 1954) that are faster than when a word has not been presented previously. Because young children cannot read, a similar task has been employed that uses pictures instead of words (Cycowicz et al., 2000). In this paradigm, children are exposed to a study list consisting of a series of pictures. This is followed with a test phase in which participants are exposed to degraded examples of stimuli from the study list along with new stimuli (see Figure 3.1).

Children demonstrate priming by recognizing old test items at a greater degree of perceptual degrade than new test items. Children as young as 3, show evidence of priming when naming recently presented pictures. In the verbal domain, children as young as 3 to 4 show advantages in reacting to a previously seen stimulus (Parkin & Streete, 1988) or in generating category exemplars to a prompt (e.g., Perez et al., 1980). Thus, if a preschooler has been presented with a picture of a 'bear' she will be able to name it more quickly later. Furthermore, recent exposure to the prime will also make it more likely that she will respond 'bear' when asked for an example of an animal.

It is important to note, however, that while these advantages are generally consistent across development for priming that is perceptual in nature

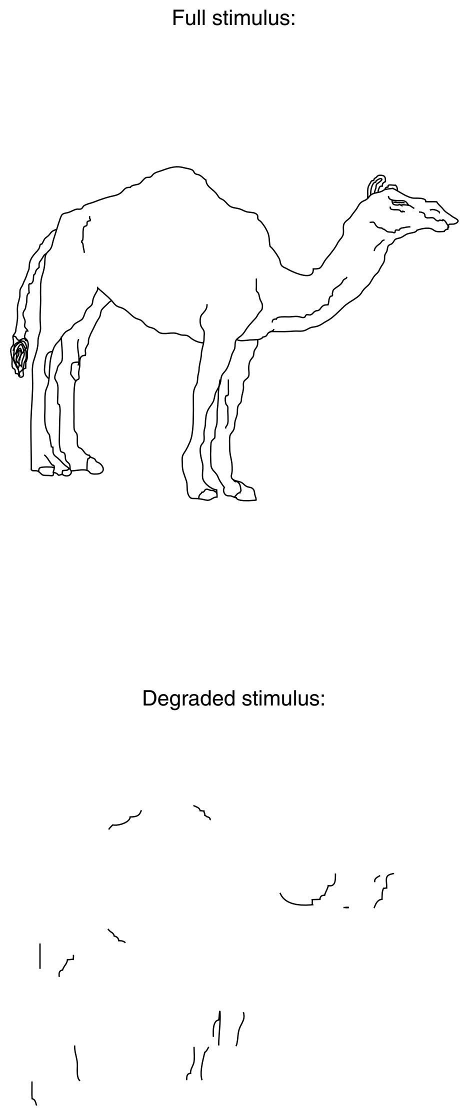

Figure 3.1 Example of the type of stimuli used in the degraded image identification task

Note: Used with permission. Copyright 2010, Life Science Associates, Bayport, NY, lifesciassoc@pipeline.com (see Cycowicz et al., 2000 for an alternate interpretation), conceptual priming seems to improve across childhood (Mecklenbrauker et al., 2003), likely due to an increased semantic knowledge base. This effect is seen directly in tasks such as stem completion, in which the child is given a word stem such as B__________, and priming is measured as the increased probability of filling in the stem with a recently presented word (Billingsly et al., 2002) as well as indirectly in studies showing that the DRM illusion increases from childhood to adulthood (Holliday & Weekes, 2006). That is, children become more susceptible to errors from semantically organized lists with age, presumably due to stronger and better organized category and concept structures in memory as well as better retrieval strategies based around these structures.

When do children attribute fluency as evidence of previous occurrence?

In contrast to the vast number of studies on priming and development, relatively little work has expressly addressed the way that children attribute processing fluency to memory. We are aware of three key studies in this domain; two using picture priming to manipulate fluency (Drummey & Newcombe, 1995; Guttentag & Dunn, 2003), and one on the attribution of fluency to autobiographical memory (Liu & Newcombe, 1999). Since the paucity of available data affords the opportunity for a complete review, we discuss each of these experiments in turn.

Drummey and Newcombe (1995) presented participants (ages 3, 5, and adult) with a list of pictures to study. Later, these pictures were presented in an incomplete form that was then slowly made complete. Participants in this experiment had two basic tasks: the first of these was an implicit identification task in which participants were asked to identify the name of the degraded image being presented. This task was followed by an explicit recognition test in which participants answered whether or not the picture had been presented on the study list. All of the participants could identify the previously studied items earlier in the completion process, demonstrating priming, but the 3- and 5-year-old children failed to use this as a guide to their recognition memory judgements. That is, they experienced enhanced processing but failed to attribute it to the recent prior experience with the item. This research suggests that children younger than age 5 do not use the fluency heuristic.

Follow-up research was able to determine that by the age of 8, children do seem to use fluency to guide memory decisions. In a procedure similar to that of Drummey and Newcombe (1995), Guttentag and Dunn (2003) gave participants a list of pictures to study. During the recognition test, the pictures were blurred and slowly brought into focus. Children of both ages (4 and 8) showed earlier recognition of studied items, but only the 8 year olds showed increased positive responses to items identified at a faster rate of focusing (more degraded pictures). This suggests that the link between implicit processing and explicit memory decisions happens somewhere between the ages of 5 and 8.

The finding that young children did not apply fluency effects to explicit memory is not limited to simple pictures that were seen once, it also occurs with naturalistic stimuli that have been repeated over many occasions at spaced delays. Lie and Newcombe (1999) tested children (the average age at time of testing was about 7 years old) on face judgements. Some of the faces presented in this experiment were those of preschool classmates who had not been seen for several years and others were novel pictures of children from the same age range. The participants made same-different judgements to pairs of faces (one frontal face, one profile face) some of which were previous classmates. All children were more successful on the judgement task when it contained former classmates. However, their explicit recognition memory for classmates versus non-classmates was unaffected by the fluent processing presumably induced by previous exposure to a classmates face. That is, whether they later correctly identified a classmate did not vary as a function of priming in the judgement task. Again, presumably, the familiar faces could be judged faster due to enhanced fluency of processing and this did not impact the recognition of a face as being from the past.

When do children learn to attribute fluency strategically?

The studies described above generally suggest that young children have a delay of up to several years between the point in time when they begin to show implicit memory effects such as priming (age 3), and the time when they begin using complex implicit cues such as fluency to guide their memory decisions (somewhere between ages 5 and 8). Due to the fact that little experimental work has examined this question, we begin by briefly examining a broader question: 'When do children begin to develop the metacognitive skills that may underpin strategies such as fluency attribution?' Typically, developing children seem to develop some basic ability to examine their own mental states by the age of 3. It is around this age that children begin to use words such as 'think' and 'know' (Flavell, 1999). Children aged 4 and 5 years begin to demonstrate some general understanding of the dynamic relationship between the variables associated with memory. For instance, when asked about their memory, they are able to verbalize that information fades rapidly from short-term memory, that more study-time results in better retention, and that the number of items to be remembered has an important effect on the probability of successful retrieval (Kreutzer et al., 1975). Additionally, children become more likely to use rehearsal to improve their memory performance between ages 5 and 9 (Flavell et al.,

1966) and source monitoring as well as reality monitoring improve between ages 4 and 6 (Drummey & Newcombe, 2002; Sluzenski et al., 2004).

Experiments aimed at assessing children's metacognitive skills have been hampered by the fact that many of the available methods of assessing metacognitive ability require basic reading skills. Such methodological limitations pose a serious problem for researchers interested in memory development: How to tell an increase in performance that arises due to advances in memory skill from an increase in performance resulting from an advance in linguistic skill. A recent example of work that cleverly avoided this conundrum is a study by Balcomb and Gerken (2008) who found that some 3-year-old children can use metamemorial strategies to guide memory decisions. In this experiment, subjects answered recognition memory questions by choosing from a target, a lure, or a third response option indicating 'not sure'. The premise is that choosing to avoid a question is suggestive of awareness that the answer is unknown. In Balcomb and Gerken's study, children learned paired associates of novel animals and objects. They were then given two tests. In their experiments, the first test was an optional memory test in which the animal was presented along with the correct object and a foil object. Both objects had appeared during the study phase but only one had been paired with the animal in question. The children could select either object as their choice or they could choose a button that skipped the question. In the second test, participants were forced to make a memory decision on each trial and could not skip a question. The results of their experiments suggested that some children were better than others at predicting when to avoid a question based on their accuracy to the test items they chose to answer. Further, the children who were more accurate in the selection process had better memory performance overall. Although the authors did not have a way to determine precisely *why* a child skipped an item, the results suggest that even at the age of 3 some children are aware of the difference between a strong memory and a weak one. Studies such as this are critical for understanding early metamemory capabilities.

Although these recent methodological advances demonstrate the potential for increased understanding of young children's metacognitive skills, at the present time there is still much to learn in this domain. Of particular note for the purposes of this chapter is that we do not know when young children are able to understand that the relationship between fluency and memory depends in part on the type of fluency and the type of previous exposure to the stimulus. Presumably, it must occur after the age at which they begin to attribute fluency to memory judgements at all (between 5 and 8 years of age). Thus, we predict that the attribution of processing fluency following a modality change should follow an inverted-U shaped function. Below the age of 6, children would fail to use fluency as a guide to memory no matter what the study/test conditions. Older children (between ages 6 and 8) should over apply fluency to memory decisions once they have learned about the link between fluency and memory. Later, the development of a more sophisticated understanding of the idea that the link between fluency and memory is partly based on the match between prior and subsequent exposure would allow a discounting of fluency after a modality change. This attributional shift relies on effective source monitoring of prior experiences, a skill that also improves across childhood (e.g., Drummey & Newcombe, 2002).

The idea of a transition into fluency attribution is consistent with Whittlesea and Williams' (1998, 2001a, 2001b) and Jacoby and colleague's (Jacoby & Dallas, 1981; Jacoby et al., 1989) discussion of the importance of attribution for fluency decisions. Effects of fluency of processing are observed before attribution to memory decisions. Whittlesea and Williams discrepancy-attribution model of fluency is also consistent with a developmental transition into fluency attribution. This model assumes a four-step process for the connection from fluency to positive memory judgements: Expectation, uncertainty, surprise, and then an attribution of familiarity. Until young children have sufficient experience with memorial tasks, it is unlikely that they could develop expectations for what a processing experience should be like. Thus, one way to test the lower boundary of fluency attribution may be to train young children on what sort of information to use and expect from memory decisions.

What is the impact of aging on the fluency heuristic?

Once a child has acquired a mature fluency heuristic, the strategy should remain stable across the lifespan. Recent reports indicate that the fluency heuristic seems to be relatively spared by normal aging (although see Yano et al., 2008 for a discussion of fluency in Alzheimer's patients) in contrast with other memory processes such as source monitoring (e.g., Craik et al., 1990) or memory binding processes (e.g., Chalfonte & Johnson, 1996). Thapar and Westerman (2009) tested older adults (M = 68) using a masked prime to enhance the fluency of some of the test items. Similar to younger adults (Jacoby & Whitehouse, 1989; Westerman et al., 2002), older adults readily applied fluency to their recognition decisions. This is not a surprising finding, as amnesiacs have also been shown to use fluency to guide recognition decisions (Verfaellie & Giovanello, 2006). Instead, the key results of the study were that older adults continued to selectively apply fluency in accordance with the initial presentation of the item during the study phase. That is, fluency effects from a masked prime during the recognition test were larger if the study list was presented visually than presented auditorily. Thapar and Westerman argue that this is due to a continued ability to use the fluency heuristic. In a second study, fluency was enhanced using a more conceptual manipulation of a lexical decision task which sometimes embedded the test item. In this case, again similar to younger adults (Miller et al., 2008), older adults readily applied conceptual fluency to a recognition decisions regardless of the modality the study list was presented in. Based on these studies, fluency attribution processes seem resistant to declines during aging.

Conclusion

Understanding the circumstances under which people attribute fluency as evidence of previous occurrence has been a major goal of memory researchers for over 20 years. It is surprising then that so little data exist examining the question of how fluency attribution skills develop. The review of the literature presented here demonstrates that young children are sensitive to previous exposure to a stimulus on measures of priming, indicating that even very young children are capable of the type of processing described by element one of the framework outlined earlier in this chapter. However, beyond this point the picture clouds considerably. Specifically, the age at which children begin to use this enhanced processing to guide memory decisions remains to be determined. Further, the ample evidence that adults will change fluency attribution strategies in order to accommodate expectations and test conditions has yet to be examined from a developmental perspective. It is our hope that the framework presented here will prove to be of value to future researchers as they attempt to resolve these important questions.

4

Attributions of Fluency: Familiarity, Preference, and the Senses

Antonia Mantonakis, Daniel M. Bernstein, and Elizabeth F. Loftus

Introduction

Consumer judgement and decision making is guided by phenomenological experiences (Whittlesea, 1997), also called 'non-emotional feelings' (Clore, 1992) associated with cognitions. These feelings, such as certainty, surprise, and confusion, are considered non-emotional because they are feelings associated with a state of knowledge (Clore, 1992), as opposed to emotional feelings of happiness, anger, and sadness, which relate to the state of a person. These feelings, which may arise from incidental exposures to contextual information (e.g., exposure to promotional materials), can influence a person's memory, and hence affect one's feelings of familiarity, preference, and sensory evaluation. The role of memory in preference is not clear in most models of judgement and decision making (although see Weber & Johnson, 2006). We believe that the concept of fluency (in general) and more particularly Whittlesea's (1997) *Selective Construction and Preservation of Experience*s (henceforth SCAPE) account may be useful as a framework for understanding consumer judgement and decision making.

We focus on fluency, which is a metacognitive cue that reflects the relative ease or difficulty that a person experiences while performing a cognitive operation, and how fluency can lead to inferences about the external environment. We organize our discussion around the role of fluency in familiarity and preference judgements, and attempt to integrate findings from cognitive, social, and consumer psychology to provide new insights into consumer behaviour. We review both laboratory and field studies and propose new ideas about the role of fluency in evaluation about experiential objects involving taste, touch, sound, and smell.

Fluency

When people make judgements about previous experiences or current preferences, they have access to both mental contents that are produced (e.g., the perception of a wine's label, including pictures), and the subjective experiences that accompany those contents (e.g., fluency of processing the label's words). The subjective experience of fluency refers to the relative ease or difficulty in processing mental contents. Fluency prompts inferences about many different aspects of the environment, including an item's value or familiarity. People attempt to attribute this ease or fluency to an appropriate source (Whittlesea & Williams, 1998), based on their intuitive theories of cause and effect (Schwarz, 2004).

Fluency can be conceptualized as falling into one of two broad categories of influence. The first category is 'perceptual,' arising from the subjective ease at processing an item's font, colour, or other visuo-perceptual details. Perceptual fluency can arise from prior exposure (Whittlesea, 1993), orthographic regularity (Whittlesea & Williams, 1998), or linguistic regularity (Shah & Oppenheimer, 2008). Perceptual fluency has been found to influence a variety of judgements, including preference (Novemsky et al., 2007). Unfortunately, there is little known about causes and outcomes of perceptual fluency involving touch, taste, smell, and sound (although see Miller et al., 2008). The second category of fluency is 'conceptual,' arising from the subjective ease at processing an item's meaning. Conceptual fluency can arise from priming an item's semantic associations (Whittlesea, 1993). Conceptual fluency has also been found to influence a variety of judgements, including preference (Lee & Labroo, 2004).

The process by which fluency is used as a cue in judgement and decision making is complex. In any type of decision, absolute judgements are much more difficult to make than relative judgements (Weber & Johnson, 2009). For example, in music, absolute or perfect pitch (the ability to name or recreate a musical note played in isolation) is far more difficult and less common than relative pitch (the ability to name or recreate a musical note played after hearing another note; Sacks, 2007). A person's reference point is determined by the context (e.g., other options in a choice set of wines), general expectations, or specific expectations; people use reference points when making judgements. A common approach is to view fluency as the difference between expected ease (which can be based on anticipation, Jacoby & Dallas, 1981; Whittlesea & Williams, 2001a, 2001b), or the context (Briñol et al., 2006; Unkelbach, 2006), and actual ease, in conditions of uncertainty. If a person had complete certainty about which wine was the best value, there would be no need to rely on other cues, including metacognitive cues such as fluency to make a wine-purchasing decision.

Inference, attribution, and construction are ideas that form the core of Whittlesea's (1997) SCAPE account of memory. According to the SCAPE account, each mental representation is preserved to serve as a resource for perception and performance on future occasions. The effects of prior experiences on current behaviour do not simply involve retrieval of a mental representation, but also pertain to the contextually driven subjective quality of that retrieval. The account posits that any mental event, regardless of whether it consists of the recall of an experience at a winery, or the identification of taste on the palate, occurs through a constructive process that involves two steps. The first is the production of mental events, whereby ideas are brought to mind. Production can be the result of an external stimulus (situational cue), such as when a consumer has to make an online stimulus-based choice (e.g., which one of the wines in this flight is my favourite?), or simply the consequence of generating a thought, such as when a consumer has to make a memory-based choice (e.g., which wine from those sampled on last week's wine tour was my favourite?; Lynch & Srull, 1982). Either way, production is guided by prior experiences.

The second step in SCAPE is the evaluation of the quality of those mental events (the fluency or elaborateness of processing), producing subjective experience. The purpose of the evaluation function is to evaluate the goodness of the mental event. Ideas about an experience, whether it is in the moment (sampling a wine flight) or a reflection of the past (remembering a winery tour) can come to mind easily, with a lot of detail. The outcome of those ideas (fluency) requires an attribution, linking the subjective experience to an internal or external source. A flood of vivid memories (coming to mind fluently) from a winery tour associated with a specific cabernet franc may lead to the conclusion that that particular wine was the best tasted on the tour.

The evaluation process is inferential and unconscious. The evaluation process is also guided by prior experiences, in interaction with current expectations, based on the context. If repeated exposure to a piece of art produces a metacognitive experience of fluency, a person might interpret that fluency as pleasantness in the context of a preference judgement, or familiarity in the context of a recognition judgement (Whittlesea & Price, 2001). Speculatively, the experience of humour is the result of a violation of expectation between an expected and actual outcome. In the context of a joke, the source of fluency is attributed to humour. The ultimate violation of expectation is randomness, explaining why comedy relies on randomness as a source of humour. To illustrate:

> 'Haikus are easy
> But sometimes they don't make sense
> Refrigerator.'
> (Anonymous quote from a T-shirt)

The reason that humour may rely on surprising fluency is that the most dramatic effects of fluency occur when it is unpredictably high or low compared to some expectation (Whittlesea, 2002; see also Labroo & Kim, 2009). Fluency must be *surprising* to show an effect on judgement (Whittlesea & Williams, 1998, 2000, 2001a, 2001b). Said another way, when people are aware of the reason for the relative ease or difficulty of processing (hence there is no perceived discrepancy in processing), there is no reason to make an attribution.

Fluency and familiarity

An essential characteristic of the remembering process is the feeling of familiarity; the feeling's source provides a useful illustration of the unconscious attribution process. Intuitively, a feeling of familiarity would arise when an event has occurred in the past; the experienced event creates a memory trace, which is activated upon encountering a cue for that event. Contrary to this idea is the notion that feelings of familiarity are mediated by an unconscious attribution process, and do not always rely on having a memory trace (Jacoby et al., 1989). The illusion of familiarity, created by the enhanced processing of a novel stimulus provides a potent illustration of how the feeling of familiarity is not always a direct result of cueing prior representations in memory. Walking into a winery in Niagara-on-the-Lake, for the first time, and seeing the tasting room (wine bar, wine bottles, oak barrels), smelling the wine, and then being asked what you'd like to try, may produce a feeling of familiarity. In this case, unexpected fluency occurs because of the sense that many of the features of the situation resemble specific features of other, already encountered situations from one's past, in a surprising way. That is, the person's fluency is surprising in the situation, consciously experienced as familiarity (see also Mandler, 1980).

Prior experience with a stimulus enhances the fluency with which that stimulus is processed in the same way that practice makes a skill easier to perform. Because of this veridical link between fluency and actual past experiences, people rely on processing fluency as a heuristic in deciding that they have experienced an event before making an attribution to prior experience (saying 'I'm sure I've been to this winery before!'). There is correlational evidence for this attribution process (Jacoby & Witherspoon, 1982) as well as experimental evidence, whereby fluency is experimentally manipulated independent of prior experience through manipulations such as visual clarity (Whittlesea et al., 1990).

It is not just perceptual fluency (Jacoby & Dallas, 1981), but also conceptual fluency of processing, manipulated without participants' knowledge, which can lead to illusions of familiarity (Whittlesea, 1993). The illusion occurs because of an unconscious attribution process that arises without access to memory contents or to the cognitive processing that may be driven by memory contents. This heuristic process can't differentiate cognitive processing that has been enhanced by actual experience (memory) versus external sources.

The fluency heuristic is used selectively for familiarity only when there is an expectation that current processing should be affected by past experience.

Assuming this, people are blind to two aspects of their own processing. People are unaware of the difference between fluency of processing arising from prior experience versus fluency of processing arising from an experimental manipulation and that they are making an inference and attributing fluency to something. If the inference becomes conscious the process does not occur (Jacoby & Whitehouse, 1989).

One corollary of the SCAPE framework is the discrepancy-attribution hypothesis. According to this hypothesis, when there is an unexpected or surprising mismatch between expected and actual performance on a given stimulus in a given context, the perceived discrepancy is consciously experienced as the feeling of familiarity, and unconsciously attributed to a prior experience of that stimulus. When a person experiences 'surprising fluency', the surprise leads to a feeling of familiarity, and the person attributes the surprise to the past. If thoughts about a winery experience come to mind with surprising fluency, it is the surprise that leads to a feeling of familiarity.

The perception of discrepancy is thought to occur when outcomes either violate or validate expectations in a surprising way (hence, 'surprising fluency'). Often this surprise occurs because the expectation is a constrained, indefinite one, so that the relationship between expectation and outcome is ambiguous (Whittlesea, 2002b).

Fluency and preference

Other illusions that arise when prior experience is manipulated independent of prior experience include illusions of truth (Begg & Armour, 1991), visual clarity (Whittlesea et al., 1990), another person's performance (Jacoby & Kelley, 1987), fame (Jacoby et al., 1989), good category membership (Whittlesea & Leboe, 2000), good choice (Bodner & Mulji, 2010; Novemsky et al., 2007), and risk perception (Song & Schwarz, 2009). To illustrate the latter, fictitious amusement park rides were rated as less risky if they were easy (Chunta) versus difficult (Vaiveahtoishi) to pronounce.

Fluency can also be misattributed to pleasantness (Whittlesea, 1993; Zajonc, 1980) likeability (Bornstein & D'Agostino, 1994), and valuation (Alter & Oppenheimer, 2006). It is not just the perceptual fluency, but also the conceptual fluency of processing, manipulated without participants' knowledge, that can produce illusions of pleasantness (Whittlesea, 1993). It has been argued that like the illusion of familiarity, the illusion of pleasantness is thought to occur because of an unconscious attribution process that occurs without access to memory contents or to the cognitive processing that may be driven by memory contents. However, the attribution is more pronounced the more 'subliminally' the items are presented (Bornstein & D'Agostino, 1994). The effect of processing fluency on pleasantness is non-monotonic: with additional exposures, boredom sets in and pleasantness ratings attenuate (Bornstein et al., 1990; see Berger & Fitzsimons, 2008, for a view of how frequent exposures to conceptual cues can 'accumulate' to influence product evaluation and choice).

The notion that previous exposures can influence liking is not a new one. One of the most studied findings in social and consumer psychology is the mere exposure effect, which is the finding that, as exposure frequency increases, so does preference (Zajonc, 1968). If a person sees a banner ad for a brand several times (even in peripheral vision), they tend to like that brand more (Fang et al., 2007). This finding occurs even if the repeatedly presented item is shown for only 5 ms at a time (Bornstein & D'Agostino, 1994). The effect on preference is observed in the absence of recognition, which poses a puzzle. Whittlesea and Price (2001) questioned why it is that fluency – an important cue to recognition – is only associated with increased preference (and not recognition) in mere exposure effect studies. To address this question, participants were shown pictures in a training phase, either one, three, or five times. When participants were asked to justify (i.e., use an analytic processing strategy) why an item was preferred (hence, to not experience the fluency), preference was at chance, whereas when participants were told that they 'would have no cues to recognize the items,' but had to use general, categorical familiarity (hence, could experience the fluency), recognition increased with repetition. It is the dimension that is made salient by the task that dictates the source to which the fluency will be attributed.

In each case, there is a misattribution of fluency to the most likely or available source (see Schwarz, 2004, for a review), due to an unconscious inference. The difference, however, is that the item's 'oldness' (recency; Lee & Labroo, 2004, or frequency; Berger & Fitzsimons, 2008), which causes fluency, is a relevant dimension for recognition, whereas an item's 'oldness' is not a relevant dimension for most pleasantness judgements (with some exceptions; wine is believed to improve with age), making the attribution erroneous. When an item is fluent due to factors other than its actual 'oldness,' the attribution to pleasantness will be erroneous (Whittlesea, 1993).

There is an alternative explanation for the mechanism underlying affective judgements, which differs from the mechanism underlying cognitive judgements (Lee, 2002). Note that cognitive judgements, such as recognition, have correct and incorrect answers (Lee, 2001; Zajonc, 1980). The explanation is as follows: both conceptual and perceptual fluency lead to more positive attitudes. However, only conceptual fluency, when associated with negatively valenced concepts, can lead to negative attitudes (Lee & Labroo, 2004, Experiment 4). Unlike previous research that shows that fluent processing is positively valenced (Harmon-Jones & Allen, 2001), conceptual fluency can be either positively or negatively valenced because of the possibility of spreading activation to positive or negative constructs in semantic memory (Collins & Loftus, 1975).

This finding that conceptual fluency, when associated with negatively valenced concepts, can lead to negative attitudes calls into question the need to use a misattribution model to account for effects of conceptual fluency on affective judgement (Lee & Labroo, 2004). According to Lee and Labroo (2004), Whittlesea's (1993) argument that conceptual fluency is misattributed to pleasantness is based on other data showing how people misattribute perceptual fluency to psychophysical judgements (loudness, brightness, etc.), and there has not been a direct test of the misattribution model on affective judgements.

In a series of studies in line with Lee and Labroo's (2004) explanation, we have examined the effect of forming either negative or positive associations in memory on preferences and behavioural intentions. In one study, participants were led to believe that as children, they had gotten sick from strawberry ice cream. After receiving the false suggestion the participants become more confident that they had gotten sick from strawberry ice cream. This false autobiographical belief resulted in a decrease in preference for strawberry ice cream and an intention to avoid strawberry ice cream (Bernstein et al., 2005a). In order to separate those who were susceptible to the false suggestion from those who were not, 'believers' were distinguished from 'non-believers' based on two criteria. First, participants must have initially indicated a low confidence rating that they had gotten sick from strawberry ice cream, with an increase in confidence after the suggestion. Second, the participants must have generated a specific memory or a non-specific belief ('I just know that it happened, but can't recall when, where, or how') that the critical event had occurred. After separating participants according to these criteria, we found that 'Believers' were more likely to 'avoid' the critical item and were less inclined to want to eat it than the non-believers. These findings show that providing a suggestion can lead to false memories for negative food-related experiences, and that certain behavioural outcomes emerge such as decreasing preferences towards the food.

In another study, participants received a suggestion that as children, they had gotten sick from dill pickles and hard-boiled eggs. By examining participants' confidence ratings before and after the suggestion we found that those who believed the suggestion significantly increased their confidence ratings that they had gotten sick from these items when they were a child (Bernstein et al., 2005b). Furthermore, participants decreased their preferences towards the items and were also more willing to avoid them.

In another study, this time using asparagus as the critical item, we suggested to participants that they loved to eat asparagus as children (Laney et al., 2008). 'Believers' not only reported more desire to eat asparagus, they rated pictures of asparagus as more appetizing and less disgusting (versus their pre-suggestion ratings). Believers also increased their willingness to pay for asparagus (versus a group that did not receive the suggestion). Later studies showed that people who were seduced by the false information actually ate less of the food (Geraerts et al., 2008).

The notion of surprising fluency can be used to account for these suggestion effects: that is, people believe the false event (e.g., getting sick from ice cream; having loved asparagus), with confidence, leading to the behavioural consequence (lowered preference for the ice cream; higher ratings towards asparagus). Although Lee and Labroo's (2004) conceptualization of fluency can also account for the finding that altered memories can lead to both increases and decreases in preference (given that conceptual fluency may entail spreading activation to positive or negative constructs in semantic memory), we believe that their explanation does not account for the full pattern of data. Rather, we think that one must induce belief through elaboration and imagination in order to observe behavioural consequences of the false memory (Bernstein & Loftus, 2009).

Fluency and perceptions about the senses

Our discussion has focused on perception (conceptually driven), and not sensation (data-driven), and on cognitions rather than experiential objects that can be consumed. An examination of perceptions relating to the senses would provide valuable insights into theories of recognition and classification that have not been examined in the context of sensory evaluation judgements, such as the aroma and taste of a wine. Much of the findings from perception may be relevant to experiential evaluations, because cognition and sensory thoughts play major roles in experiential evaluations. Advertisements that mention multiple sensory experiences (smell, sight, and sound) versus a single sensory experience (taste only) can increase the number of positive sensory thoughts a person generates, and subsequently leads to increased perceptions of taste, as measured by tastiness ratings (Elder & Krishna, in press).

Perception and taste, touch, smell, and sound

Given that cognitions relating to sensation influence taste perception, can metacognitive experiences also influence perception that is used in recognition or classification judgements relating to experiential items? When classifying the taste of a wine as one grape varietal over another, what process(es) does a person use? Does fluency play a role in taste judgements? If so, what would 'experiential fluency' be?

One potential avenue to address this question comes from the work of Oppenheimer and Frank (2008), who examined the effects of the metacognitive cue of fluency on categorization judgements (perception). The logic used was that typical exemplars of a category are frequently experienced, easily accessible, and the most primed by their associates in memory. Therefore, over the course of one's lifetime, the metacognitive experience of fluency co-occurs with the judgement of good category membership, making fluency a valid cue to category membership. In their studies, they used natural and artificial categories (mammals, vehicles) and used a fluency manipulation that varied font type. Participants had to perform exemplar verification (bird?) or feature verification (has wings?). Typicality ratings for exemplars were significantly lower in the low fluency condition (10-point Mistral font type) than the control condition (12-point Times New Roman font type).

Oppenheimer and Frank's (2008) findings challenge traditional theories of categorization, including prototype theories (Rosch & Mervis, 1975), exemplar theories (Brooks, 1978; Medin & Schaffer, 1978), and theory and knowledge-based models (Rehder & Hastie, 2001). In general, these models do not take into account metacognitive information in classification. These models may be able to account for fluency if they add it as a 'feature' or 'knowledge,' however future research is needed to examine the role of fluency in sensory thoughts.

An examination of sensory thoughts would provide insights not only into the theories of recognition and classification, but also on the role of metacognitive experiences in sensory evaluations. Sensory scientists do not consider the role of metacognitive experiences in sensory evaluation, as they often assume that consumers are rational decision makers (Köster, 2003). While much is understood about bottom-up processing, such as genetic differences leading to variations in taste sensation (Bartoshuk et al., 2005), top-down processing, such as the role of visual and verbal cues on cognitive processes that can lead to biased judgement in sensory evaluation, are only beginning to be understood.

In terms of top-down processing, research has shown that a brand name can influence taste perception (Allison & Uhl, 1964), and extrinsic cues, such as information pertaining to ingredients (Lee et al., 2006), or visual cues such as colour (Hoegg & Alba, 2007) can change one's taste experience. For haptics (touch), visual ads (e.g., showing a kitten) versus verbal ads can result in higher perceived softness for a product (Mitchell & Olson, 1981).

There has been very limited research examining how metacognitive information independent of sensory evaluation can lead to errors in perceptions of taste, touch, smell, and sound (although see Krishna, 2006). Does a person's fluency of processing cognitions relating to a given sensation (touch, smell, taste, or sound) bias judgements of the quality of that experience? Can the context alter a person's expectations of a sensory experience, thereby altering perception?

When it comes to actual consumption, some have argued that sensory systems have been optimized by evolution (Abdi, 2002), and that sensory inputs are inherently evaluable (Hsee et al., 2009). Thus 'sensory utilities' (versus prediction or memory utilities) should not be biased by contextual factors (Hsee et al., 2009). However, would fluency, a metacognitive cue arising from context, affect perception of taste, touch, or smell of the item? If so, is it the physiological perception or intensity that is affected, or the hedonic response, which may include liking or acceptability?

In terms of metacognitive ease, one issue that has arisen in the Niagara region in Canada is the debate between whether to focus marketing efforts at promoting our Riesling versus focusing efforts at promoting our Gewürztraminer. Critics of the latter idea claim that 'no one can even pronounce it,' 'it's so frustrating to say,' or 'how can someone enjoy something they can't even read?' It is a question open to investigation: Does fluency of the grape varietal name lead to changes in taste perception, or discrimination between the two grape varietals? One way to test this notion is to provide participants with the identical wine, however, with labels of grape varietals that are either familiar (Merlot), fluent but non-familiar (Moscato), or disfluent and non-familiar (Mtsvane). Then ask participants to rate their preference and willingness to pay for the wine. Would people's taste experience (physiological, or hedonic) be altered by the label manipulation? We have studies in progress exploring this issue.

Sensation: sensing touch, taste, smell, and sound

Bottom-up sensory evaluations are made using a variety of tasks including identification and discrimination. In discrimination tests, participants are provided with three samples whereby two are the same, and one is different. This is typically called a triangle test. Participants have to identify which of the three samples is the different one (Raghubir et al., 2009). The number of people who correctly choose the different sample must be corrected for guessing. In the repeated-pair test, participants are asked to evaluate two different samples (using preference, identification, or discrimination) repeatedly, and the consistency of their ratings is examined (Buchanan & Henderson, 1992). In another version of the discrimination task that uses a scale, several pairs of samples are provided to participants (some same, some different), and participants make similarity ratings on a scale (1 = *identical taste*; 7 = *different taste*; Hoegg & Alba, 2007).

Although these tasks involve low-level sensory discrimination, it may be the case that even low-level sensory discriminations can be influenced by top-down, metacognitive influences. There are virtually no studies examining the influence of metacognitive experiences on sensory discrimination. Could fluency influence changes in sensory discrimination? This is an important question because accurate sensory discrimination is the basis for differences in consumer preference (Raghubir et al., 2009). One way to address this issue is to manipulate fluency related to sensory experience, and then assess whether fluency can affect discrimination. The example above with pronouncing grape varietal names is one in which the fluency experience is rather divorced from the sensory judgement, but it might still affect the sensory experience of discrimination. One way to test this would be to give participants a repeated-pair test, and examine discrimination ability. One pair would contain a wine in a juice glass and a juice in a wine glass. The same pair would be repeated, although on the second trial, the juice would be in a juice glass and the wine would be in a wine glass. Here, the experience of the drinks (in terms of evaluation, preference) might be affected by one's expectations about what will be tasted and the actual taste.

One could extend the notion of fluency into auditory sensory judgements: Listening to a choral concert in a medieval church versus in a baseball stadium, for example. The former produces a rich sensory and perceptual experience, while the latter does not. The same idea applies to touch: How do people experience the feel of accidentally stepping on a rogue tomato in a produce store versus stepping on the same tomato during the energizing Tomatina Festival in Spain?

Taken together, the metacognitive cue of fluency that a person experiences while performing a cognitive operation can lead to inferences about the external environment. According to the SCAPE account, each mental representation (whether motoric, cognitive, or sensory) is preserved as a resource for perception and performance on future occasions. The effects of prior experiences on current behaviour are the result of a constructive process that involves two steps, production of mental events, and the evaluation of the quality of those mental events (the fluency or elaborateness of processing), producing subjective experience. The purpose of the evaluation function is to evaluate the goodness, or the source, of the mental event, based on current expectations, which are guided by the context, which can be an event, or an item associated with the to-be-judged object.

The metacognitive experience of surprising fluency, which may arise from incidental exposures to contextual information, false feedback, or other experimental manipulations discussed here, can influence a person's memory, and therefore one's feelings of familiarity, preference, and sensory evaluation. The ideas presented here contribute to the literature on the role of memory in preference judgements (Weber & Johnson, 2006). We believe that the concept of fluency (in general) and more particularly Whittlesea's (1997) *Selective Construction and Preservation of Experiences* account may be useful as a framework for future research on consumer judgement and decision making.

Note

Antonia Mantonakis wishes to acknowledge the Chazen Institute for International Business at Columbia Business School, and the support of a Discovery Grant from the Natural Science and Engineering Research Council of Canada and a Standard Research Grant from the Social Sciences and Humanities Research Council of Canada. The authors wish to thank Reh Mulji and Jason Leboe for comments on an earlier draft of this chapter.

5
The Role of Familiarity in Implicit Learning

Zoltan Dienes, Ryan B. Scott, and Lulu Wan

Introduction

On the conference circuit it was always an invigorating experience to be 'Whittlesead,' as the saying went – to bump into Bruce at coffee, ask him a question, and hear at high speed how the issue interconnected with Bruce's often insightful take on all the various workings of the mind. Whether or not you agreed with him, or even knew whether you agreed with him, he always provided a fresh view worthy of serious thought.

Bruce's innovative reworking of theoretical approaches to memory included the key distinction between the production of a representation of the world, on the one hand, and an evaluation phase that determined the conscious phenomenology associated with that representation, on the other (see e.g., Mantonakis et al., 2008, for a recent overview of Bruce's ideas). We will follow this same broad distinction. That is, we accept that a mental state that just represents the world is not in itself conscious. It is only a second evaluative process that represents one as being in a certain mental state – of remembering, knowing, seeing, somehow experiencing – that allows the experience to seem a certain way, that is, to have conscious phenomenology; indeed, this assumption has been key in our own work (see e.g., Dienes, 2008, for a review).

In understanding memory and cognition, Bruce further highlighted both the relevance and irrelevance of fluency as a feature of the production phase. Fluency is the speed with which the production is formed – which when evaluated as a sign of oldness may lead to feelings of familiarity. We will also sceptically explore the role of fluency in forming people's feelings of familiarity. In fact, the aim of this chapter is to investigate feelings of familiarity as a common feature of implicit learning, a line of inquiry that enjoys a particularly high risk of getting one Whittlesead – if only Bruce could still be found at coffee breaks on the conference circuit.

In this chapter, we will argue that familiarity refers to a unidimensional signal of degree of constraint satisfaction. (Bruce might regard this constraint-satisfaction signal as similar to his notions of coherence or integrality – all these notions refer to global often unconscious evaluations of a production.) Familiarity is 'unidimensional' in that one can have more or less familiarity but it is not otherwise structured. It indicates 'constraint satisfaction' in that it indicates how predictable components of the stimulus are given past experience. The constraints so satisfied can refer to the occurrence of commonly co-occurring elements (chunks) but also to other more abstract patterns. We argue that people are often aware of this signal as a feeling of familiarity, though it can influence classification decisions unconsciously, and people are often unaware of the constraints whose satisfaction resulted in the feeling of familiarity. Despite the intuitions of many researchers, we find the signal bears no detectable relationship to fluency. Finally, and despite the intuitions of many researchers, we find familiarity can be used to control which of two bodies of implicit knowledge are used to make a classification.

Implicit learning and familiarity

Implicit learning is a process by which one 'learns about the structure of a fairly complex stimulus environment without necessarily intending to do so and in such a way that the resulting knowledge is difficult to express' (Berry & Dienes, 1993, p. 2). For example, people exposed to strings of letters generated by an artificial grammar can later recognise new strings as grammatical or not, while finding it hard to describe what distinguishes grammatical from non-grammatical strings (Reber, 1967). Artificial grammar learning is a paradigm that both we and Whittlesea have used extensively as a testing ground for theories about implicit learning.

There is some debate over what can be implicitly learned (see Pothos, 2007, for a review). Reber proposed we learn abstract knowledge; for example, the allowable bigrams and their frequency in the grammar (Reber & Lewis, 1977). Whittlesea favoured the idea that we store only the individual experiences of each training episode (e.g., Whittlesea & Dorken, 1994), continuing (and elaborating) a Canadian tradition started by Brooks (1978). Brooks proposed that people store each training item, and the similarity of a test item to training items can be used to classify without the need for explicitly represented rules, contra Reber. We have favoured neural network models of implicit learning (e.g., Dienes, 1992; Kuhn & Dienes, 2008), which can flexibly learn a number of regularities, approximate processing principles of the brain, and can have a degree of abstractness along a continuum from exemplar models (like Brooks) to symbolic representations of rules (Cleeremans, 1993).

Whatever the precise basis of the learning mechanism, modellers have found it useful to postulate a summary signal of the net coherence a test item has with the structure of the training items. For example, in neural network models of artificial grammar learning (see Cleeremans & Dienes, 2008, for a review), the network attempts to reproduce the presented test stimulus. The more the test stimulus satisfies all constraints learned from the training items, the better the reproduction. In order to classify, the correlation or match between the reproduction and the stimulus is calculated. Thus, the correlation or match is a unidimensional signal of degree of constraint satisfaction. The greater the signal, the greater the probability that the test item is classified as grammatical.

Brooks' (1978) key insight that people can show rule-like behaviour by learning only specific exemplars has been implemented in a number of computational models. For example, Hinztman's (1988) MINERVA model, used recently by Jamieson and Mewhort (2010) to model artificial grammar learning, stores all training items. A test item will produce an echo whose *content* reflects which parts of the test item resonate best with all training items, and whose *intensity* reflects the overall similarity of the test item to all training items. Incidentally, the echo content comprises a constant automatic construction of abstract representations (in principle unconscious until 'evaluated'), in a similar way as a neural network automatically abstracts. The echo intensity is a unidimensional signal of degree of constraint satisfaction, also just as in the neural network case.

What does this signal represent? Across a range of different models of implicit learning, it is a qualitative non-conceptual indication of the degree to which the test item follows the same constraints as previously encountered items. It is thus a fine-grained signal of how old the test stimulus is, or how old its features collectively are, given that the constraints satisfied are those defined by old stimuli. In this sense, the signal represents events in the world: The previous meeting of the subject with the item or its features. It thus represents the familiarity of the subject with the test material.

'Familiarity' in the way just used refers to a state of affairs in the world, the fact of having been in contact with something or not. 'I am familiar with the Simpsons' means I have had contact with the Simpsons. 'My familiarity with Bach is not great' is a statement of the extent of my contact with Bach's music. In contrast, psychologists typically use the word 'familiarity' to refer to a mental state or process. Familiarity in this sense refers to the person's representation of the fact of them having had prior contact with the presented material. It is *subjective* familiarity. In this sense, the constraint-satisfaction signal *is* the person's familiarity with the test item.

The signal is a *first-order* representation in the sense that the representation is about the world. As such, according to higher order theories of consciousness (Carruthers, 2007; Dienes, 2008; Rosenthal, 2002), it is an unconscious representation (or mental state) until the person is aware of being in that state – by representing to themselves that they have the first-order representation. Representing 'I have a feeling of familiarity' makes the familiarity conscious. Consistent with Whittlesea's notion of an evaluation stage, it is this second-order representation that creates the *conscious* feeling of familiarity. Familiarity as a feeling, a first-order representation, is the second sense of familiarity; familiarity as specifically a conscious feeling (i.e., as a second-order representation) constitutes a third possible sense of the word familiarity. But we will mainly use familiarity in the second sense, subjective familiarity, the subject's take on the extent of their acquaintance with the test material, a feeling that will often be conscious but need not be.[1]

Familiarity can in principle be based on many sorts of constraints: The relevant constraints cannot be determined a priori. Servan-Schreiber and Anderson (1990) were the first to suggest people may use feelings of familiarity in the artificial grammar learning task, and they suggested such feelings were based on the frequency with which chunks of letters in the test item had occurred in the training phase. Indeed, in classifying test items, people do become sensitive to the frequency of training chunks as suggested (e.g., Knowlton & Squire, 1994). But we now know that in learning artificial grammars people can become sensitive to other features as well: For example, the pattern of repetition structure in a string of letters (e.g., Vokey & Higham, 2005; Whittlesea & Dorken, 1994) or, more abstractly, global symmetries in musical stimuli (Kuhn & Dienes, 2005). That is, familiarity might be based on a range of possible constraints. Familiarity is a particular similarity function mapping from test items to training items, and what that function is in any particular context needs to be determined empirically. Anecdotally, it seems people have little insight into the nature of the similarity function. They don't know what or even what sort of feature might make a person or object seem familiar or unfamiliar, and have little control of its terms (i.e., of the features and their weightings contributing to the feeling of familiarity). We show people are indeed often unaware of the features upon which familiarity is based. The lack of control seems so self-evident that Jacoby (1991) took uncontrollability to be definitional of familiarity. But later in this chapter we turn it into an empirical issue and show people do have some broad control over the terms of the function (even if they do not know what its terms are).

Persuasive evidence that people do rely extensively on a unidimensional signal in classifying test items in the artificial grammar learning paradigm comes from analyses of receiver operating characteristics (ROCs). ROCs provide an indication of the underlying cognitive processes on which judgements are based. For example, Kinder and Assmann (2000) demonstrated that the ROCs for an artificial grammar learning task are consistent with a signal detection model that assumes a continuous underlying dimension, which they postulate to be familiarity. Lotz and Kinder (2006) further demonstrated that ROCs remain consistent with a continuous underlying dimension under transfer conditions, that is, when tested on materials employing the same grammar but instantiated in a different letter set from the training stimuli. Tunney and Bezzina (2007) employed ROC analyses to show that in artificial grammar learning there was not exclusive reliance on a continuous underlying dimension but such reliance increased with delay after the training phase.

If we are interested in the role of subjective familiarity in implicit learning and artificial grammar learning, we need a direct test of that role. The most direct way of measuring the extent to which a person represents their acquaintance with test material is to ask them to rate their feelings of familiarity. Higham and Vokey (2004) collected such ratings in the context of a word recognition experiment. Norman, Price, Duff, and Mentzoni (2007) did not take ratings but found that at the end of an implicit learning experiment, people often stated they used feelings of familiarity to determine their responses. In a series of papers, we asked people to rate their familiarity on a 100-point scale after each test item on an artificial grammar learning task (Scott & Dienes, 2008, 2009, 2010a, 2010b, 2010c) so we could explore the determinants and the role of such feelings of familiarity in implicit learning. We found rated familiarity of test strings was predicted by various structural properties of the training strings ($R = 0.41$). Familiarity was correlated with repetition structure, the frequency of bigrams and trigrams in the training phase, and the overlap with specific training strings (in that order). That is, familiarity becomes sensitive to the structural properties of the items on which the person was trained. Further, rated familiarity strongly predicted classification responses ($r = .70$). When subjects said they were using rules or recollections, familiarity was just a partial predictor of grammaticality judgements; the actual grammatical status of the strings had additional predictive power, indicating in these cases subjects used knowledge sources other than familiarity – consistent with what subjects said they were doing. However, when subjects said, in contrast, that they were just guessing, or using intuition or familiarity, familiarity predicted grammaticality judgements without grammatical status having any extra predictive power. In these cases, familiarity seemed to be the sole determinant of subjects' responses that allowed discrimination between grammatical and non-grammatical strings.

We will now discuss the relationships we found between familiarity and awareness, familiarity and fluency, and familiarity and control.

Familiarity and awareness

Implicit learning occurs, in our lights, when a person has acquired unconscious knowledge – unconscious even when they are actively using it. For example, the weights in a neural network represent constraints that a person actively uses in classifying but the person may not in any way represent that they know each constraint. A person may use the constraint that 'MT' is a bigram in the training phase, but not be aware of this knowledge even as they use it to classify a string. Whittlesea has argued a different definition of implicit learning: It occurs when people learn knowledge for one purpose and do not know that they can use it for another. They may learn that MT is a bigram in the training phase, but not know that this knowledge is relevant to classifying a string as grammatical, even if they use it to do so. See Whittlesea and Dorken (1997) and Dienes and Berry (1997) for an exchange on our approaches to implicit learning. In many cases the two ways of putting it will agree. Consistent with Whittlesea's formulation, we have shown that sometimes people consciously know a rule (or consciously have a memory) but also consciously think it is not relevant and hence say that they are completely guessing (Dienes & Scott, 2005; Scott & Dienes, 2008). However, and there may be no disagreement with Whittlesea here, people sometimes apply knowledge when they consciously believe they are producing responses literally at random: It doesn't seem to them that they are using knowledge thinking it is irrelevant, it seems to them, so they say, that they are using nothing at all (Scott & Dienes, 2008). To illustrate a stronger contrast between the two definitions, we also showed that unconscious knowledge can be formed even when it is learned for the very purpose for which it is tested: People asked to find the rules of a grammar in the training phase, can still learn to classify test items as rule-governed or not while believing they are purely guessing or using intuition, and not using memory or rules at all (Dienes & Scott, 2005; Scott & Dienes, 2008). That is, while the acquisition of unconscious knowledge can be incidental, it need not be. Further, Whittlesea's approach might imply that one can only have implicit knowledge when it is an indirect consequence of something attended to and consciously encoded as such. Yet, important as it is, attention does not completely determine the acquisition and application of implicit knowledge. People systematically misrepresent the perceptual variables controlling interception of balls, so they cannot be consciously attending to those variables as such, even as they successfully use those variables to intercept balls (Reed et al., 2010).

Subjective familiarity may be the meeting ground of believers and sceptics of unconscious knowledge. A defining feature of implicit learning is the phenomenology: When people apply their knowledge they feel like they are guessing or using intuition or familiarity. But if in all these cases people had conscious feelings of familiarity, and these conscious feelings accounted for people's ability to discriminate grammatical from non-grammatical strings, the sceptic can say 'You see, there are conscious feelings, the knowledge is expressed consciously even in the defining cases of implicit learning' (cf. Berry et al., 2008; Shanks, 2005). Consistently, Reber, who has always been a strong advocate of unconscious knowledge, has long said that people know *that* they know something, they just do not know *what* they know (e.g., Reber, 1989). Norman et al., (2007) expressed this sentiment by suggesting familiarity was a prime example of *fringe* consciousness: A conscious feeling acting as a pointer to the existence of unconscious knowledge. They argued such fringe feelings mediated responses on implicit learning tasks.

What does the evidence say? As indicated above, Dienes and Scott (2008) found that when people said that they were guessing or using intuition or familiarity, people's rated feelings of familiarity accounted for grammaticality judgements. Even when people said they were responding randomly, that they were not using feelings of familiarity, their rated feelings of familiarity nonetheless accounted for their grammaticality judgements. In these cases, it seems people were not only unaware of the relevance of their feelings of familiarity, they were unaware that they were even using them. Nonetheless, people, on the face of it, had conscious feelings that contained relevant information.[2] Both sceptics and believers can be happy.

However, if familiarity is not intrinsically conscious one might expect to find cases where unconscious familiarity controls performance. We present evidence for two such cases. Scott and Dienes (2010b) used a transfer paradigm where people were tested on a different letter set than they were trained on (but using the same grammar). In this case, we found that it was only when people said they were responding randomly that they performed above chance (60%); when people said they were using any other basis such as rules or recollection they were responding at chance. Importantly, when people said they were responding randomly, their rated feelings of familiarity did not completely account for their ability to discriminate grammatical from non-grammatical items; the grammatical status of the items had predictive power above and beyond rated familiarity in accounting for grammaticality judgements. What other source of knowledge could people have been using? People claim not to be using any source at all. We suggest (because we believe that implicit learning is based on a neural network) that the source is unconscious familiarity, familiarity not expressed in people's ratings. Indeed, we showed that beyond rated familiarity it was the presence of novel bigrams and trigrams that predicted people's grammaticality judgements and accounted for their additional accuracy above that achievable through rated familiarity. Novel chunks would be expected to influence familiarity; but they didn't in this case influence conscious familiarity. Why does unconscious familiarity outperform conscious familiarity in this case? We will argue below that the similarity function defining familiarity can be changed by the subject's focus of attention. It may be that when subjects allow their responses to happen 'at random' the similarity function is less influenced by conscious hypotheses that, given the difficulty of the transfer task, are irrelevant or wrong.

Wan, Dienes, and Fu (2008) presented evidence for the control of the similarity function defining familiarity (we discuss this study further below). When people were trained on two grammars, they could successfully choose to endorse one or other grammar. The important point for here is that while rated familiarity distinguished the grammars when people said they were using familiarity to do so, it did not when people said they were literally guessing. Again in this case, we postulate people were guided by unconscious familiarity, the unidimensional output of neural networks primed to be differentially biased to one grammar rather than another. Just as allowed by higher order theories of consciousness (a version of which Whittlesea subscribes, as noted above), people can be guided by first-order representations alone.

Whether or not the feeling of familiarity is conscious, the knowledge on which familiarity is based can be conscious or unconscious. Whether or not this knowledge consists of exemplars, processing episodes or something more abstract, we have asked subjects to indicate trial by trial whether they are aware of the basis of their judgements. We have consistently found that people use familiarity even when claiming they have no idea of the basis of their grammaticality judgements, nor of the basis of their familiarity judgements. The description of familiarity as a fringe feeling by Norman et al., (2007) seems apt. It can point to unconscious structural knowledge. Nonetheless Scott and Dienes (2009) suggest that, as exposure to training and test material continues, people can use differential familiarity of parts of a string as a basis for formulating hypotheses with a separate system capable of dealing with hypotheticals. In this way, familiarity can be used to gradually explicate one's structural knowledge (cf. Matthews et al., 1989). (On multiple systems, Scott & Dienes, 2009, could be contrasted with e.g., Whittlesea & Dorken, 1994, 1997.)

Familiarity and fluency

A dominant view is that familiarity is based on fluency of processing, that is, the speed with which it takes processing to complete. Jacoby and Dallas (1981) suggested that when processing an item with relative ease, people may attribute this to having encountered the item before and experience it as familiar. Whittlesea and Williams (2000) elaborated this position by arguing that it was not fluency itself but surprising or discrepant fluency that can lead to feelings of familiarity.

Applying the fluency account of familiarity to models of implicit learning requires that an additional assumption be made. For example, in the context of connectionist models it requires the assumption that the constraint-satisfaction signal be correlated with the speed that the network arrives at its judgement. While there is no compelling a priori reason why this assumption should be true (Bullinaria, 1994), it has proved effective in modelling human responding in the sequential reaction time task (Cleeremans, 1993). However, the implicit knowledge expressing itself in judgement tasks (like artificial grammar learning) may be quite different from implicit knowledge expressing itself in perceptual motor tasks like the sequential reaction time task (Seger, 1994). Further, fluency is just one a priori possible basis for familiarity. Fluency is a property of the representational vehicle of one's knowledge: How fast the representation takes to form. Another vehicle property that could be used as a basis for familiarity is the total amount of activation. Conversely, familiarity may be carried purely by the content of a representation, rather than any content-irrelevant vehicle property. Indeed, in most computational models of implicit learning, processing fluency is literally the same for grammatical and non-grammatical stimuli, if not longer for grammatical rather than non-grammatical items in chunking models (Boucher & Dienes, 2003). In typical models (see Cleeremans & Dienes, 2008), familiarity is carried purely by a representation having the required content. Nonetheless, there has been a persistent assumption in the artificial grammar literature that item familiarity equates to fluency, more specifically perceptual fluency (e.g., Buchner, 1994).

Buchner (1994) employed a perceptual clarification task to explore the potential relationship between perceptual fluency and responding in AGL. The perceptual fluency of test strings was assessed based on the speed with which participants were able to make out the strings as they gradually emerged from behind a mask. The results provided mixed evidence; while grammatical strings were identified more rapidly than ungrammatical strings these differences in fluency were not found to correlate with participants' grammaticality decisions. Further the grammatical and non-grammatical strings were not counterbalanced and relevant item properties that would plausibly influence fluency were not controlled.

A relationship between perceptual fluency and judgements of grammaticality has been observed under conditions where fluency has been artificially manipulated so causal conclusions can be drawn. Kinder, Shanks, Cock, and Tunney (2003) employed a perceptual clarification task similar to that used by Buchner (1994) but manipulated perceived fluency by having half the strings clarify at a faster rate. Consistent with an influence of fluency, faster clarifying strings were endorsed as grammatical more often than slow clarifying strings and the authors concluded that fluency was the default mechanism for making grammaticality judgements.

Scott and Dienes (2010a) showed that with counterbalanced grammatical and non-grammatical items, naturally occurring fluency as measured in Buchner's (1994) perceptual clarification task was not related to grammaticality. Further, fluency bore a tiny albeit significant relation to rated familiarity (r =.08). Familiarity was related to various structural variables, but in a way that was independent of measured fluency. In addition, when fluency was manipulated, following the same procedure as Kinder et al. (2003) it only influenced familiarity and grammaticality decisions when subjects were given limited time to make judgements (as in Kinder's, 2003, procedure). Otherwise, in the normal conditions of artificial grammar learning experiments, even fluency manipulated to be 15 times the difference observed to naturally occur between Buchner's grammatical and non-grammatical stimuli, influenced neither familiarity ratings nor grammaticality judgements. Scott and Dienes (2010b) further showed that the limited influence of fluency extends to transfer conditions; fluency again bears a tiny, albeit significant, relation to rated familiarity (r =.04) and is unrelated to grammaticality. In sum, fluency carries no useful information in the artificial grammar learning task and people barely use it either for grammaticality judgements or as a basis of feelings of familiarity. These conclusions are consistent with Whittlesea and Leboe (2000) who showed with a different classification task and fluency manipulation that when participants can exploit either fluency or structural similarity to make classification judgements that they reliably favour the latter (cf. also Johansson, 2009). Perceptual fluency is a *dumb heuristic* influencing responding only in the absence of actual implicit knowledge (Higham, unpublished).

Familiarity and control

In an elegant series of experiments, Whittlesea and his colleagues have shown that the structural level to which a subject attends (e.g., letters, syllables, words) changes the structural properties to which a person is sensitive in classifying stimuli (Whittlesea & Dorken, 1993; Whittlesea & Wright, 1997). Higham and Vokey (2004) showed that the duration with which an item is briefly displayed as a prime before presenting it for judgement influences rated familiarity. That is, it is likely that the constraints determining familiarity can be influenced by the focus of attention at training and test and by other manipulations independent of the training phase. These considerations lead one to predict that familiarity should be partially controllable, even if one is unaware of its basis: The structural features upon which it is based may shift with context. Indeed, connectionist networks are notoriously context-sensitive, so the constraint satisfaction signal should be sensitive to context. Thus, imagining one context may make the items from that context differentially familiar.

However, the dominant approach to familiarity does not allow the conclusion that familiarity could be controllably higher for one context rather than another. Jacoby (1991) defines familiarity to be that memorial process that does not allow control. In that light, the artificial grammar learning literature contains an intriguing set of results: Dienes, Altmann, Kwan, and Goode (1995) found that when subjects are trained on one artificial grammar after another, they can choose almost perfectly which grammar to use in endorsing grammatical items. That is, they can choose which context – the first or second set of training strings – is used in determining their choices in the test phase. On the Jacoby approach, familiarity, by definition, can not be used to achieve this discrimination between the grammars. Scott and Dienes (2008) by contrast found that people use, and say they use, familiarity as the main means for distinguishing grammatical from non-grammatical items when trained on one grammar. So can familiarity be used to discriminate two grammars or can't it?

Wan et al. (2008) trained people on two grammars and took familiarity ratings, plus attributions from subjects concerning the basis of their judgements (random, intuition, familiarity, recollection, rules). They found that people could discriminate between the grammars almost perfectly and they could do so whatever the perceived basis of their judgement. When people say they are using recollections or rules, it is not surprising that they can discriminate. However when people say they are using familiarity they can also discriminate. Crucially, when they say they are using familiarity they almost exclusively pick the string with the highest rated familiarity. Further, the rated familiarity for test strings consistent with their chosen grammar was greater than that for strings from the other (counterbalanced) grammar. Familiarity, subjectively defined, is sensitive to intentions and can play a key role in strategic control.

We found no tendency for the to-be-ignored grammar to be endorsed more than entirely non-grammatical items: No evidence for familiarity in the Jacoby sense. In contrast, Higham, Vokey, and Pritchard (2000) using more confusable artificial grammars found substantial endorsements of a to-be-ignored grammar compared to non-grammatical strings: On the Jacoby approach these confusions are based on familiarity. This result allows an interesting contrast between familiarity subjectively defined as we do and the Jacoby approach. The Jacoby approach would classify knowledge as based on a familiarity process even if subjects stated on each trial that their response was based on conscious rules or recollection, and discrimination between the grammars failed simply because the rules were not perfectly discriminating. Thus, a mental state is one of familiarity or not depending on a specific state of affairs external to the subject, namely, on whether some strings are classified as grammar A or B according to an experimenter's grammars – which might have been different and thus could have been classified otherwise. This is an example of externalism in the philosophy of mind (see e.g., Wilken et al., 2008, entries on 'externalism' and 'representations, problems'). If the experimenter happened to classify all the new strings in the same way as the subject, there would be no familiarity. According to our approach, a state being one of subjective familiarity depends not on how the experimenter classifies strings, only on the mechanisms operating within the subjects' mind (namely, only on whether a mechanism was used that represents oldness, regardless of whether this mechanism makes the same choices as the experimenter). While this argument biases our own preferences for a definition of familiarity, it should not be obvious that it should. Externalism as a view concerning the content of mental states is the current dominant (though not exclusive) view amongst philosophers. Jacoby's definition applies externalism in an interestingly different way, to the type of mental state one is in independent of its content;

in fact, by Jacoby's approach, the very conscious status of one's mental states can be so externally determined.

Conclusion

In summary, we hope we have drawn out arguments for the role of familiarity in implicit learning that interweave with Whittlesea's ideas in interesting ways (even if in ways he wouldn't approve). But hopefully that makes our conclusions surprisingly easy on the mind given what they are: That familiarity is typically conscious but can be unconscious; that it is often a conscious fringe feeling signalling that there exists further unconscious knowledge; that it might bring sceptic and believer together in harmony; that in cases where one has structural knowledge fluency plays almost no role in determining familiarity; and that familiarity can be controlled by intentions and used to control choice of items from different contexts, even as one remains unaware of its basis.

Notes

1. In the second sense of familiarity, familiarity can be unconscious. This may strike some readers as odd. How can a feeling be unconscious? According to higher order theories, any mental state, including a feeling, is unconscious unless you are aware of having it. Thus, a blindsight patient can see – but the seeing is unconscious because the person is unaware of seeing. Operationally, unconscious familiarity could allow a person to discriminate which of two stimuli were old while the person believes they are literally guessing.
2. Being conscious of using familiarity does add something to using it only unconsciously: In Experiment 3 of Scott and Dienes (2008), when people said they were using familiarity, the average rated familiarity on a 100-point scale was 56 (SE = 2.5), significantly higher than the average familiarity when people said they were responding literally randomly (47, SE = 3.6) or even than when people said they were using intuition (46, SE = 2.5), p's < .05. More importantly, the mean correlation between familiarity ratings and grammaticality judgements was higher when people said they were using familiarity (r = 0.56, SE = .07) than when they said they were responding randomly (r = .33, SE =.10), which did not differ from when people said they were using intuition (r = .38, SE = .07). Thus, consciously using familiarity involves more use of feelings of familiarity than its unconscious use.

Part III

Inferential Processes and Recollection/Retrieval

6
The Constructive Nature of Recollection

Tamara L. Ansons and Jason P. Leboe

Introduction

In dual-process theories of recognition (Jacoby, 1991; Jacoby & Dallas, 1981; Mandler, 1980; Tulving, 1985), people have two bases for identifying a stimulus as previously encountered. *Old* judgements can arise from either a feeling of *familiarity* for the stimulus or from successful *recollection* of the context surrounding a prior exposure to that stimulus (see Yonelinas, 2002, for a review). This dual-process approach tends to orient researchers toward differences between the processes of familiarity and recollection. For example, feelings of familiarity are commonly seen as originating from a heuristic attribution process that is prone to error (Jacoby & Dallas, 1981; Jacoby & Whitehouse, 1989; Whittlesea, 1993; Whittlesea & Williams, 1998, 2000). In contrast, recollection is often seen as the more dominant and accurate basis for making recognition judgements. Indeed, rather than generated by a faulty heuristic attribution process, recollective experiences are seen as arising from a successful search of memory and a direct retrieval of details about the past into consciousness.

Whittlesea's *Selective Construction and Preservation of Experience* (SCAPE) model of memory (Whittlesea, 1997) was unique in challenging these assumptions about the distinction between familiarity and recollection. This account conceived of all instances of remembering as a constructive, inferential process. That is, whether remembering involves a non-specific feeling of familiarity or a vivid recollection, it is the interpretation of current processing as originating from some prior experience that defines the subjective experience of remembering. A comparatively innovative, constructive approach to remembering can also be found in the *Source Monitoring Framework* (*SMF*; Johnson et al., 1993; Lindsay, 2008). By noting this basic similarity between all instances of remembering, the SCAPE account provides a motive for investigating how people infer details of their past based on the qualitative aspects of their recollections. That is, the account makes predictions about the heuristics that people rely on when using their recollections to make recognition judgements and when they employ them. We test the predictions outlined by this account by examining the conditions that guide individuals to use recollections inferentially when forming judgements about their past.

The current study

Despite the implications of the SCAPE approach to memory, very few studies have directly examined the recollection-based heuristics that people will use to guide their recognition judgements (for exceptions, see Dodson & Schacter, 2001; Kleider & Goldinger, 2006; Leboe & Whittlesea, 2002; Whittlesea & Leboe, 2000, 2003). In the experiments described below, we developed a procedure designed to cause some recognition test items to sponsor detailed recollective experiences, independent of whether those items were presented in the target list. Our experiments illustrate that, like familiarity, recollection can be an important source of error in remembering. This being so, the central question then becomes, what are the rules that determine *when* the generation of detailed recollections will contribute as a source of error in recognition judgements? The experiments described in the current chapter represent an initial step in acquiring a more comprehensive understanding as to how people use recollected detail to make inferences about the past. Specifically, our results indicate that, in the context of a recognition test, people's use of recollections reflect a metacognitive understanding of the principle of transfer-appropriate processing. We refer to this influence as a *resemblance of processing* heuristic, conceiving of this heuristic as a refinement of the resemblance heuristic described previously by Whittlesea and Leboe (2000). Whittlesea and Leboe's emphasis was on similarities between recognition or classification test items and the common structural or semantic properties of items used to construct a preceding study list. We consider the application of the resemblance heuristic to be much broader, involving not only similarities in the structural or semantic properties of stimuli, but also similarities between the mental processes engaged during encounters with stimuli in a study phase and the mental processes stimulated by the appearance of a stimulus during a test phase. It is our desire to orient the reader to the primacy of processing resemblance over stimulus resemblance that motivated our use of the label *resemblance of processing* heuristic in this context.

Experiment 1

Method

Participants

Thirty-one participants were recruited from Introduction to Psychology classes at the University of Manitoba.

Apparatus and stimuli

A MacIntosh G3 computer installed with FutureBASIC II 2.07 and connected to a keyboard and a 15-inch colour monitor was used to present the target words and record the participants' responses. A set of 900 moderate to highly imageable words, consisting of 180 sets of five semantically related nouns ranging between 3 and 12 letters in length, were generated for use in Experiment 1.

Procedure

Phase 1: Initial Encoding Phase

Prior to beginning this phase, participants were informed that they would be presented with a series of trials, each consisting of a word presented alone or as the right-hand item in a word pair. They were told that when the word appeared alone, their task was to read the word silently to themselves; however, when the word appeared as one member of a pair, they were to generate a mental image of the two objects interacting. To illustrate what was meant by the instruction, *generate an interactive image*, participants were provided with the following example: 'Suppose the two words were KITE – DOG. With these words, you could generate an image of a dog chasing a kite.' Participants were also told that singly presented words and the right-hand word of word pairs would form the basis of a later memory test. Using a similar procedure in a previously published study (Leboe & Ansons, 2006), we confirmed that interactively imaged items produced more vivid recollections with the presentation of those items at test, than items that were merely read before.

In a different random order for each participant, *target* words were presented either three times as a member of a different pair of semantically related words (the *3X-Imaged* condition), three times in isolation (the *3X-Read* condition), or once in isolation (the *1X-Read* condition). Forty target words were assigned to each of these three conditions, generating a total of 280 trials. These trials were displayed in a random sequence for 5 seconds each.

Phase 2: Recency Phase

Once the *Initial Encoding Phase* was complete, participants were informed that a series of word pairs would be presented at the centre of the screen and their task was to generate an interactive mental image of the two objects. Half of the target words from each of the conditions generated by the *Initial Encoding Phase* were randomly selected for presentation in the *Recency Phase*. These items appeared once for 5 seconds each, generating a sequence of 60 trials. Each target word appeared to the right-hand side of a semantic associate that did not appear in the preceding phase.

Phase 3: Test Phase

All of the target words from Phase 1 were presented in random sequence at the centre of the computer screen, generating a total of 120 trials.

Participants were informed that half of the words from the first phase were also presented during the second phase of the experiment, and that their task was to judge whether or not the target word was presented *recently*, during Phase 2 of the experiment. Participants made their response by pressing a button on the keyboard.

Within our experimental session, participants experienced a more controllable variant of the situation they find themselves during more traditional recognition tests. When encountering a recognition test item in those studies, the participant is faced with the challenge of deciding whether that item was recently encountered during the experimental session. If not, the alternative is that the item is merely something the participant has encountered in contexts other than the list they studied, during perhaps hundreds of experiences with that item outside of the laboratory. Our procedure was meant to achieve greater control over the remembering processes that stem from encounters with test items that originate outside of the context that participants are asked to base their recognition decisions upon. Sometimes the earlier processing of items within irrelevant contexts will sponsor detailed recollections during a subsequent recognition test, whereas sometimes those earlier encounters will not sponsor much in the way of detailed recollections. Likewise, targets assigned to our *3X-Imaged* condition should sponsor more detailed recollections during the *Test Phase* of our experiment than targets assigned to our *3X-Read* and *1X-Read* conditions, given the abundance of evidence revealing the effectiveness of imagery encoding at promoting success in future recall (Leboe & Ansons, 2006; Lutz & Lutz, 1977; Paivio et al., 1966). As an initial test of our procedure, we designed Experiment 1 to determine whether recollection of an interactive image originating from the *Initial Encoding Phase* would be sufficient to convince participants to identify *3X-Imaged* items as having been presented recently, even for the subset of those items that did not actually appear during the *Recency Phase*. We expected that this would occur because the generation of interactive images characterized participants encoding of targets during the second phase of the experiment. It should be relatively difficult to reject items as having been presented recently when they sponsor the recollection of the same type of process that was engaged in response to all displays during the *Recency Phase*.

Results and discussion

The mean proportions of *recent* judgements for each condition are presented in Table 6.1, with between-participant standard errors provided in parentheses.

An initial Analysis of Variance (ANOVA) revealed a significant *Initial Encoding X Presentation Recency* interaction, $F(2, 60) = 4.85$, $MSe = .02$, $p < .05$, so we tested the effect of Initial Encoding separately for the *Recent* and *Not Recent* targets. For the *Recent* targets, participants were 14% more likely

Table 6.1 Proportion of *recent* judgements in Experiment 1 as a function of Initial Encoding (1X-Read/3X-Read/3X-Imaged) and Presentation Recency (Recent/Not Recent)

	Initial Encoding			
Presentation Recency	**1X-Read**	**3X-Read**	**3X-Imaged**	**Overall**
Recent	.61 (.03)	.58 (.02)	.72 (.03)	**.64 (.02)**
Not Recent	.20 (.04)	.27 (.04)	.45 (.03)	**.30 (.03)**
Overall	**.40 (.02)**	**.43 (.02)**	**.58 (.02)**	

to judge *3X-Imaged* words as presented recently than *3X-Read* words (.72 vs. .58), $F(1, 30) = 24.26$, $MSe = .01$, $p < .001$, and they were 11% more likely to judge *3X-Imaged* words as presented recently than *1X-Read* words (.72 vs. .61), $F(1, 30) = 12.33$, $MSe = .02$, $p < .005$. For *Not Recent* targets, *3X-Imaged* words were 18% more likely to be judged as presented during the *Recency Phase* than *3X-Read* words (.45 vs. .27), $F(1, 30) = 17.39$, $MSe = .03$, $p < .001$, and were 25% more likely to be judged as *recent* than *1X-Read* words (.45 vs. .20), $F(1, 30) = 32.65$, $MSe = .03$, $p < .001$. For both *Recent* and *Not Recent* targets, participants were about as likely to judge *3X-Read* and *1X-Read* items as appearing in the *Recency* phase, $p > .05$, in both cases.

In Experiment 1, during an *Initial Encoding Phase*, some words were initially encoded by generating interactive mental images, whereas others were merely read. Half of the imaged and read words for that initial phase then appeared during a *Recency Phase*, whereas the other half did not. We presume that generating an interactive image involving target words during the *Initial Encoding Phase* would sponsor more detailed recollections during the final test phase than merely reading targets during that phase. The consequence was that participants were more likely to identify *3X-Imaged* targets as being present during the *Recency Phase* than both *3X-Read* and *1X-Read* targets. This result illustrates how processes of recollection can readily be a source of error in the generation of recognition judgements, contrary to the more frequent treatment of recollection processes as serving as a more valid basis for making recognition judgements than familiarity. The experience of familiarity is prone to error because it is an inferential basis for making recognition judgements. An item can generate a strong feeling of familiarity for reasons other than the presence of that item within a prior study list. As our Experiment 1 demonstrates, processes of recollection can lead to errors in recognition judgements for precisely the same reason.

It is perhaps quite sensible that our participants often made the error of thinking that *3X-Imaged* targets appeared in the *Recency Phase* even when they did not actually appear in that phase. Targets presented in the *Recency*

Phase were also interactively imaged. In consequence, the type of recollection stimulated by *3X-Imaged, Not Recent* targets would be perfectly consistent with the type of recollection that would be stimulated by targets actually presented during the *Recency Phase*. By now, comparable source confusion errors have been extensively documented across a range of contexts (for reviews see Johnson et al., 1993; Mitchell & Johnson, 2000). A major premise of Whittlesea's (1997) SCAPE account of memory is that people base their decisions about the past on inferences about the significance of qualitative aspects of what comes to mind during an act of remembering. In part, this proposal was inspired by other key remembering frameworks, such as the *SMF* (Johnson et al., 1993) and attribution approaches to remembering (Jacoby et al., 1989). The influence of the *SMF* is particularly relevant to the current study. Research guided by this approach to remembering has demonstrated that recollections originating from an imagined event may be mistaken as recollections originating from an event that actually occurred to the extent that they possess qualities that are typical of true remembrances, such as vividness of sensory detail (Johnson et al., 1988). Likewise, the results of our Experiment 1 could be seen as a form of source-monitoring error.

An alternative explanation for the results of Experiment 1 is that detailed recollections sponsored by prior generation of interactive images made *3X-Imaged* targets compelling candidates for generating *recent* judgements, regardless as to the nature of processing that occurred during the *Recency Phase*. We conducted Experiment 2 to investigate whether participants would also be more likely to judge previously imaged targets as *recent* even when they were not required to generate any interactive images during the *Recency Phase*. In Experiment 2a, targets presented in the *Recency Phase* were presented alone and participants were required to simply read them. In Experiment 2b, targets presented in that phase were also presented alone, but participants were instructed to generate a mental image of that object.

Experiments 2A AND 2B

Method

Participants

Twenty undergraduate students participated in Experiment 2a and thirteen students participated in Experiment 2b.

Apparatus and stimuli

The same apparatus and stimuli that were used in Experiment 1 were also used for Experiments 2a and 2b.

Procedure

The procedures of Experiments 2a and 2b were identical to the procedure used in Experiment 1 except for the way participants were instructed to encode target words during the *Recency Phase*. During that phase in Experiment 2a, participants were asked to simply read each word aloud as they appeared for 5 seconds each on the computer screen. During the *Recency Phase* of Experiment 2b, participants were asked to use the 5 s time period that the target persisted on the screen to generate a mental image of that object.

Results and discussion

Experiment 2a

The mean proportion of *recent* judgements for each condition of Experiments 2a and 2b are presented in Tables 6.2 and 6.3, respectively, with between-participant standard errors for each condition provided in parentheses.

Our analysis revealed a main effect of *Initial Encoding*, $F(2, 38) = 8.10$, $MSe = .01$, $p < .005$, and a main effect of *Presentation Recency*, $F(1, 19) = 76.85$, $MSe = .04$, $p < .001$. In the most critical contrast, for which frequency of target presentation during the *Initial Encoding Phase* was equivalent, participants were

Table 6.2 Proportion of *recent* judgements in Experiment 2a as a function of Initial Encoding (1X-Read/3X-Read/3X-Imaged) and Presentation Recency (Recent/Not Recent)

	Initial Encoding			
Presentation Recency	**1X-Read**	**3X-Read**	**3X-Imaged**	**Overall**
Recent	.46 (.04)	.50 (.04)	.43 (.03)	**.46 (.03)**
Not Recent	.14 (.03)	.23 (.04)	.13 (.03)	**.17 (.03)**
Overall	**.30 (.03)**	**.37 (.04)**	**.28 (.03)**	

Table 6.3 Proportion of *recent* judgements in Experiment 2b as a function of Initial Encoding (1X-Read/3X-Read/3X-Imaged) and Presentation Recency (Recent/Not Recent)

	Initial Encoding			
Presentation Recency	**1X-Read**	**3X-Read**	**3X-Imaged**	**Overall**
Recent	.76 (.04)	.82 (.05)	.70 (.04)	**.76 (.04)**
Not Recent	.17 (.04)	.27 (.06)	.07 (.03)	**.17 (.04)**
Overall	**.47 (.03)**	**.54 (.05)**	**.39 (.03)**	

9% more likely to judge *3X-Read* words as presented recently than *3X-Imaged* words (.37 vs. .28), $F(1, 19) = 14.95$, $MSe = .01$, $p < .005$. Thus, in contrast to the outcome of Experiment 1, the recollection of having previously engaged in interactive imagery discouraged participants from judging targets as having been presented in the *Recency Phase*, even when those targets actually were presented during that phase. Participants were also 7% more likely to judge *3X-Read* words as presented recently than *1X-Read* words (.37 vs. .30), $F(1, 19) = 8.83$, $MSe = .01$, $p < .01$, but they were about equally likely to judge *3X-Imaged* and *1X-Read* targets as presented during the *Recency Phase*, $p > .05$.

Experiment 2b

Our analysis of the results of Experiment 2b yielded a main effect of *Initial Encoding*, $F(2, 24) = 11.08$, $MSe = .01$, $p < .001$, and a main effect of *Presentation Recency*, $F(1, 12) = 140.20$, $MSe = .05$, $p < .001$. Consistent with the outcome of Experiment 2a, participants were 15% more likely to judge *3X-Read* words as presented during the *Recency Phase* than *3X-Imaged* words (.54 vs. .39), $F(1, 12) = 14.83$, $MSe = .01$, $p < .005$, and 8% more likely to judge *1X-Read* words as presented during the *Recency Phase* than words that were *3X-Imaged* (.47 vs. .39), $F(1, 12) = 9.91$, $MSe = .004$, $p < .01$. In addition, participants were 7% more likely to judge *3X-Read* words as being presented recently compared to *1X-Read* words (.54 vs. .47), $F(1, 12) = 5.66$, $MSe = .01$, $p < .05$.

In Experiments 2a and 2b, we observed that participants were no longer more likely to judge words that were *3X-Imaged* during the *Initial Encoding Phase* as presented recently when words were merely read during the *Recency Phase*. In Experiment 2a, compared to words that were *3X-Read* during the *Initial Encoding Phase*, *3X-Imaged* words that possessed a higher likelihood of sponsoring rich and detailed recollections during the test phase were *less* likely to be identified as encountered recently. Given that words from the *Recency Phase* were merely read, it is sensible that the recollection of interactive images created during the *Initial Encoding Phase* would not be considered strong evidence for the presentation of targets during that phase. Likewise, even though imagery generation did occur in the *Recency Phase* of Experiment 2b, the type of imagery was not precisely the same as the interactive imagery generation that characterized earlier processing of *3X-Imaged* items. However, since targets that were interactively imaged before were as likely to appear in the *Recency Phase* as not, the most rational strategy would have been for participants to simply treat the recollection of interactive images as irrelevant to the judgement they were asked to make. Participants in our Experiments 2a and 2b did not adopt this optimally rational strategy. Instead, the recollection of previously generated images served as a basis for rejecting targets as having appeared in the *Recency Phase*. Put another way, in Experiment 1, when the *Recency Phase* consisted of interactive image generation, the unsurprising consequence was that participants often made errors of *commission*, resulting in participants frequently making the error of judging *3X-Imaged, Not Recent* targets as presented *recently*. Experiments 2a and 2b represent a reversal of this bias, simply through the absence of interactive imagery generation during the *Recency Phase*. In these cases, the outcome was an increase in errors of *omission*; the bias to reject targets that sponsored recollection of interactive images resulted in a greater tendency to incorrectly identify *3X-Imaged, Recent* targets as *not* presented recently. Together, we suggest that these results reflect participants' expectation that recollective processes should be compatible with the type of processing that occurred during the critical to-be-remembered event. We interpret this influence of processing expectations on recognition judgements in the current studies as representing a heuristic application of the TAP principle; a *resemblance of processing* heuristic in the use of recollective processes to guide judgements about past events.

Experiment 3

Method

Using a similar procedure to Experiment 2a (items merely read during the *Recency Phase*), in conducting Experiment 3 our goal was to determine a possible boundary condition for participants' reliance on the *resemblance of processing* heuristic. In particular, we were interested in the circumstances that would cause participants to abandon their reliance on this heuristic. Consider that, in Experiment 2, the proportion of interactively imaged items that appeared in the test phase was .33, whereas .33 of test items were read once and .33 of test items were read three times during the *Initial Encoding Phase*. In consequence, a relatively low proportion of test items sponsored detailed recollections of having been interactively imaged. Thus, when participants recollected an interactive image at test, it violated expectations for two reasons: 1) that type of recollective process was not consistent with processing that occurred in the *Recency Phase* and 2) most test items did not sponsor that type of recollection. We were curious as to what the outcome would be if *most* test items did sponsor a recollection of having previously generated interactive images. Under those conditions, recollecting an interactive image at test would violate expectations based on processes that occurred during the *Recency Phase*, but would be perfectly consistent with more general expectations for the set of items that appeared in the test phase. When most test items promote recollection of an interactive image, such recollections might not be as compelling a basis for rejecting test items as having appeared in the *Recency Phase*.

Consider the following illustration: a person encounters a woman on the street and their friend asks, 'Was she at the birthday party last week?' In response to the question, one possible outcome is that the person generates a recollection consistent with things that happen at birthday parties. Supposing that the woman missed that particular party, the person might incorrectly respond, 'Yes, she was!', if their recollection originated from some party other than the one that happened last week. This is the type of commission error that we observed in our Experiment 1. It is also the type of source confusion that has already received considerable attention from researchers (Lindsay, 2008).

As an alternative possibility, the question might prompt recollection of having encountered the woman in a shopping mall. Suppose that the woman was actually at the party; however, the person might incorrectly respond, 'No, she wasn't there.' based on the recollection of having encountered the women in a mall coming to mind. The inference, in that case, is that the person has been encountered before, but in some other context than at the party. It is this type of reasoning that provided the source of omission errors in Experiment 2 (rejection of *Recent* items as presented during the *Recency Phase* when they are accompanied by detailed recollections). Now suppose that the question prompts recollection of having encountered the woman on a university campus *and* most of the people one knows are expected to sponsor that type of recollection (a situation many professors and students find themselves in). It might be more difficult to reject the woman as having been at the party based on a campus-based recollection; a frequent association between people and campus might cause such recollections to be perceived as a constant that would not be a sensible basis for excluding people as having also been present in other contexts. In Experiment 3, we tested this notion by manipulating the proportion of words that would sponsor detailed recollections at test. If participants are able to flexibly adapt their use of the *resemblance of processing* heuristic in a way that is dependent on their expectations about the processing of test items, the use of detailed recollections should not be used exclusively as evidence to reject items. Instead, as with the example discussed above, when detailed recollections occur frequently, this type of processing might be treated as a constant and should no longer be used as a basis to reject items as having been presented recently.

Participants

Twenty-one students participated in the *.75-Imaged* condition and 22 participated in the *.25-Imaged* condition.

Apparatus and stimuli

These were the same as in the preceding experiments.

Procedure

The procedure of Experiment 3 was nearly identical to the procedure used in Experiment 2a; however, we eliminated the *1X-Read* condition so that participants either generated interactive images (*3X-Imaged* condition) or read target words three times (*3X-Read* condition). Participants in the

.75-Imaged condition generated interactive images for 75% of the target words and merely read the remaining 25% of the target words. Participants in the *.25-Imaged* condition generated interactive images for 25% of the target words and merely read the remaining 75% of the target words. Next, participants completed the *Recency Phase*. For all participants, half of the words presented during the *Recency Phase* were previously interactively imaged and half of the words were previously read. There were a total of 80 target words presented in the *Recency Phase*. As a consequence of the proportion manipulation applied to the first phase, 40 (or 1/3) of the previously imaged words and all 40 of the previously read words were presented in Phase 2 for the *.75-Imaged* condition. The reverse was true for the *.25-Imaged* condition. Finally, participants completed the *Test Phase*, during which they were presented with all of the target words from the *Initial Encoding Phase* and were asked to judge which of those targets appeared in the *Recency Phase*.

Results and discussion

The mean proportions of *recent* judgements for each condition are presented in Table 6.4 with the between-participant standard error for each condition provided in parentheses.

Due to the proportion manipulation, for the *.75-Imaged* condition all read words were presented in Phase 2, and for the *.25-Imaged* condition all imaged words were presented in Phase 2. With missing cells for *Not Recent* items, we focused our analysis on the proportion of *Recent* items judged as recent as a function initial encoding and the proportion of previously imaged items. The critical result of Experiment 3 was a significant interaction between the proportion of words imaged during the *Initial Encoding Phase* and the type of *Initial Encoding*, $F(1, 41) = 137.09$, $MSe = .01$, $p < .001$. In the *.25-Imaged* condition, participants were 25% more likely to judge *3X-Read, Recent* words as having appeared in the *Recency Phase* than *3X-Imaged,*

Table 6.4 Proportion of *recent* judgements in Experiment 3 as a function of the proportion of words imaged in the Initial Encoding Phase (.25-Imaged/.75-Imaged), Initial Encoding (3X-Read/3X-Imaged) and Presentation Recency (Recent/Not Recent)

	Initial Encoding Phase				
	.25-Imaged		**.75-Imaged**		
Presentation Recency	**3X-Read**	**3X-Imaged**	**3X-Read**	**3X-Imaged**	**Overall**
Recent	.56 (.03)	.31 (.03)	.34 (.02)	.49 (.04)	**.42 (.02)**
Not Recent	.25 (.02)			.20 (.02)	**.23 (.01)**
Overall	**.40 (.02)**			**.35 (.03)**	

Recent words (.56 vs. .31), $F(1, 21) = 102.02$, MSe=.01, $p < .001$. This result is similar to the outcome of Experiment 2a when the proportion of previously imaged words appearing at test was also relatively low. By contrast, in the *.75-Imaged* condition, participants were 15% more likely to judge *3X-Imaged, Recent* words as having appeared in the *Recency Phase* than *3X-Read, Recent* words (.49 vs. .34), $F(1, 20) = 41.30$, MSe=.01, $p < .001$.

The results of Experiment 3 confirmed our suspicion that participants would abandon their reliance on a *resemblance of processing* heuristic when most of the items presented at test sponsored interactive imagery generation. Indeed, when 75% of test items were imaged in the *Initial Encoding Phase*, participants were actually *more* likely to judge previously imaged items as presented during the *Recency Phase* than previously read items. This bias is in direct conflict with the type of processing participants engaged in during the *Recency Phase*. Participants merely read words encountered in that phase, so participants should not have expected recollection of interactive images at test to be diagnostic of the prior inclusion of words in that phase. In the *.75-Imaged* condition, perhaps the frequency of interactive imagery generation at test compelled participants to abandon their reliance on a *resemblance of processing* heuristic in favour of a heuristic centred on recollection vividness. Mainly, detailed recollections were treated as a constant and were no longer used as a basis to reject items as having been presented during the *Recency Phase*. In fact, the expectation of this type of processing lead participants to reject previously read words at a higher rate than previously imaged ones.

Returning to the example described above ('Was she at the party?'), failure to generate campus-based recollections might be used as a basis for rejecting the woman in question as having attended the party. If most of the people one knows can be expected to sponsor campus-based recollections, those who fail to do so may be rejected as having been present in other context as well ('It seems that I don't really know that person very well, so she probably was not at the party.'). Although not reflecting reliance on the *resemblance of processing* heuristic that guided recognition judgements in Experiments 1 and 2, processing expectations are still at the root of this type of inference. When there was a general expectation for test items to sponsor the recall of interactive images in Experiment 3, previously read items may have been rejected at a higher rate because they violated this expectation.

Concluding remarks

Dual-process approaches to recognition (Jacoby & Dallas, 1981; Mandler, 1980; Tulving, 1985) have assigned the recollection process a special status. Unlike recognition judgements guided by a vague sense of familiarity for the stimulus, recollection permits conscious access to details about the context in which that stimulus was encountered. The result is that recognition judgements based on recollection are often more valid and less prone to error than recognition judgements based solely on familiarity. Development of the process-dissociation procedure (Jacoby, 1991) further highlighted this typically dominant role of the recollection process in supporting accurate recognition judgements. That framework placed emphasis on the corrective function of recollection; participants may use their recall of contextual details to avoid claiming a recognition test item as *old*, even when that item generates a strong feeling familiarity. Recollection provides a useful basis for rejecting a stimulus as having appeared in some context ('She wasn't at the party.') by providing knowledge about the presence of the stimulus in some other context ('I remember seeing her at the mall.').

Even so, the role of recollection in promoting accurate recognition judgements can be overstated. It is possible for the information provided by the recollection process to be insufficient for determining whether some stimulus was encountered in one or another context (Gruppuso et al., 1997). It is also possible for people to use the familiarity they experience when presented with a stimulus as a basis for inferring the context in which that item occurred (Diana et al., 2008; Dodson & Johnson, 1996). Evidence also reveals that false recollections can occur (Brainerd & Reyna, 2005; Brainerd et al., 2001; Higham & Vokey, 2004), and that experiencing a recollection as a true reflection of the past depends on the application of heuristics that are prone to error (Kurilla & Westerman, 2008; Leboe & Whittlesea, 2002; Lindsay & Kelley, 1996; Whittlesea & Leboe, 2000, 2003). Supported by these studies, the inferential basis of recollection serves to blur the clear distinction that is often presumed to exist between the processes of recollection and familiarity (see also Kurilla & Westerman, this volume). Nevertheless, the more typical conception of remembering as strictly dichotomous (familiarity is a heuristic, error-prone basis for making recognition judgements; recollection is direct retrieval of the details surrounding an experience with a stimulus), has discouraged research into the inferential origins of recollective experiences, at least in the context of the recognition task. Investigations into how accurate recollections form the basis of inferences about past events are also extremely rare (but see Dodson & Schacter, 2002; Kleider & Goldinger, 2006, for notable exceptions).

In the experiments described in the current chapter, we tested whether people would use detailed recollections when making judgements about the past, even when the content of those recollections was non-diagnostic of the remembering task they were asked to perform. Under those conditions, participants used their recollection of interactive images generated in an *Initial Encoding Phase* to infer either the presence or absence of items in a separate *Recency Phase*, depending on whether the type of processing engaged in during the *Recency Phase* was consistent with the recollection of interactive images at test. We described this bias as reflecting a metacognitive understanding of the principle of transfer-appropriate processing; that is, participants will rely on a *resemblance of processing* heuristic when using detailed recollections to decide on the context in which some stimulus occurred (see Westerman et al., 2002, for evidence that a similar inferential process can determine whether people will rely on feelings of familiarity to make *old* judgements). We also observed that application of this heuristic occurred mainly when test items were less likely to promote recollection of interactive images, whereas participants abandoned reliance on this heuristic when a high proportion of test items were capable of promoting such recollections. Beyond the current results, however, our broader goal is to encourage a more central focus for the constructive nature of recollection and recollection-based remembering by memory researchers. A general note of caution: even when the recollective process happens to yield accurate details about the past, that process can still provide the basis for remembering errors.

Note

We would like to thank Shohan Illsley and Geoffrey J. Palmer for their assistance with data collection.

7

Inferential Processes in Subjective Reports of Recollection

Brian P. Kurilla and Deanne L. Westerman

Preface

Bruce Whittlesea has been instrumental to the process of developing a theory of memory that attempts to explain the relationship between implicit and explicit processes. His research findings on processing fluency, his novel insights regarding attribution processes in memory, and the creativity of his experimental methods have greatly influenced our work. We are grateful for his contribution to the field and for the inspiration that his research has provided to us as we continue to try to understand the basis of recognition memory and the conscious experience of remembering.

Introduction

The ability to distinguish between stimuli that have previously been experienced and those that are novel is a fundamental aspect of human functioning and provides a foundation for higher level cognitive processes. Evidence of some form of recognition memory is present from very early in life (e.g., Fagan, 1970) and accuracy declines relatively little across the lifespan (see Craik & Jennings, 1992). The processes and brain structures that contribute to recognition memory have been the subject of an enormous amount of research from a variety of scientific perspectives. Most current theories of recognition memory propose the operation of two separate and independent processes: namely, recollection and familiarity (e.g., Jacoby, 1991; Mandler, 1980; Yonelinas, 1994, 2002). According to such dual-process theories, recollection and familiarity contribute very different types of information to a recognition decision. Whereas familiarity is thought to provide only a vague sense of undifferentiated recency, recollection is thought to allow retrieval of specific contextual details from the original encoding event, which is generally taken to mean that recollection is more useful than familiarity on tasks that require memory for source (Yonelinas, 1999). Consequently, the subjective experience that accompanies recognition based on familiarity is generally thought to be very different from that which accompanies recognition primarily based on recollection (Gardiner, 1988; Tulving, 1985).

In addition to the contributions of familiarity and recollection, recognition decisions are also influenced by heuristics, strategies, and biases. For example, the *fluency heuristic* is the finding that stimuli that are processed relatively fluently tend to receive more positive recognition judgements presumably because of an attributional process that interprets fluency as evidence of past experience (Jacoby & Dallas, 1981). Traditionally, such attributional processes are thought to influence the perception of familiarity. The idea that attributional processes contribute to the experience of recollection, although advanced by some researchers (e.g., Leboe & Whittlesea, 2002; Whittlesea, 2002, 2003), has gained much less traction within the field of recognition memory. This is curious given the general consensus that memory is subject to a great deal of reconstruction and inference (e.g., Schacter, 2001), and given that some of the most well known studies in cognitive psychology are those that have demonstrated the 'recollection' of events that did not occur. For instance, in the now famous study by Loftus and Palmer (1974), participants were more likely to falsely recollect having seen broken glass at the scene of an automobile accident when they inferred that the cars had been travelling at higher speeds. Similarly, many of the participants in Neisser and Harsch's (1992) study of flashbulb memories reported clear recollection of details that did not occur. These experiments, and many others, have shown beyond dispute that people sometimes experience recollection of details and possibly entire events that never occurred (see Brainerd & Reyna, 2005). However strong this consensus may be in general, it has not been reflected in current theories of recognition memory. On the contrary, very little attention has been paid to false claims of recollection or the ways in which attributional processes may contribute to a subjective sense of recollection in recognition memory. To the extent to which false recollection has been studied, the examination has been focused on how semantic similarity among items at encoding impacts false recollection (e.g., Payne et al., 1996). In this chapter we seek to review studies that have discovered *test* conditions that elevate reports of recollection in laboratory-based recognition tasks. These studies suggest that recollection, like familiarity, is a product of inferential processes and suggest that a subjective feeling of recollection can be influenced by a variety of factors.

In what follows, we review evidence from several studies that have examined the importance of various heuristics in determining people's reports of conscious recollection. Next, we will attempt to generalize from these findings and to formulate a compelling argument for why claims of conscious remembering might also sometimes be influenced by *perceptual fluency*, that is by the relative ease with which processing occurs for the perceptual characteristics of a stimulus. In doing this, we argue in accord with a functionalist account of the distinction between experiences of familiarity and conscious recollection, namely that any memorial information is likely to be interpreted as conscious remembering when it permits determination of the prior source of an encounter (Bodner & Lindsay, 2003). Finally, after reviewing the relevant research, we discuss the implications our proposed account holds for dual-process theories of recognition memory.

Many of the experiments that we will review have investigated subjective experiences in recognition memory using Tulving's (1985) remember-know paradigm. In this paradigm, participants are given a recognition memory test and asked to indicate with a yes or a no response whether stimuli were recently studied. For any stimulus given a yes response, participants are asked to introspect on the nature of their conscious experience and to report whether they 'remember' or merely 'know' that the item was previously studied. Participants are asked to give a remember response if they believe that they can consciously recollect details of the item's prior occurrence and to give a know response if they fail to recollect such details but are nonetheless confident that it was previously studied. While many dissociations between remembering and knowing have been reported (e.g., Gardiner, 1988; Gardiner et al., 2006; Rajaram, 1993, 1996), use of the remember-know paradigm as a tool for measuring the contribution of different processes to recognition memory has been extremely controversial, mainly because some have argued that remembering and knowing merely reflect different levels of confidence (e.g., Donaldson, 1996). Furthermore, even though memory for the source of an event is often thought to require recollection, recent findings suggest that familiarity can also support above chance accuracy on a source discrimination task (Diana et al., 2008). Indeed, several findings have demonstrated above-chance source discrimination for both remember responses and know responses in the remember-know paradigm (Conway & Dewhurst, 1995; Perfect et al., 1996; Wais et al., 2008). If source discrimination can be accomplished through a variety of processes, then perhaps conscious experiences of recollection do not reflect the contribution of a single process, such as episodic retrieval. An additional challenge to the claim that different subjective experiences reflect different underlying processes is that both experiences seem to be based, in part, on the outcome of heuristic decision strategies. If remembering and knowing are influenced by similar heuristic strategies and similar 'familiarity-based' processes, then this seriously weakens the claim that these subjective reports reflect independent memory processes. In the next section, we review reports of heuristics that have been found to reliably increase subjective claims of recollection on a recognition memory test.

Evidence of heuristics that lead to claims of remembering

Several heuristics have been shown to increase subjective states of recollection. One theme that emerges in these studies is that recollection responses are elevated when a stimulus is identified from an impoverished or incomplete cue. For instance, on a recognition task, people may rely on a perceptual identification heuristic, in which positive recognition responses in general and reports of recollection in particular are more likely to be assigned to briefly presented test stimuli that can be correctly identified compared to briefly presented stimuli that cannot be correctly identified (Higham & Vokey, 2004).

A similar phenomenon demonstrates that a feeling of discovery can sometimes be misinterpreted as conscious remembering (Dougal & Schooler, 2007). When participants are asked to complete an anagram before making a recognition decision, successful completion of the anagram leads to increased positive recognition responses in general and increased reports of remembering in the remember-know paradigm. Presumably, these heuristics are useful for making yes/no recognition decisions because stimulus identification and word fragment completion are generally superior for stimuli that were seen before compared to stimuli that are novel (e.g., Tulving & Schacter, 1990). Therefore, it makes sense that easy stimulus identification would be used as a cue for recognition. However, what is unclear is why perceptual identification and a feeling of discovery trigger conscious experiences of remembering rather than vague feelings of familiarity. Indeed, other studies similar to those described above have found that stimulus generation only influences claims of knowing in the remember-know paradigm. For example, Lindsay and Kelley (1996) gave participants a cued recall test with word fragments as retrieval cues. Half the fragments were created by deleting only a single letter (easy fragments) and half were created by deleting two letters (difficult fragments). Unbeknownst to the participants, some of the fragments could only be completed with a non-studied word. Relative to difficult fragments, easy fragments received more know responses but an equivalent number of remember responses presumably because participants interpreted easy stimulus generation as a cue for familiarity. This was true regardless of whether fragments could be completed with a previously studied word.

Conscious experiences of remembering, whether real or illusory, might be triggered when easy stimulus identification/generation is perceived to be relatively spontaneous and when the generated material is idiosyncratic rather than meaningfully related to the cues provided for retrieval. In the study by Lindsay and Kelley (1996), stimulus generation may have appeared somewhat less impressive than that which occurred in the studies by Higham and Vokey (2004) and Dougal and Schooler (2007) because most of the stimulus was present prior to making a retrieval attempt (i.e., even in the case of difficult fragments, which only had two letters deleted). Furthermore, participants merely generated the core word in response to its fragment, but as the authors note ‘illusions of recollection would require manipulations that cause context- and source-relevant thoughts and images to come readily to mind as well as the core item itself' (Lindsay & Kelley, 1996, p. 208). This latter possibility is consistent with Whittlesea's (2003) framework of *Selective Construction and Preservation of Experience* (SCAPE), which regards the subjective experience of remembering as based, in part, on an inference concerning subtle changes in ongoing stimulus processing, rather than a retrieval process per se. SCAPE contends that the primary function of the memory system is to reconstruct prior interactions with stimuli, and that subjective feelings of familiarity and recollection are based on inferences concerning the quality of these reconstructions. Thus, the SCAPE framework advocates a functionalist account of the distinction between remembering and knowing (Bodner & Lindsay, 2003; Bodner & Richardson-Champion, 2007; Gruppuso et al., 1997) and contends that subjective experiences of remembering can be supported by any process or processes that permit determination of the unique and distinctive spatio-temporal context of a memory.

A study by Leboe and Whittlesea (2002) highlights the importance of stimulus generation and the reconstruction of distinctive encoding contexts to subjective interpretations of memory. Participants studied three different types of word pairs: semantic associates (e.g., LION–TIGER), unrelated items (e.g., SUMMER–TABLE), and words paired with a non-distinctive cue (e.g., XXXX–HOUSE). Thus, while each item in the first two types of pairs was presented in a unique context, each word in the third type of pair was presented in a context that was shared by many other items in the study phase. At test, participants were re-presented with the second item of the study pairs (e.g., TIGER) and were asked to recall the item with which it was previously paired. Participants were also asked to rate the extent to which they felt that their act of generating a response constituted a genuine instance of remembering. The results showed that participants were most likely to provide a related word, somewhat less likely to provide a row of XXXX's, and least likely to provide an unrelated word. However recall accuracy was highest for unrelated words, somewhat less accurate for a row of XXXX's, and least accurate for related words. Most interesting though is the comparison of these data with rated confidence in genuine remembering or recollection. Participants were most confident that they were actually recollecting a prior event when they reported an unrelated word (a pattern that held even when the word that was reported was incorrect), somewhat less confident when they reported a related word, and least confident when they reported a row of XXXX's. Leboe and Whittlesea argued that these findings suggest an account of recollection that is grounded in construction and inference. Unrelated information rarely spontaneously comes to mind when attempting to remember, which is why this was the least frequent type of response. However, when an unrelated word *is* spontaneously generated, people are unlikely to see any reason for why this might have occurred, other than because the two words may have been paired together during study. Thus, when unrelated items were spontaneously generated, they were accompanied by feelings of high confident remembering. On the other hand, many words in the study phase were accompanied by a row of XXXX's, making this a relatively frequent response on the recall test. Presumably participants understood that this was a non-distinctive context and that generation of a row of XXXX's might have been due to guessing. As such, these reports were not as likely to be interpreted as genuine remembering even when they were correct.

The findings reviewed above suggest that surprisingly easy generation of information that is related to a cuing stimulus only by virtue of a shared episodic encoding history is an important contributing factor to experiences of conscious recollection. We will next turn to research that has examined how enhanced *processing* speed of an intact stimulus affects recognition decisions generally and subjective experiences of recollection and familiarity specifically.

Processing fluency and claims of remembering

One of the most extensively studied heuristics in memory is the fluency heuristic, so it makes sense to consider whether there is any evidence that this might also factor into conscious experiences of remembering. As mentioned above, fluency refers to the ease with which stimulus processing takes place, and it is now well established that people are more likely to claim to recognize stimuli that are perceived as being easy to process compared to stimuli that are perceived as being less easy to process (Jacoby & Dallas, 1981).

Fluency is widely regarded as a useful cue for inferring that one has had prior experience with a particular stimulus, but it has not been thought to be useful in determining the specific context in which that prior experience occurred. Thus, dual-process theories of recognition typically regard fluency as a contributing factor to feelings of familiarity but not conscious recollection (Yonelinas, 2002). This line of thinking comes, in part, from studies that have used the remember-know paradigm in conjunction with artificial fluency manipulations. Test items that are made more fluent with a brief presentation of a matching prime (e.g., Jacoby & Whitehouse, 1989) tend to receive more know responses but equivalent remember responses relative to less fluent items, suggesting that participants experience enhanced fluency as familiarity rather than conscious recollection (Kinoshita, 1997; Rajaram, 1993). The idea that fluency only contributes to feelings of familiarity has also been advanced by the source monitoring framework, which assigns a very minimal role to fluency in memory tasks that are comprised entirely of previously studied items and therefore require participants to remember the prior context of an event (Johnson et al., 1993). However, the contribution of fluency to source discrimination is not zero. Indeed, fluency must play some role in determining the context of a particular memory because all recognition memory tests are, to some extent, source-monitoring tasks. Participants are not asked merely to report whether they have ever seen a particular word before, but rather whether they think they may have seen it before in a particular context, namely during the study phase of the experimental session.

Even this concession may not give fluency enough credit though. Indeed, the potential role that fluency might play in determining the source of a particular memory is demonstrated in a set of experiments by Kelley, Jacoby, and Hollingshead (1989), which showed that people are especially likely to regard perceptually identified items on a visual test as having been studied earlier on a visual list rather than an auditory list. The authors concluded that perceptually identified words are perceived as being more familiar than unidentified words and that this familiarity is taken as evidence that the stimulus was presented in the same modality during the test phase as during the study phase. These results suggest that fluency might actually be quite useful in determining the context of a particular memory, at least under certain circumstances. Furthermore, if determination of the spatiotemporal context of a memory determines whether one experiences a subjective feeling of conscious recollection as opposed to vague familiarity (see Bodner & Lindsay, 2003; Gruppuso et al., 1997), then this could mean that fluency is sometimes a contributing factor to conscious experiences of remembering and not always associated with vague undifferentiated feelings of familiarity.

The notion that fluency plays a minimal role in determining the spatiotemporal context of a memory is also contrary to a set of findings showing that the use of the fluency heuristic is source specific. For instance, although robust illusions of recognition memory are found for perceptually fluent test items when the study and test list are both presented visually, this does not occur when the study phase is presented aurally (e.g., Miller et al., 2008; Westerman et al., 2002), or when both phases are visual, but the stimuli themselves are in a different perceptual form at study and test (e.g., pictures at study and words at test; Westerman et al., 2003). It seems that people are reluctant to use fluency as a cue for recognition when there is a change in sensory modality between study and test, presumably because enhanced perceptual fluency on a visual recognition test is unlikely to have resulted from prior study if the study list was presented aurally. Furthermore, Westerman and colleagues (Westerman et al., 2002, 2003) reported similar results in experiments that used 'counterfeit' study lists instead of actual study lists. Participants who were under the impression that they had viewed a visual subliminal study list (when, in actuality, they only viewed static) were more likely to interpret perceptual fluency as a cue for prior occurrence on a visual test than participants who were under the impression that they had heard an auditory subliminal study list. Interestingly, a recent study with amnesic participants showed that, compared to healthy adults, patients with amnesia used the fluency heuristic, but did not demonstrate source specific modulation of the fluency heuristic (Willems et al., submitted). This was particularly true for amnesic participants who scored low on a metamemorial belief scale that was designed to measure their knowledge of how previous exposure could impact later processing, thereby supporting the idea that metamemorial factors modulate use of the fluency heuristic. For healthy participants, the strategic use of the fluency heuristic in the manner described above suggests the operation of metacognitive processes that monitor the conditions that are likely to give rise to fluent processing of stimuli. When a potential source for enhanced fluency has been identified by this metacognitive monitor, whether it is prior study in general, prior study in a particular context, or some other source entirely, attribution to that source occurs, resulting in a sense of undifferentiated familiarity, familiarity and confidence that the stimulus was studied in one particular context, or a feeling that it was not studied at all.

Additional support for the notion that fluency might contribute to conscious recollection comes from a study by Whittlesea (2002; Experiment 6), which investigated subjective experiences associated with manipulations of *conceptual fluency* (i.e., the relative ease with which the meaning of a word can be accessed). Participants in this experiment studied a list of single words and then received a recognition test. Half of the items on the test were preceded by a sentence stem that was highly constrained and therefore predictive of the final word (e.g., The stormy seas tossed the…BOAT). The rest of the test items were preceded by sentence stems that were less highly constrained and therefore non-predictive of the final word (e.g., She saved her money and bought a…BOAT). Once participants read the sentence stem they were required to press a key to reveal the final test word and then to report whether that word was *actually remembered*, *familiar*, or *new*. Whittlesea found that predictive sentence stems increased claims of remembering the final word regardless of whether it was actually studied, and that they had no effect on claims of familiarity (Experiment 6A). These findings support the notion that fluency may be a contributing factor to conscious experiences of recollection, but they do not speak to the issue of whether perceptual fluency also contributes to these subjective experiences. This distinction is relevant because findings from our own lab suggest that conceptual fluency and perceptual fluency differ in important ways (Miller et al., 2008).

We recently investigated whether perceptual fluency contributes to claims of conscious recollection when participants report their subjective experiences using an independent ratings method developed by Higham and Vokey (2004). Unlike the standard remember-know paradigm, which requires participants to choose between two mutually exclusive alternatives (e.g., remember versus know), Higham and Vokey's independent ratings method allows participants to rate each test item on a four-point scale in terms of both recollection and familiarity (where 1 = *definitely not recollected* and 4 = *definitely yes recollected* on the recollection scale, and 1 = *definitely not familiar* and 4 = *definitely yes familiar* on the familiarity scale). Participants were given a standard recognition memory test that incorporated the Jacoby and Whitehouse (1989) priming procedure to increase the perceptual fluency of some of the test items. Kurilla and Westerman (2008) found that the fluency manipulation led to higher average ratings for both feelings of familiarity and feelings of recollection. This conflicts with the earlier reports we described above (e.g., Kinoshita, 1997; Rajaram, 1993), which showed that matching primes exclusively elevate subjective feelings of familiarity in the remember-know paradigm. The reason for this conflicting set of results may be that the independent rating method permits reports of conscious recollection that differ in terms of confidence while the standard remember-know paradigm does not. Consistent with this possibility is the fact that, in our Experiment 2, fluency primarily affected lower confidence 3 responses but not higher confidence 4 responses on the recollection rating scale.

Finding that artificial fluency manipulations increase claims of remembering provides indirect evidence in favour of the idea that the conscious experience of recollection may partly be the result of the strategic use of fluency, which appears to be governed by metacognitive assessments of the conditions that are most and least likely to lead to fluent stimulus processing. Arguably, if fluency is attributed to the prior circumstances that are most likely to have given rise to enhanced fluency (e.g., prior visual study rather than prior auditory study; Kelley et al., 1989; Westerman et al., 2002), then this could provide one with the means to pinpoint in space and in time the origins of a particular memory, which, according to a functionalist account of the distinction between remembering and knowing, could be sufficient to trigger a subjective feeling of recollection (Gruppuso et al., 1997).

The relationship between feelings of recollection and feelings of familiarity

The findings we have reviewed so far suggest that reports of conscious remembering and reports of vague familiarity are not all that different in that each can be triggered by the same type of information depending on experimental conditions. This suggests that the same event that is confidently recollected in one test context may be perceived as being only vaguely familiar in a different test context. This has been confirmed in experiments demonstrating that reports of conscious remembering can be influenced by expectations concerning to-be-tested material. In a study conducted by Bodner and Lindsay (2003), participants studied two separate lists of words under different encoding conditions. One group studied one list in a medium level-of-processing (LOP) and the other in a shallow

LOP (i.e., the medium-with-shallow group), and a second group studied one list in a medium LOP and the other in a deep LOP (i.e., the medium-with-deep group). During the following recognition memory test, participants from both groups were presented with a list of words comprised of items from their respective study lists mixed with novel lures and were asked to provide judgements of remembering and knowing. Comparisons across groups revealed that medium LOP items received more remember responses when they were tested in the context of weak targets (i.e., the medium-with-shallow group) than when they were tested in the context of strong targets (i.e., the medium-with-deep group). This difference in remembering could not be due to differences in encoding because claims of remembering were equivalent when the two groups received a test list comprised entirely of medium LOP items mixed with novel lures (Bodner & Lindsay, 2003, Experiment 2). Therefore, it seems that the surrounding test context is an important factor in determining how participants define 'remembering;' the same recognized event can be experienced as either vaguely familiar or consciously recollected depending on whether other targets on the test are relatively strong or weak. In fact, the same pattern of results can be found when one merely expects a subsequent recognition test to be comprised of strong items (McCabe & Balota, 2007). A study by Rotello, Macmillan, Reeder, and Wong (2005) found that remember responses are also influenced by the type of instructions that participants receive. When participants were required to provide a rationale for remembering, the number of remember responses they assigned to targets and lures was reduced relative to a condition in which no such rationale was required.

These findings demonstrate that reports of conscious remembering are influenced by decision strategies and expectations about the relative strength of targets. Furthermore, these findings suggest that a report of conscious remembering is triggered neither by a single process responsible for retrieving episodic detail nor by the retrieval of information beyond some rigid threshold.

Implications for dual-process theories of recognition memory

If processing fluency, stimulus identification, and stimulus generation reliably contribute to conscious experiences of recollection, then this means that feelings of recollection are not based on a source of information that is entirely separate and independent from that which supports vague feelings of familiarity. This falls perfectly in line with a unidimensional signal detection account of the distinction between remembering and knowing (Donaldson, 1996; Wixted & Stretch, 2004). A unidimensional SDT account of recognition memory contends that old and new items are normally distributed along a common continuum of evidence, such as memory strength, and that recognition decisions are made by setting a response criterion somewhere along that continuum. Items that are located above the criterion are called 'old' and items that are located below the criterion are called 'new.' From a unidimensional SDT perspective, remembering and knowing merely reflect two decision criteria along the memory strength continuum, with the remember criterion located to the right of the know criterion. However, while it is common to assume that this necessarily means that remembering is simply based on more of the same 'stuff' as knowing, it is quite possible and indeed highly probable that the memory strength decision variable is multidimensional in nature, reflecting a complex combination of multiple sources of information (Wixted, 2007). As such, different points along the memory strength continuum are likely to reflect different combinations of differentially weighted sources of evidence.

Therefore, we do not dispute the general dual-process idea that recognition decisions are supported by two processes. Rather, we agree with Wixted's (2007) proposal that recognition decisions are probably based on a combination of familiarity and recollection because neither is a perfect indicator of prior study. As such, we do not regard subjective experiences of familiarity and recollection as entirely separate. Instead, we think that they may depend on many of the same component processes and sources of information, such as repetition priming and processing fluency. Therefore, we regard subjective experiences of both familiarity and recollection as completely constructed and as by-products of complex and strategic inferences concerning subtle changes in ongoing stimulus processing and perception. From this perspective, a complete understanding of recognition memory can only come from a complete understanding of the functionally heterogenous component processes that are combined to form the decision axis of a unidimensional SDT account. Research into the way in which this information gets combined will no doubt be incredibly useful.

Conclusion

We have reviewed several findings that suggest that a number of heuristic decision strategies and unconscious sources of information may factor into conscious feelings of recollection, and that, as such, these subjective experiences may not be based on information that is entirely separate and independent from that which supports feelings of familiarity, as commonly assumed by some dual-process theories of recognition. One critical factor that we believe may encourage the experience of conscious and confident recollection is the extent to which heuristics are used strategically because doing so restricts their use to circumstances where they are most appropriate and consequently may provide one with a means for assigning spatiotemporal context to a particular memory. According to a functionalist account of the distinction between remembering and knowing, relative success in coming up with spatiotemporal context is the primary factor determining whether a memory is experienced as vaguely familiar or confidently and clearly remembered (Bodner & Lindsay, 2003; Bodner & Richardson-Champion, 2007). In the case of the fluency heuristic, we reviewed work from our own lab suggesting that perceptual fluency might satisfy this requirement and therefore be sufficient to trigger conscious experiences of recollection in some cases. This proposal is consistent with existing theories of memory, in particular Whittlesea's (2003) SCAPE theory, which views subjective experiences of vague familiarity and conscious recollection as by-products of complex metacognitive inferences regarding subtle changes in ongoing stimulus processing and perception. We feel the collection of findings reviewed here is important because it highlights something that, despite having long been acknowledged in the wider field of memory research, has been greatly overlooked by current dual-process theories of recognition, namely that subjective experiences of recollection are as much a product of inference and active construction as subjective experiences of familiarity.

8
Metacognitive Processes before and during Retrieval

Giuliana Mazzoni and Maciej Hanczakowski

Introduction

> The idea that people have expectations, developed in previous experiences, or on the fly during the current event, which they use to evaluate the goodness and significance of their current processing, adds a degree of freedom in the evolution of subjective reactions that is absent in other accounts (of retrieval; Whittlesea, 2004, p.15).

Whittlesea's statement suggests that retrieval can depend on expectations, and that a degree of freedom can be added to the retrieval processes. This additional freedom provides room for voluntary pre-retrieval strategies used when responding to an environmental cue. These pre-retrieval strategies are the focus of this contribution.

Memory is crucial in negotiating life and maintaining one's identity. Being able to remember the past, both factual and personal, helps individuals in most everyday endeavours, and provides a sense of stability and continuity essential to one's survival. Retrieval, the key process in remembering, has been mostly conceived as a mandatory process automatically activated by a cue or a question. In this chapter we discuss some evidence that favours a different conception of retrieval, one that sees retrieval as a resource-demanding gated process that is activated only when specific criteria are met. It is not our intent to claim that retrieval is never automatic. Rather, what is stressed in this chapter is that retrieval can *also* be initiated non-automatically, and can be subjected to voluntary control. Our claim is that metacognitive judgements help establish when these requirements, necessary to initiate retrieval, are met. In our view, metacognitive control over retrieval does not only occur at the output level, for example, deciding whether to volunteer or withdraw a memory candidate (see e.g., Koriat & Goldsmith, 1996), it occurs also before retrieval, in deciding whether to start the retrieval process or not.

To this aim we make ours and combine three ideas already present in the literature that help understand metacognitive control processes exerted over retrieval. The first is the distinction initially proposed by Richardson-Klavehn, Gardiner, and Java (1996) between a voluntary type of retrieval initiated and controlled by the subject, and a more automatic involuntary retrieval, which manifests itself predominantly in indirect/implicit tests. The second is the idea of familiarity as a gate element to initiate more effortful retrieval. This idea was first introduced by Atkinson and Juola (1974), according to whom familiarity precedes recollection, which starts only if evidence from familiarity is equivocal. Familiarity as a criterion to trigger search permission for retrieval has been studied recently by Malmberg (2008). We then add some recent data on item plausibility (Mazzoni, 2007) that confirm that voluntary retrieval occurs, and is controlled by pre-retrieval evaluations about the nature of the information to be retrieved.

These ideas are discussed in the context of Bruce Whittlesea's approach which posits that subjective judgements (evaluation) are separate from retrieval (production) and can influence how retrieval proceeds (Whittlesea, 1997; Whittlesea & Leboe, 2000). We propose that the memory framework proposed by Whittlesea is a good way to look at control mechanisms in memory retrieval due to its inherent selectivity. As described in Whittlesea's initial quote, his approach reveals how memory processes are controlled by expectations and attributions. As such, it is the appropriate tool for describing how individuals exert control over their own cognitive processes.

Voluntary and involuntary retrieval

Retrieval is the process by which information is accessed in memory and subsequently output. In recall, retrieval represents the essential process, whereby the content of a memory, previously encoded, is activated and accessed, and eventually output. Retrieval is usually conceived as the result of the interaction between a cue and the content of a memory. For example, according to Moscovitch (1992, 1994) the core process of retrieval is based on an automatic *ecphoric system* in which a specific cue interacts with information stored in memory.

However, ecphoric processes are not the only processes involved in retrieval. Moscovitch (1992, 1994) postulates an additional *strategic system*, which is responsible for effortful and goal-directed retrieval. These can include pre-ecphoric operations of choosing and elaborating retrieval cues that would constrain the set of information accessed in memory, and post-ecphoric operations, such as monitoring, that would ensure that retrieved information fulfils the goals of retrieval. In this chapter we want to focus on a specific pre-ecphoric control operation, namely a gating process, that we think is responsible for the way in which retrieval proceeds. However, we will treat automatic and effortful components of retrieval as two distinct retrieval processes. So we will differentiate between involuntary retrieval based solely on ecphoric process and voluntary retrieval based on ecphoric process together with control operations exerted over this process.

The distinction between direct and indirect tests of memory provides the theoretical background for distinguishing between voluntary and involuntary retrieval processes. Direct tests of memory require participants to think back to the study session and remember the material presented. In indirect memory tests, no reference is made to the study phase at the moment of retrieval. Control processes that may be involved in memory retrieval are more likely to be employed when memory is tested with a direct test, than in indirect tests, as in this latter case participants are usually not aware that their memory is tested.

Many studies have shown that performance in those two kinds of tests can be dissociated (see Richardson-Klavehn & Bjork, 1988 for a review). Some have postulated that retrieval processes engaged in those tasks tap different memory storages (Schacter & Tulving, 1994), or different memory representations (Nelson et al., 1998). In a slightly different approach, some have proposed that separate retrieval processes operate on the same memory representations, but these representations are accessed in different ways depending on the type of memory test used (Jacoby, 1998). Because indirect tests are defined as unconscious, retrieval processes involved in those tests have also been referred to as unconscious, in contrast to conscious retrieval involved in direct tests. However, Richardson-Klavehn et al. (1996) pointed out that retrieval processes are always unconscious and hence the distinction between conscious and unconscious can refer only to a state of mind reflecting the relationship between products of retrieval and the related study phase, not to the processes themselves. They proposed that rather than been described as conscious and unconscious, retrieval processes should be conceived as *voluntary* and *involuntary*, and showed that it is volition of retrieval, not consciousness about the way in which the product of retrieval refers to the study phase, that is responsible for the dissociation between direct and indirect memory tests (Richardson-Klavehn et al., 1994).

Here we adopt the distinction between involuntary and voluntary retrieval processes (Richardson-Klavehn et al., 1996), and focus our attention on control mechanisms in voluntary retrieval. We believe retrieval is a gated controlled process that can be initiated or not initiated depending on some preliminary metacognitive judgements. In the next paragraphs we will discuss familiarity and plausibility as two metacognitive criteria used in controlling voluntary retrieval, but preliminary to that we need to show that the voluntary-involuntary distinction adopted in this chapter is compatible with several memory models that apparently do not include this distinction. Several memory models distinguish between two kinds of retrieval processes (e.g., Jacoby, 1991; Nelson et al., 1998). Although they describe those processes in different ways and take different assumptions on how those processes relate to each other, their conception is still broadly consistent with the distinction between involuntary and voluntary retrieval adopted here. In the present chapter we do not want to choose one model over the other as we are less interested in the detailed description of those processes than in the basic distinction according to which one process is involuntary and cannot be controlled, whereas the other is voluntary and its application is controlled by the individual who remembers.

When two different memory processes are proposed, their reciprocal relationship should be considered. In distinguishing between involuntary and voluntary retrieval processes, Richardson-Klavehn et al. (1996) examine three forms of relationship, that is, redundancy, independency, and exclusivity. Although these had already been proposed in the literature to describe how conscious and unconscious processes might relate to each other, we agree with Richardson-Klavehn et al. (1996), who state that they should not be used when referring to conscious versus unconscious memory, but to voluntary versus involuntary retrieval. Redundancy would mean that voluntary retrieval is dependent on involuntary retrieval. Hence involuntary retrieval can occur without voluntary retrieval, while the reciprocal (voluntary retrieval without involuntary retrieval) cannot occur. Redundancy is present in the model by Moscovitch (1992, 1994) described earlier, in which the core automatic retrieval process depends on the medial temporal lobe (that we call involuntary), while the more effortful and strategic working-with-memory process depends on the prefrontal cortex (we call this voluntary).

The principle of independency implies that voluntary and involuntary retrieval run in parallel, and either of them can produce a memory. Also independency is present in several memory models, and, for example, represents a key assumption in a model developed by Jacoby (1991, 1998), in which it makes it possible to compute estimates of the relative contribution of the two retrieval processes towards the observed memory performance (see Richardson-Klavehn et al., 1996 for a discussion of this model). Similarly, Nelson et al. (1998) in their PIER2 model propose two independent retrieval processes working in parallel. In this case they operate on different memory representations. In both Jacoby's and PIER2 models, processes are described as conscious or unconscious, but if the arguments made by Richardson-Klavehn et al. (1996) are correct, which we think is the case, those processes should rather be described as voluntary and involuntary. That distinction certainly is in agreement with Jacoby's framework which refers to automatic and controlled influences on memory. In PIER2, control of what we call involuntary retrieval, operating on implicit representations, occurs only at encoding (by means of inhibition), which justifies calling this kind of retrieval involuntary and automatic. This model is then congruent with the distinction between voluntary and involuntary processes that we adopt in this chapter.

Of the three principles, only exclusivity cannot not be applied to accurately describe the relationship between voluntary and involuntary retrieval, as it would require a situation in which voluntary retrieval is initiated and involuntary retrieval is not. It would mean that control mechanisms in retrieval would at the same time be used to stop involuntary retrieval and initiate voluntary retrieval, a situation difficult to imagine. It is questionable if involuntary retrieval can be controlled at all (but see Bergström, de Fockert, and Richardson-Klavehn, 2009 for an example of control over involuntary retrieval in a think/no-think task), as it is probably impossible to stop and initiate retrieval at the same time. The exclusivity assumption, which was present in the first version of the PIER model (Nelson et al., 1992), has now been changed into an independency assumption in the new version of this model. On the whole, more than one memory model is compatible with the voluntary/ involuntary distinction in retrieval.

Familiarity and recollection as products of attributions

In recognition, the idea that there are two separate processes in memory has also long been prevalent. Two processes in recognition have been proposed by several researchers (Atkinson & Juola, 1974; Mandler, 1980), who claim that recognition involves retrieval processes which provide information about the contents of the memory trace (e.g. Mandler, 1980; Reder et al., 2000; see Yonelinas, 2002, for a review). Both familiarity and recollection processes are involved when trying to recognize an item. Most researchers treat these as separate retrieval processes, which access different kinds of information stored in memory (Jacoby, 1991; Yonelinas, 1994). Familiarity has been described as a fast and automatic process that sometimes leads to false alarms in recognition tests. The automatic nature of familiarity is evident in several models of recognition. In Mandler's model (1980, see also Mandler, 2008), for example, familiarity is a function of integration of a memory trace to which a probe is matched. In a model developed by Jacoby (e.g., Jacoby & Dallas, 1981) familiarity is a function of the fluency with which the memory probe is processed. Recollection has been described as slow, effortful and controlled process that can help to avoid commission errors (e.g., Jacoby, 1991; Rotello & Heit, 2000).

Others have defined familiarity and recollection also as subjective states, as we do in the present chapter (for various understandings of those concepts see Yonelinas, 2002). We reserve the names of familiarity and recollection to results of attributional process operating on products of automatic and controlled retrieval (that we refer to as involuntary or voluntary retrieval after Richardson-Klavehn et al., 1996). We believe that the two distinctions (voluntary-involuntary and recollection-familiarity) are to a certain degree correlated but they do *not* overlap. Indeed, products of automatic, involuntary retrieval can more often give rise to a feeling of familiarity, and products of controlled, voluntary retrieval can more often result in a sense of recollection (Gruppuso et al., 1997).

The role of subjective judgements in determining the sense of familiarity should not be overlooked. For example, according to the framework developed by Whittlesea (Whittlesea & Williams, 1998), familiarity results from subjective evaluations of the fluency experienced in processing a retrieval cue. Fluency is evaluated taking into account the specific context in which it occurs and the expectations connected to it. These evaluations determine the level of familiarity connected with a certain cue. As Whittlesea claims,

> In contrast, if you instead meet that clerk on a bus, dressed in civilian clothes, you may experience a powerful feeling of familiarity. One's general expectation about a bus is that all passengers should be strangers and hence not fluently processed. The prior experience of the clerk's face may facilitate integration of his or her features into a unitary percept (although probably to a lesser degree than in the store, with support from a uniform and surrounding context). If it does so, the fluency of processing that face will exceed the norm for the bus context. This discrepancy attracts attention to that face and requires explanation. It could be attributed to exceptional quality of the clerk's features and experienced as a feeling of handsomeness. Alternatively, it could be attributed to prior experience. In that case, one will experience a feeling of familiarity. (Stimulated by that feeling, one might go on to regenerate details of the previous encounter, such as store and uniform, resulting in actual recall or fail to regenerate such detail, resulting in a continued state of perplexing familiarity.) That is, the fluency of processing is evaluated relative to expectations aroused by the context (Whittlesea & Williams, 2001, p.12).

Therefore, according to Whittlesea, familiarity is dependent on judgements made by a subject. The same probe may elicit a feeling of familiarity, or fail to elicit it, depending on the context in which it is encountered, and the expectations held by the subject. Hence, the feeling of familiarity is the result of an evaluation that occurs well before any actual recognition takes place. This feeling of familiarity can be one of the criteria that influence control over retrieval processes, as recently indicated by Malmberg (2008), who, within the context of a global matching model of recognition, showed that familiarity modulates search permission in memory. Within the claim that memory 'control processes generally produce the input for the retrieval process, and they make use of the output from the retrieval process to govern the completion of a memory task' (Malmberg, 2008, p. 266), he showed that time to search memory increases as the familiarity of the cue increases. Therefore, he concludes that evaluating familiarity is a metacognitive judgement that determines if an item is searched in memory. Before commenting these data more in depth, however, it is necessary to explain in greater detail the idea of metacognitive judgements and metacognitive control.

Metacognitive control processes

According to the model developed by Nelson and Narens (1990), metacognition contains two key components, a set of monitoring processes and a set of control processes that act over more basic cognitive processes (Mazzoni & Nelson, 1998; Metcalfe, 1996; Nelson, 1996b). Monitoring involves assessing information about one's knowledge and performance. Control, on the other hand, involves self-regulative processes that direct and modify one's behaviour, such as processes that govern the selection of strategies used to accomplish cognitive tasks and reach cognitive goals (Cary & Reder, 2002). These two processes are central to everyday cognitive activity, from learning in an academic environment to making mental calculations. They are also interconnected with each other, so that the results of monitoring processes represent the input for control processes. The outcome of control processes is in turn the object of monitoring, making it possible to assess the extent to which the goal of the cognitive activity is achieved. Much research in metacognition has focussed on monitoring processes, attempting to understand their role and their function. Monitoring processes have been mostly conceived as conscious (however see Reder & Schunn, 1996), and in line with this conception, they have been measured as judgements of learning (JOLs) activated at the time of encoding, as feeling-of-knowing judgements of items that could not be previously retrieved, or as confidence ratings in the goodness of the reported memory output.

Metacognitive control processes have been the object of substantially less research, which has mainly focussed on strategy selection and strategy use. One example of research on metacognitive control is represented by the work on how study time is allocated during encoding, with the aim of understanding the rules governing this very basic learning strategy (Mazzoni et al., 1990). Even more rarely research in metacognition has dealt with control processes in retrieval from memory. Some substantial contributions in this area examined the interplay between monitoring outcome and control. The Koriat and Goldsmith (1996) model on decisional processes, which establish which items to volunteer and which to withhold at retrieval, is a good example of this type of work. It represents also a good example of the fact that metacognitive control processes occur at retrieval, and has highlighted that retrieval can be divided into two separate processes, but has mainly focused on control processes at work after retrieval is initiated.

Here we propose that metacognitive control processes can affect the very initiation of voluntary retrieval as well as how this kind of retrieval develops in time. In order to understand control of voluntary retrieval we need to consider that voluntary retrieval is likely to be imposed by the specific task at hand. This includes all direct tests used in laboratory paradigms that specifically require participants to retrieve information from memory. This also includes a variety of other tasks that can be completed only if participants access previous experiences.

Control exerted over voluntary retrieval in direct tests has been a focus of memory research since the work conducted by Reder in the 1980s and 1990s (see Cary & Reder, 2002, for a review). Reder (1982) asked what happens when people are requested to retrieve information from memory. Surprisingly, she found that participants do not always do what they are asked. When asked a question that requires retrieval from memory, people sometimes use other means rather than retrieval. For example, at times, when queried about the occurrence of a particular event, people first judge if this event is plausible given their knowledge, rather than searching their memory for the event. From these results Reder proposed that not only retrieval, but also plausibility reasoning is a strategy that can be used to answer a question. She also proposed that cue familiarity can be responsible for choosing an appropriate strategy for a given question (Reder, 1987).

Recently, Reder's idea of familiarity as a control mechanism for retrieval has been developed by Malmberg (2008; see next section). However it is important to notice that both Reder and Malmberg treat cue familiarity as a simple function of the number of its presentations. In this chapter we instead are in line with the implications of the framework developed by Whittlesea (1997), according to which familiarity should be understood as the result of attributional processes. We propose that familiarity is only one instance of a class of attributions that can control voluntary retrieval. In a following section we propose that plausibility of an event, which was described by Reder as a strategy of answering a question, at times should be conceived as a control mechanism over voluntary retrieval.

Metacognitive control processes before retrieval: familiarity

The idea that familiarity is a gate to retrieval has been recently explored by Malmberg (2008), who reported the results of a few studies examining how familiarity can modulate control processes that occur at the beginning of retrieval. The idea in this case is that familiarity modulates search permission in memory. Assuming, along with many memory models, that memory control processes generally produce the input for retrieval processes, in the REM model of cued recall, for example, Diller, Nobel, and Shiffrin (2001) claimed that the amount of time people are willing to search their memory is a positive function of the familiarity of the retrieval cue. The idea that willingness to search in memory for the response to a specific cue is a function of some subjective criterion is not new in metacognitive research. Several researchers in the 1980s and 1990s examined, for example, the extent to which the result of monitoring processes (e.g. feeling of knowing) are positively related to response time in recall (e.g., Nelson et al., 1984; Reder, 1987; see also Barnes et al., 1998). Feeling of knowing represents the subjective aware response to the question of whether some unrecalled item could be later recognized. Some studies have shown that this judgement is based mostly on the familiarity of the cue, and not on the familiarity of the unrecalled target (e.g., Metcalfe et al., 1993; Schwartz & Metcalfe, 1992).

The idea that feeling of knowing mediates the time spent searching in memory for a specific target item follows from the fact that a 'feeling-of-knowing' (FOK) evaluation is made whenever retrieval fails, and the time to provide a *don't know* response is a positive function of the FOK rating, with the higher the FOK rating, the longer the time spent searching memory for the target. Time is best measured in 'don't know' responses, as they are rather immune to the influence of other variables which instead might heavily affect response time when items are successfully retrieved. Nelson et al. (1984) reported a positive and significant correlation between FOK judgements and the length of search for responses to general knowledge (factual) questions, such as 'What is the capital of Australia' or 'How old is Queen Elizabeth II' or 'Who is the richest man on earth.'.

Differently from Nelson et al.'s (1984) studies, in which the mediating role of FOK judgements was assessed in semantic memory, in his more recent studies Malmberg (2008) assessed in episodic memory the extent to which memory search is a function of the familiarity of the cue. He assessed whether the familiarity of the retrieval cue influences the time spent searching in memory for a target item, which would represent an indication of familiarity controlling search permission in memory. Latencies of correct responses do not represent reliable indicators of willingness to search in memory, because in such cases search is interrupted by the retrieval of a candidate which is good enough to be reported (Gillund & Shiffrin, 1984; Koriat & Goldsmith, 1996; Nelson & Narens, 1990). Willingness to search in memory is instead measured as response time for 'don't know' responses (Glucksberg & McCloskey, 1981), which do not suffer from this limitation. If don't know responses are longer for items which are more familiar (e.g., primed items), then one can conclude that memory search is a function of familiarity.

This hypothesis was in general confirmed by Malmberg's (2008) studies, which showed that cues having a higher level of familiarity, i.e. primed cues, similar cues and cues studied for longer, produced longer search for the corresponding targets (before responding don't know) than other less familiar cues. However, and crucially, it was also found that familiarity does not always produce the effect, which seems to be obtained only when participants *believe* that familiarity represents the degree of memorability of the item. Therefore, it seems that length of memory search is not determined by familiarity per se, but by what participants believe makes items more memorable or, in other words, by what in their opinion makes search in memory worthwhile.

When the studies conducted by Malmberg (2008) are linked with our previous considerations and with the framework developed by Whittlesea, two conclusions can be drawn. First, voluntary retrieval is controlled by familiarity. Independently from whether the search is in semantic or episodic memory, cue familiarity determines if voluntary retrieval is initiated and for how long it proceeds. Second, pre-retrieval control is indeed a complex mechanism. It is not that only one simple judgement is required before voluntary retrieval starts. The pattern of metacognitive judgements that determine retrieval is complex. The first step, as Malmberg (2008) showed in one of his experiments, is a global judgement on whether familiarity of a cue is an appropriate base for determining the content of memory. When familiarity manipulation is salient and participants cannot expect any correlation between familiarity of a cue and memorability of an item, cue familiarity will be discounted as a mechanism controlling voluntary retrieval. The second step is another global judgement, which determines the criterion along the dimension of familiarity. Participants have to decide how much familiarity elicited by a cue is sufficient to justify voluntary retrieval. But it is important to notice here that familiarity does not only represent a gate to voluntary retrieval, it may also determine its temporal dynamics. So we can also imagine a function being constructed that relates the familiarity of a cue to the duration of memory search. In the third step, information retrieved involuntarily as a response to a given cue has to be attributed to a past experience giving rise to familiarity. As described by Whittlesea, experienced familiarity is in fact the result of a metacognitive judgement at the level of the individual cue. As this judgement is the result of expectations held by participants (see Whittlesea & Williams, 2001), it is possible that it is somehow related to the first global judgement about the usefulness of familiarity in a given task.

Therefore, a picture in which familiarity is a straightforward gate to voluntary retrieval is too simple. In fact there are situations in which familiarity will not be used, as described by Malmberg (2008), and there are probably also situations in which familiarity will not occur due to specific expectations held by the participants. So the question remains about what makes memory search worthwhile, if it is not just familiarity. In the next section we propose that event plausibility may be one more mechanism controlling voluntary retrieval.

Metacognitive control processes before retrieval: plausibility

In this section we claim that any characteristics that participants believe would make the presence of an item in memory more likely has the potential of determining if search is initiated at all, and will increase the length of time spent searching memory. And any characteristic that participants believe would make search in memory worthless would diminish the time spent, to the extreme situation when the probability of recalling an item is believed to be so low that search in memory is deemed worthless and is not even initiated. In this case it is clearly shown that the judgement can be conceived as a gate to retrieval, whereby retrieval is initiated or not initiated depending on the result of the judgement.

Up to now we described how a cue familiarity judgement (based on involuntary retrieval) can control retrieval. When attributional processes lead to the belief that a cue can be remembered, searching memory becomes worthwhile. However, cue familiarity represents only one specific case in which voluntary retrieval is regulated by a subjective feeling of remembering. Other attributions made on products of involuntary retrieval can also have a regulatory influence on voluntary retrieval. Here we want to present another product of this evaluative process, that is, item plausibility. We argue that when presented with a question about an event from our biography, we involuntarily retrieve information that helps us to understand that question. The retrieved information is then subjected to evaluative processing which can lead to feeling of implausibility of the queried event. Thus, the perceived plausibility of an event described in a memory probe is an important regulator of retrieval from autobiographical memory.

When we talk about item plausibility we talk about either the plausibility of a memory query or the plausibility of an item within a given context. For example, a question like 'Did you go to the beach last year?' should trigger a search in autobiographical memory. But, we claim here, the search is activated only when going to the beach is a plausible event. In the case of people who have lived their entire life in the steppes or in the mountains, the question refers to an event which for them is implausible. Therefore, it would make little sense to search memory (i.e., to initiate retrieval) for an event that certainly never happened. Events range greatly in terms of what has been called elsewhere 'personal plausibility' (Scoboria et al., 2004), which depends on the specific culture, the specific social group, and the individuals' life history. Among events, some are plausible for one group of people (e.g., those who believe in alien abduction, Newman & Baumeister, 1998) but not for another (e.g., sceptics). Therefore a question about alien abduction would trigger a retrieval attempt among believers but not among sceptics.

Plausibility can be one of the characteristics that modulate memory search. For example, previous studies have shown that preliminary plausibility judgements may precede slower memory retrieval in sentence verification tasks (Reder, 1982). The idea of plausibility judgements as a gate to retrieval is akin to the idea that time to give 'don't know' responses to memory questions depends on whether the individual believes that some information is present in memory (Gentner & Collins, 1981;

Glucksberg & McCloskey, 1981; Klin et al., 1997; Koriat & Lieblich, 1977). In these studies it was found that very fast responses are obtained when no relevant information is present in memory, whereas the provision of information slows response times, even if the information is irrelevant and uninformative. These results have been explained by postulating the presence of metacognitive processes that provide a fast preliminary evaluation of the stimulus or the content of memory. Whether further search in memory or other cognitive processes are activated depends on the output of these fast preliminary monitoring processes (also see Metcalfe, 1993).

The hypothesis about item plausibility as a gate to retrieval initiation predicts that time spent searching in memory is a function of the level of the perceived personal plausibility of an event. If an event is deemed highly implausible, then no memory search is triggered, and a very quick 'No' response is output, directly derived from the decision about plausibility. Only when the event is deemed plausible a search in memory is activated. This process is illustrated in Figure 8.1.

The left branch represents the case in which the event is deemed implausible; the right branch represents the case in which the event is considered sufficiently plausible to deserve a memory search. As the figure shows, in case of a clearly implausible event, no further processes are activated, and the response to the question 'Did the event happen to you?' should be a very quick 'No.' In case of a plausible event, the response could be of either type (Yes or No), and, more important, it should be much slower because more processes are at play. The time to say whether an event had happened should be proportional to the perceived level of plausibility, so that the greater the plausibility of an event, the greater the time to come up with a response.

This process was tested in one experiment (Mazzoni, 2007) in which participants were asked to state whether each of six events had happened to them (ratings of likelihood of occurrence). Events ranged from highly plausible (losing a toy as a kid) to highly implausible (witnessing demonic possession), and response latencies were measured, along with plausibility ratings and ratings in likelihood of occurrence. For one event (Demonic possession) participants provided a wide range in plausibility ratings (from highly plausible to highly implausible). Latencies were measured only for low likelihood of occurrence ratings, corresponding to 'No, the event did not occur' responses. In other words, latencies were measured only for events that participants believed had not happened to them, as only in this condition latencies for highly implausible events should reflect only the judgement, and no search in memory should occur. The hypothesis that 'No' response latencies when the event was deemed highly implausible are significantly faster that 'No' response latencies when the same event is considered plausible was confirmed. When the event was deemed highly implausible, the average latency for 'No' responses was 270 ms, whereas when the event was

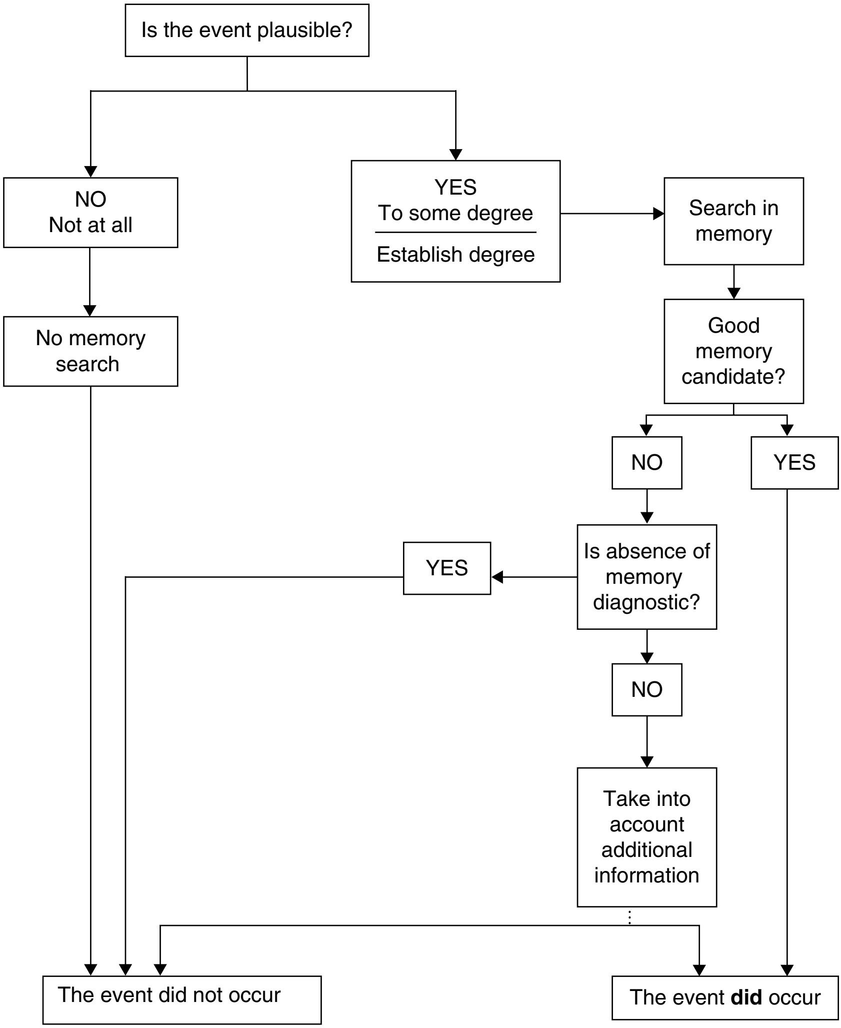

Figure 8.1 Item plausibility as a gate to retrieval initiation

deemed not implausible latencies ranged between 324 ms. and 1203 ms, with a linear increase as plausibility ratings increased. In a second study (Mazzoni, in preparation), ten new events were used, ranging again from highly plausible (having a baby tooth removed) to highly implausible (seeing a ghost). They were presented with the same procedure used in the previous study, in which participants were asked to rate the plausibility and the likelihood of occurrence for each event.

Again, it was found that latencies to rate the event as not having occurred were significantly shorter in participants who deemed the events implausible than in those who deemed the event plausible. In addition, there was a significant increase in latencies as the rated plausibility of the event increased.

These results indicate that for events that are considered implausible participants either stop their search in memory very soon, after approx 200 ms, or they do not even initiate any search in memory. In either case, these results represent a demonstration that search in memory (i.e., retrieval) is controlled by pre-retrieval judgements.

Future directions

In experimental research the focus is usually on a limited number of aspects of memory functioning. The conceptual framework developed by Whittlesea allows us to take a broader perspective and try to understand varieties of subjective feelings that are determined by ever-fluctuating expectations and attributions. Such a perspective may prove useful for understanding phenomena which not are easily subjected to the experimental method.

One phenomenon that may be an interesting case for applying this kind of approach is déjà vu. This elusive feeling has long perplexed students of memory. It has been proposed that déjà vu may be defined as a strong feeling of familiarity without recollection (for a review see Brown, 2003). In a recent study, Cleary, Ryals, and Nomi (2009) developed a laboratory procedure to elicit déjà vu in a simple recognition test and showed that reports of déjà vu are in fact related to feeling of familiarity. However, we believe that this account may not be complete. It may prove useful to consider attributions made after involuntary retrieval as judgements that can lead to a variety of feelings experienced simultaneously. For example, it is our conviction that people can experience simultaneously feelings of familiarity and implausibility, as both stem from attributional processes operating on the same information retrieved from memory. From this point of view, déjà vu could be a feeling that an environmental cue is strongly familiar but encountering it in the past is in fact highly implausible. Those opposing feelings may be responsible for the element of 'surprise' that accompanies déjà vu.

Alternatively, it should also be considered that it is the feeling of recollection, rather than familiarity, mixed with personal implausibility that could account for the phenomenon of déjà vu. In the framework developed by Whittlesea and adopted here, familiarity and recollection are attributions that may be in fact wrong – they may be evoked by cues that are actually novel. Given that false feeling of familiarity is a common experience, it may seem unlikely that it gives rise to the relatively rare experience of déjà vu. False feeling of recollection, on the other hand, is a rare experience, particularly if coupled with perceived personal implausibility of the event that evoked it. The possibility that déjà vu consists of false feeling of remembering accompanied by a feeling of implausibility of an event could represent a contribution of attributional models to our understanding of such exceptional experience as déjà vu.

In similar vein we could also consider another memory phenomenon, namely the tip-of-the-tongue (TOT) state (see Brown, 1991, for a review). Under certain conditions people experience a strong feeling that some information is stored in memory, although they are not able to retrieve it at the particular moment. We could look at this phenomenon as another example in which both familiarity and plausibility attributions play a role. This time we could imagine that TOT stems from a strong feeling of familiarity together with a positive plausibility attribution. It is interesting to notice that in this case a mixture of those two attributions serves as the control mechanism described in this chapter, as people during the TOT experience are highly motivated to retrieve information that would match a cue.

At the end it is also worth noting that whenever cognitive control mechanisms are described, a question of breakdowns in this control arises, and that some forms of psychopathology might be linked to deficient gating mechanisms that control voluntary retrieval. Although this might occur when memory cannot be not stopped or blocked (e.g., PTSD or some exceptional cases, such as the case of AJ, a woman whose uncontrollable memory forces her to relive and re-experience in great detail every single day of her life), some caution should be used in distinguishing between those cases in which post-retrieval control fails and conditions in which the problem is specific to pre-retrieval control. The question if Whittlesea's framework can be useful in linking studies on memory and clinical psychology should be addressed by future studies.

Conclusion

Here we proposed a view of retrieval as a gated, controlled, process, and stressed how voluntary retrieval can be controlled by metacognitive evaluations such as cue familiarity and event plausibility. This proposal does not have to be limited to voluntary retrieval activated by instructions in the lab, but can be extended to spontaneous voluntary retrieval in everyday life. Spontaneous voluntary retrieval occurs in indirect tests which are 'contaminated' by voluntary retrieval because of the cues used to make participants aware of the importance of the study phase of the experiment. Such spontaneous retrieval is also important in everyday life in which people voluntarily use memory not only to answer direct questions but also to guide their behaviour in all sorts of situations. Such contamination might prove useful in investigating spontaneous voluntary retrieval. Control mechanisms in spontaneous voluntary retrieval have not been subjected to systematic research yet. The reason may lie in the difficulty to analyse in the laboratory a process that is initiated spontaneously rather than triggered with experimental instructions. However, this can be accomplished by combining the application of the involuntary/voluntary distinction for indirect tests of memory (Richardson-Klavehn et al.,1996) with Whittlesea's understanding of familiarity.

Part IV

Inferential Processes and the Regulation of Accuracy

9

Accuracy Discrimination and Type-2 Signal Detection Theory: Clarifications, Extensions, and an Analysis of Bias

Philip A. Higham

Introduction

In the beginning, we spoke of NOBAL and FURIG. As Bruce's undergraduate honours student, I found these discussions fascinating and exciting. In fact, dialogue with Bruce about NOBAL and FURIG, as well as the many other stimulating discussions we had, led me to pursue a career in research. As I wrote in my honours thesis acknowledgments all those years ago, 'His enthusiasm for research has been instilled in me and it made my thesis exciting and fun instead of tiresome.'

Now that my research career is established and my vocabulary has developed, my graduate students and I speak of HENSION and STOFWUS. These discussions are also exciting, but now tinged with sadness because we realize that there is no more to come. There is no doubt that Bruce' remarkable contribution to memory research will be sorely missed.[1]

Metacognitive tasks as response accuracy discrimination

The focus of my chapter is on an important type of discrimination that provides the basis of many metacognitive tasks: people's discrimination of their own accuracy. For example, researchers might assess the extent to which retrospective confidence ratings (RCRs) about the accuracy of answers to general knowledge questions predict the answers' actual accuracy. Other metacognitive measures include judgements of learning (JOLs; e.g., Weaver & Kelemen, 2003), feeling of knowing (FOKs; Hart, 1965), and decisions about whether to report versus withhold responses (e.g., Higham, 2002; Koriat & Goldsmith, 1996). Although the tasks have clear differences, they can all be used to measure participants' ability to discriminate correct answers from incorrect ones, a type of discrimination often dubbed *resolution*.

In the signal detection theory (SDT) literature, requiring observers to discriminate the accuracy of their own (typically stimulus-based) responses is a type-2 discrimination task, first identified in the 1950s (e.g., Clarke et al., 1959; see Galvin et al., 2003, for a review). Type-2 SDT is not well known and there have been some egregious misinterpretations of it in the literature (see Galvin et al., 2003, for examples). Perhaps for these reasons, the connection is seldom made between the type-2 SDT discrimination task on the one hand and the assessment of resolution in current metacognitive research on the other. Consequently, the valuable tools of SDT are seldom applied to metacognitive data.

However, there is some evidence that this state of affairs may be changing. In our own work, we have been using type-2 SDT to estimate retrieval, resolution, and control processes in cued recall (Higham, 2002; Higham & Tam, 2005, 2006) and formula-scored testing (Higham, 2007), both of which incorporate RCRs and an explicit report option (i.e., the opportunity to pass). Additionally, we have used type-2 SDT to examine the processes involved in test revision, in particular, those processes underpinning the changing of previously selected answers on multiple-choice tests (Higham & Gerrard, 2005). We have also used the type-2 SDT model to estimate the level of bias that will maximize scores on formula-scored tests (Higham & Arnold, 2007), and more recently, to examine monitoring processes in old/new recognition for which no report option is provided (Higham et al., 2009). All of this work involves RCRs, but now type-2 SDT has been extended to cover other metacognitive judgements. For example, Masson and Rotello (2009) and Benjamin and Diaz (2008) both reanalyzed previously published JOL data using type-2 SDT.[2] These reanalyses, in particular the Receiver Operating Characteristic (ROC) curves generated from the JOL data, provided clear evidence that SDT-based measures of resolution such as d_a are preferable to the measure that is currently most prevalent in the metacognitive literature: the Goodman-Kruskal gamma correlation coefficient (*G*; Goodman & Kruskal, 1954), an ordinal index of metacognitive accuracy advocated by Nelson (1984).

Nevertheless, there is still some progress to be made before type-2 SDT becomes commonplace in metacognitive research, and one goal of my chapter is to contribute to this progression by correcting some misconceptions about type-2 SDT. In doing so, I will also extend the type-2 SDT model into new territory and compare SDT and non-SDT measures of response bias.

The remainder of my chapter is divided into seven sections. In the first section, I will describe the two current frameworks that can model the strategic regulation of accuracy using a report option: Koriat and Goldsmith' (1996) *quantity-accuracy profile* (QAP) methodology and type-2 SDT. In the second section, I will outline some misconceptions about recent applications of type-2 SDT to accuracy regulation. Partly in an attempt to clear up some of these confusions, the following four sections are dedicated to extending the type-2 SDT model to the quantity-accuracy trade-off and to discussion of response bias measurement. Finally, the chapter will wrap up with some concluding remarks.

Frameworks for the strategic regulation of accuracy

An important accuracy-discrimination task that has received a reasonable amount of attention for over a decade involves use of a report option (i.e., the decision to report versus withhold a candidate response). Koriat and Goldsmith (1996; see also Goldsmith & Koriat, 2008) introduced the first theoretical framework for modelling the strategic regulation of accuracy involving a report option. This framework uses QAP methodology and incorporates common concepts from the literature on metacognition. For example, it distinguishes between a *monitoring* mechanism that subjectively assesses the accuracy of candidate responses retrieved from memory, and a *control* mechanism that decides whether or not to volunteer the candidate response by comparing the assessed accuracy probability (P_a) against a pre-set report criterion (RC) probability (P_{rc}). The candidate is withheld unless P_a is equal to or exceeds P_{rc}, in which case it is reported. The value of P_{rc} is malleable and is influenced by situational demands and payoffs. Koriat and Goldsmith (1996) argued that a report option is used to regulate the accuracy of information that is reported and leads to the *quantity-accuracy trade-off* (see below).

The second framework is one based on type-2 signal detection theory (e.g., Higham, 2002, 2007). Because the task involving a report option is to report correct responses and withhold incorrect ones, accuracy discrimination is required, which makes the report/withhold decision type-2. Figure 9.1 shows a type-2 SDT model that might be used to generate predictions and interpret data from accuracy-regulation experiments. The model in Figure 9.1 is an equal-variance Gaussian type-2 SDT model similar to that specified in Higham's (2007) research. In the model, there is an underlying dimension of subjective confidence over which the participant's own correct and incorrect candidate responses are distributed.[3] It is assumed that resolution is above chance such that the correct candidate distribution has higher mean confidence than the incorrect candidate distribution. Indeed, the distance between the means of these two distributions, *d′*, is a good index of resolution in this model. To decide whether or not to report a candidate response, a criterion is adopted, shown as the vertical line in Figure 9.1. If a particular candidate response has confidence equal to or higher than the criterion, the candidate is reported. Otherwise, it is withheld. The criterion splits both of the two distributions, yielding four areas. The proportions of correct candidates that are reported versus withheld constitute the hit

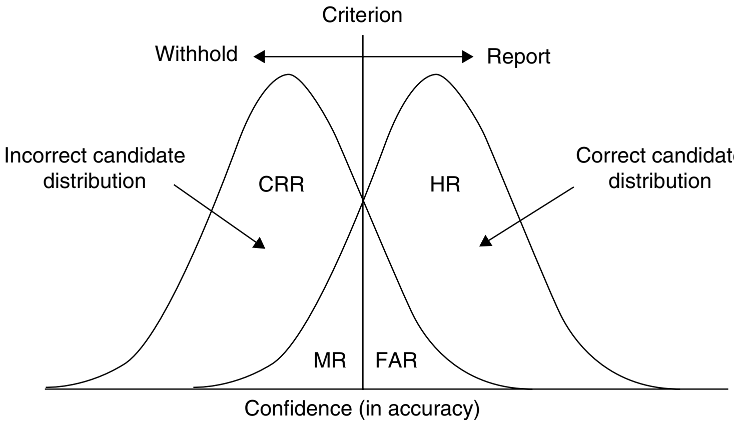

Figure 9.1 A type-2 signal detection model used by Higham (2007) to analyse performance on formula-scored tests that include a report option. Correct and incorrect candidate memory responses form equal-variance Gaussian distributions on the dimension of subjective confidence. A report criterion is adopted on the confidence dimension and if a candidate response is equal to or higher than the criterion, it is reported. Otherwise, it is withheld. HR = hit rate = proportion of correct candidates that are reported; FAR = false alarm rate = proportion of incorrect candidates that are reported; MR = miss rate = proportion of correct candidates that are withheld; CRR = correct rejection rate = proportion of incorrect candidates that are withheld.

rate (HR) and miss rate (MR), respectively. The analogous rates for incorrect candidates constitute the false alarm rate (FAR) and correct rejection rate (CRR), respectively.

Although there are obvious similarities between the two models, Koriat and Goldsmith (1996) explicitly rejected SDT as a viable model of accuracy regulation, which is at least one reason that they developed the QAP methodology instead. However, since then, Higham (2007) has pointed out that the bases of their complaints applied only to type-1, not type-2 SDT. In type-1 SDT, the observer's task is to discriminate between the presence/absence of *stimuli*, not the accuracy of *responses*. For example, in recognition memory research, the experimenter defines which stimuli on the recognition memory test are old (signal present) and which are new (signal absent). However, Koriat and Goldsmith (1996) argued that with SDT analyses of recognition data, there is no distinction between *memory retrieval* (overall memory strength) and *resolution* (the extent to which confidence discriminates between old and new items); they are operationally and conceptually the same thing, both measured with a single index of discrimination such as *d′*. Hence, they concluded that SDT was not useful for separating metacognitive and retrieval processes that are both involved in accuracy regulation.

These arguments are partially correct with respect to type-1 SDT, but are definitely incorrect with respect to type-2 SDT.[4] As long as the task involved requires discrimination of response *accuracy* and not *stimuli*, the measure of discrimination does not index retrieval but is an uncontaminated measure of resolution. Retrieval is indexed by *f*, the proportion of S (correct) trials in the experiment (more on this below). Type-2 discrimination is also theoretically independent of report bias, which in this case, reflects participants' overall willingness to report candidate responses. The SDT framework for accuracy regulation is less reliant than QAP methodology on the concepts and methods specific to the metacognitive literature (although there is nothing precluding their use). Indeed, the fact that SDT is already well developed, at least in its type-1 form, means that there is a large, true-and-tested arsenal of tools available to researchers interested in questions regarding the strategic regulation of accuracy.

As a case in point, consider Higham's (2007) use of Receiver Operating Curve (ROC) analysis to investigate the nature of the underlying evidence distributions involved in the strategic regulation of accuracy with formula-scored tests. Two curves, one corresponding to the feedback group, and another to the no-feedback group (both given low incentives) are shown in Figure 9.2. (Ignore the labels for the moment.) Both curves fall on the same ROC, suggesting comparable resolution. Also, the ROC is curvilinear and not skewed and the overall mean slope of the best-fitting line for the *z*-ROC was very close to one. Together, these results indicated that an equal-variance Gaussian evidence model was appropriate which then specified a measure of resolution (i.e., type-2 discrimination index) that would be suitable (d').[5]

Choosing an index of metacognitive accuracy based on the distributional characteristics of one's current data set is unheard of with QAP methodology. Indeed, since Nelson (1984) advocated *G*, very few alternative measures of resolution have even been considered. This failure to consider other measures is most likely due to the fact that ROC analysis, which is specific to SDT, and which is commonly used by SDT theorists to specify the evidence model and discrimination/bias measures most suitable for a given set of experimental data, is mostly absent in the current metacognitive literature.

Misconceptions about type-2 SDT

There are a number of misconceptions that have appeared in the recent literature about type-2 SDT, the measures that can be derived from it, and the meaning of those measures (e.g., see Goldsmith & Koriat, 2008). In this

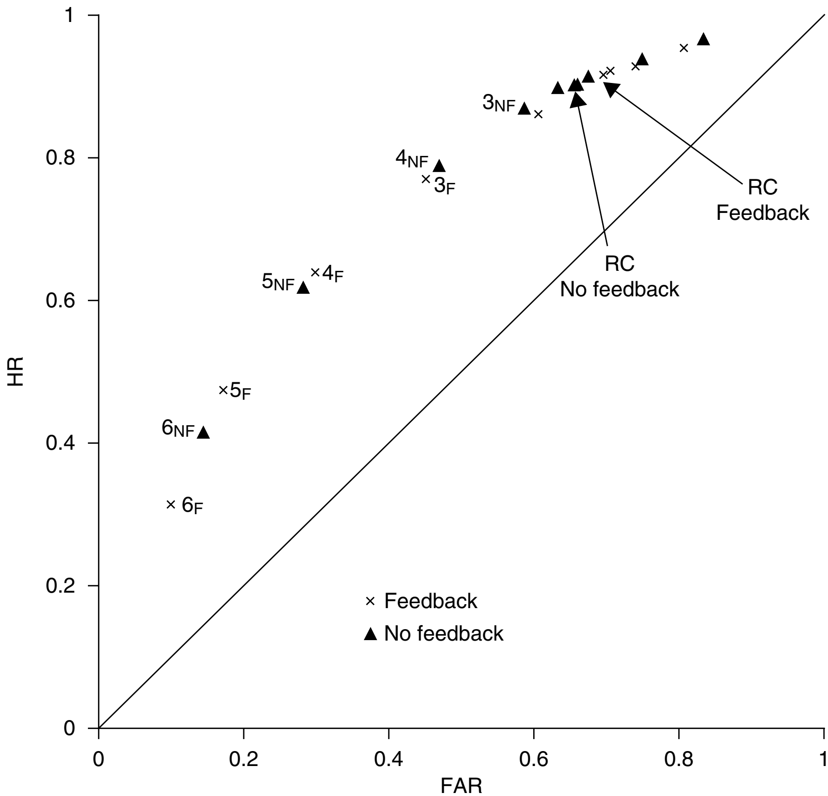

Figure 9.2 Receiver operating characteristics for the feedback- and no-feedback groups of Higham (2007, Experiment 2, low-incentives). HR = type-2 hit rate (P[high confidence|correct response]); FAR = type-2 false alarm rate (P[high confidence|incorrect response]); RC = report criterion; F = feedback; NF = no feedback. The values 3–6 correspond to specific cumulative confidence levels

section I will first describe these misunderstandings and then attempt to set the records straight.

1. Resolution is necessarily measured with respect to RC rather than confidence ratings. It is true that resolution has commonly been computed with respect to RC in published research. Indeed, I have even advocated using RC as the point at which to compute resolution and bias because of the relevance of the report/withhold decision to participants (e.g., see Higham, 2002). However, just as with type-1 SDT, confidence ratings can also be used to construct type-2 ROC curves without the need for RC (see Higham, 2007, Appendix, for how to compute type-2 ROC curves for RCR data either with or without RC). Once the ROC is computed, a variety of discrimination measures can be derived. Higham et al. (2009) calculated the corrected-HR (HR – FAR) collapsed across values of the ROC, whereas both Benjamin and Diaz (2008) and Masson and Rotello (2009) have computed d_a from JOL data. In none of these cases was an explicit report criterion even defined.

2. QAP methodology distinguishes between more performance parameters than type-2 SDT. In the recent accuracy-regulation literature in which type-2 SDT methods were used, the focus has been mostly on retrieval at forced-report, resolution (discrimination), and criterion setting (bias). On the other hand, QAP methodology distinguishes between these three parameters in addition to calibration (over/underconfidence) and control sensitivity (degree to which RC is sensitive to resolution). In a similar vein, output-bound memory accuracy (and most notably, its relationship to quantity) has not been computed in any published work on type-2 SDT.

The fact that more performance measures have been defined with QAP than with type-2 SDT is more a function of the kinds of research questions being asked than it is an indication of anything inherently limited by the type-2 SDT approach. Space constraints make it impossible to address all of these issues, so in the next section, I will focus on extending the type-2 SDT framework to model the relationship between quantity and accuracy.

3. QAP methodology measures 'true' criterion shifts (with P_{rc}) whereas SDT bias indices (e.g., B'_D) measure reporting tendencies. The latter part of this chapter is dedicated to an analysis of bias measurement from each perspective. In this analysis, assumptions are made regarding the type of criterion shift or distribution shift under consideration and the impact the shift has on various bias measures. These analyses will demonstrate that under some circumstances, P_{rc} will indicate a shift when in fact one has not occurred. The same problem is shown to be absent for type-2 SDT.

The quantity-accuracy trade-off and type-2 SDT

To understand the quantity-accuracy trade-off more fully, consider an eyewitness in the courtroom giving testimony. If the eyewitness is cautious about reporting errors, she may offer fewer answers, choosing instead to respond, 'I don't know' unless she is certain the answer is correct. This added caution would have two effects. The first is the intended effect; that is, the *output-bound accuracy* of the testimony (i.e., the quality what is actually reported) will increase. The typical cost of this accuracy increase, however, is a decrease in the *input-bound quantity* of information stored in memory that is reported. Thus, the eyewitness can be described as having the competing goals of quantity on the one hand versus accuracy on the other, which occurs because of the *quantity-accuracy trade-off* (see Goldsmith & Koriat, 2008, for a review). In general, the trade-off means that the more often the eyewitness selects to say 'I don't know,' the more accurate the selected output, but the lower the quantity.

To extend type-2 SDT to model the quantity-accuracy trade-off, first consider input-bound quantity performance under forced-report conditions. This would be indicated in Figure 9.1 by locating RC to the far left (liberal) position, such that all candidates are reported. Forced-report quantity (and accuracy) performance under these circumstances is simply the proportion of *all* items (both reported and withheld) that are correct. Higham (2007) identified this proportion as f (for forced report), which can take on any value from zero to one (0 = *all incorrect*; 1 = *all correct*).

General formulas for quantity and accuracy

If the criterion is moved to a more conservative position, some correct items will be withheld, and free-report quantity performance decreases. In particular, it decreases according to the HR, such that,

$$Q = HR \times f \quad (1)$$

where Q is quantity performance.

Only the areas above the criterion are relevant if considering output-bound accuracy performance (i.e., the proportion of reported candidates that are correct). In particular, accuracy is equal to the HR (weighted by f) divided by the proportion of all reported answers. In other words,

$$A = HR \times f / [HR \times f + FAR \times (1 - f)] \quad (2)$$

where A is output-bound accuracy.

Just as with QAP methodology, the SDT framework makes specific predictions about how resolution affects the quantity-accuracy trade-off. These relationships are shown in Figure 9.3 for chance-level resolution (i.e., HR = FAR, regardless of the criterion placement; panel A), excellent resolution (HR = 1; FAR = 0 for non-extreme criterion placement; panel B), and moderate resolution (i.e., 0 < FAR < HR < 1 for non-extreme criterion placement; panel C). Unlike Koriat and Goldsmith's (1996) QAP simulations of the quantity-accuracy trade-off, the SDT simulations involved equal-variance Gaussian distributions, meaning that the x-axes in all graphs are measured in standard-deviation units. For the models shown on the left in Figure 9.3, −4 represents maximum uncertainty ('guessing') whereas +4 represents maximum certainty ('certain correct'). For

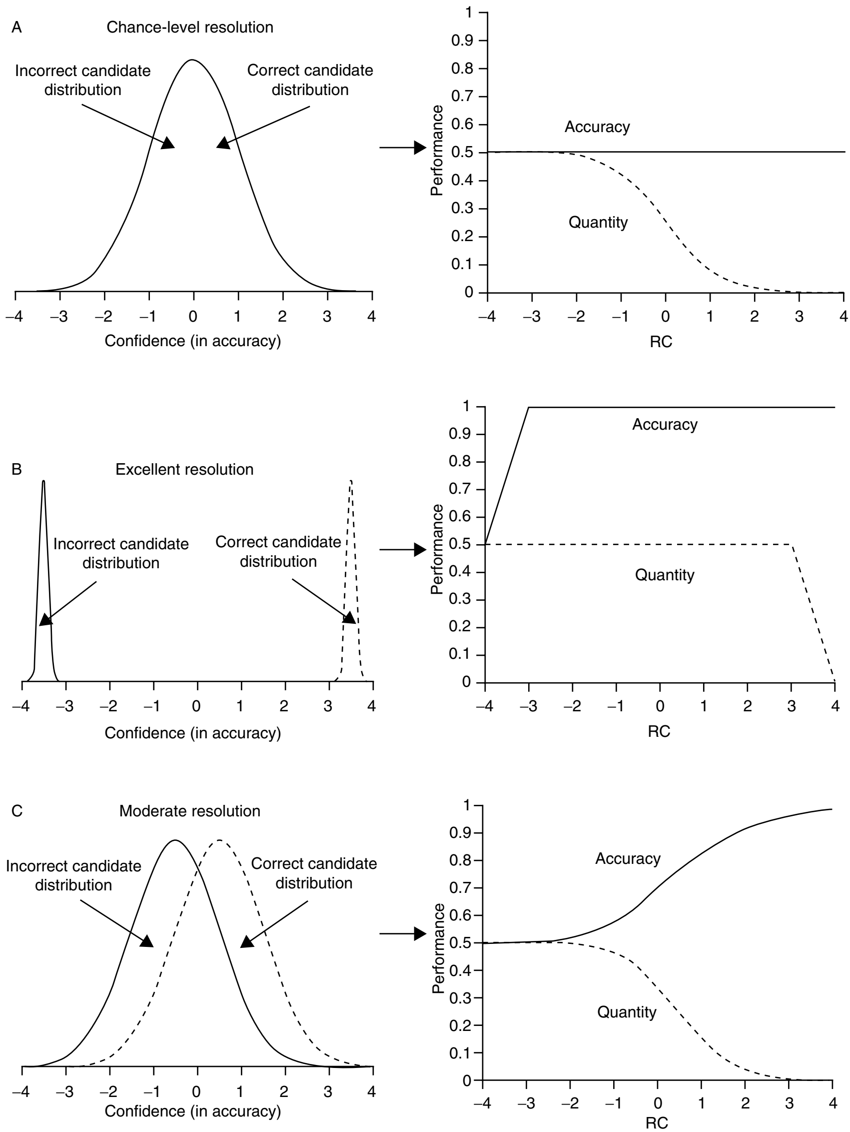

Figure 9.3 Quantity and accuracy performance based on chance-level resolution (panel A), excellent resolution (panel B), and moderate resolution (panel C). f = 0.5 in all cases. Equal-variance Gaussian distributions are assumed in all cases and the underlying dimension of confidence is measured in standard-deviation units. RC = report criterion

the graphs on the right-hand side of Figure 9.3, quantity and accuracy performance is shown for different levels of RC with f fixed at 0.50.

Forced report

If the criterion is placed at −4, then all candidates are reported and Q = A = f = 0.50. The fact that Q and A both converge on f with a maximally liberal criterion setting can be derived from Equations 1 and 2. For the case of a maximally liberal criterion placement (e.g., −4), HR = FAR = 1. Thus, Equation 1 for quantity simplifies to,

$$Q = f \quad (3)$$

and, similarly, Equation 2 for accuracy simplifies to,

$$A = f \quad (4)$$

Chance-level resolution

For chance-level resolution shown in panel A of Figure 9.3, both the correct and incorrect item distributions overlap, such that no matter where the criterion is placed, HR = FAR. Equation 1 shows that Q will decrease with more conservative criterion placements because the HR decreases. On the other hand, with chance-level resolution and HR = FAR, then both HR and FAR cancel out in Equation 2 to produce Equation 4. Consequently, no matter where the criterion is placed, A = f. This relationship is shown on the right-hand side of panel A in Figure 9.3.

Excellent resolution

Now consider the case of excellent resolution, shown in panel B of Figure 9.3. In this situation, the HR = 1 and FAR = 0 for all but the most liberal and conservative criterion placements, such that Equation 1 (for quantity) simplifies to Equation 3, and Equation 2 (for accuracy) simplifies to,

$$A = 1 \quad (5)$$

This relationship is shown in right-hand side of panel B in Figure 9.3.

Moderate resolution

Finally, for moderate resolution, shown in panel C of Figure 9.3, the correct and incorrect candidate distributions are separated by one SD (i.e., $d' = 1$). The function for quantity performance is a non-linear decrease to quantity as the criterion becomes more conservative. However, the function for accuracy rises with a slope that is between chance-level resolution and excellent resolution. It is also non-linear.

The results of these simulations are broadly consistent with those of Goldsmith and Koriat (2008; see also Koriat & Goldsmith, 1996). For example, they too found that better resolution tended to boost the free-report accuracy advantage over forced-report performance. They also found that the quantity-accuracy trade-off was nonexistent over most levels of the criterion if resolution was near perfect. However, the real value of the results of these simulations is that they demonstrate that an extension of the type-2 SDT framework to the quantity-accuracy trade-off is not impossible or even awkward, but quite straightforward. The simplicity of the model means that many other relationships involved in the strategic regulation of accuracy can be investigated as well.

Bias measurement in metacognitive research

I turn now to the issue of how best to measure bias in metacognitive research. Although measures of discrimination (resolution) have been under scrutiny lately (e.g., Benjamin & Diaz, 2008), there appears to have been very little consideration of bias *at all*.[6] This oversight is problematic because simulations have shown that response bias has an effect on G if plausible assumptions are made about the underlying evidence distributions (e.g., Masson & Rotello, 2009; Rotello et al., 2008).

I am aware of only one non-SDT bias index of metacognitive performance: P_{rc}, Koriat and Goldsmith's (1996) QAP measure of the RC in studies examining the strategic regulation of accuracy. The main difference between P_{rc} and SDT measures of bias lies in what the criterion is measured relative to. In QAP, P_{rc} is computed relative to the confidence ratings. To calculate P_{rc}, the proportion of items that conform to the decision rule is calculated for various criterion confidence levels. Conforming to the decision rule means that the confidence assigned (P_a) to withheld candidates is less than the criterion level of confidence, whereas P_a for reported candidates is greater than or equal to the criterion. The criterion level of confidence that maximizes this proportion (known as the fit rate or fit ratio) represented P_{rc} for a given participant (see Koriat & Goldsmith, 1996, for more detail).

On the other hand, SDT bias indices are measured relative to the evidence distributions, primarily being sensitive to the joint magnitude of the HR and FAR. There are many bias indices to choose from (e.g., C, ß, lnß, B'_D, and so on) each measuring different aspects of the model and each with different distributional assumptions. For example, C assumes an equal-variance Gaussian model, and represents the distance (in SD units) between the criterion and the intersection point of the evidence distributions (where C = 0). However, they all share the characteristic of indicating more liberality as the HR and/or FAR increase. Hence, SDT bias indices all differ from P_{rc} in that they are based on the *rates of high-confidence responding* (e.g., rates of equalling or exceeding a criterion level of confidence), not on the process of pinpointing the scale value corresponding to RC. This difference means that SDT bias measures are more generalizable than P_{rc}, which was designed only to apply to accuracy-regulation experiments involving an RC. Indeed, like discrimination, SDT bias indices can be computed even if there is no explicit RC and only confidence ratings are available. The levels of the confidence scale can be used to define criteria, yielding HRs and FARs from which bias can be computed.

Detection of criterion shifts

In this section, I will consider two criterion-shift scenarios. In one scenario, depicted in Figure 9.4, RC is shifted, but the criteria associated with assigning levels of confidence (10% versus 20%, and so on) are static. In other words, a greater level of subjective confidence (shown as the solid horizontal lines) is needed before a report decision is made (i.e., RC shifts), but the confidence criteria are fixed. P_{rc} will detect the shift in this case because it is measured against a fixed confidence scale. As shown in Figure 9.4, P_{rc} increases from 50% to 60%.

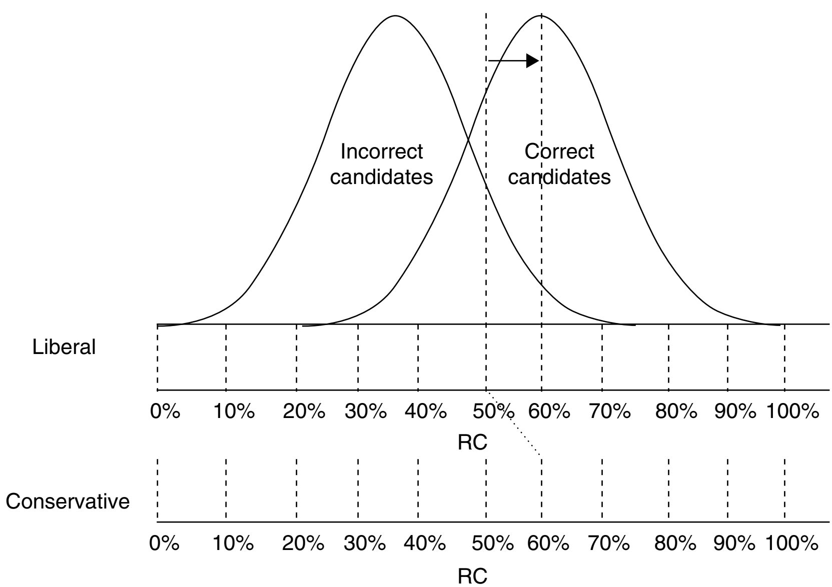

Figure 9.4 SDT depiction of a criterion shift whereby the report criterion (RC) becomes more conservative, but the confidence criteria are static. A criterion shift of this sort will be accurately detected by both P_{rc} and by SDT indices measured with respect to RC

Importantly, SDT measures of bias will also detect this criterion shift as long as the HRs and FARs are measured with respect to RC.[7] Note that both the HR and FAR for RC are larger in the liberal case compared to the conservative case. If, instead, they are measured with respect to any of the confidence criteria (e.g., 70%), then no shift will be detected because the confidence criteria are static. Indeed, this is an accurate reflection of what is occurring: the amount of subjective confidence needed to assign, say, 70% to a given response has not changed.

Now consider the scenario depicted in Figure 9.5. In this case, the confidence criteria have shifted, but RC has not. In other words, more subjective confidence is needed to assign a given level of confidence (e.g., 70%) to a candidate, but the evidence needed for a report decision hasn't changed. Because P_{rc} is measured relative to the confidence ratings, as the confidence criteria shift up the scale, RC becomes more liberal with respect to them. Thus, in this example, the conservative shift of the confidence criteria without a corresponding shift in RC will cause P_{rc} to decrease from 50% to 40%.

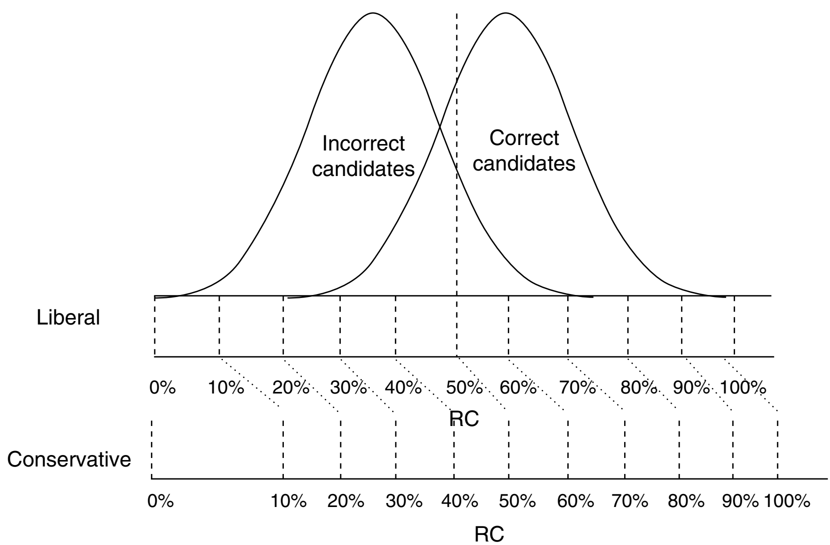

Figure 9.5 SDT depiction of a criterion shift whereby the report criterion (RC) remains static, but the confidence criteria become more conservative. Under these circumstances, P_{rc} will incorrectly indicate more liberality in the conservative condition (40%) than in the liberal condition (50%). Conversely, SDT indices will accurately indicate no change of RC, but a conservative shift of the confidence criteria

Figure 9.5 depicts a particularly dangerous scenario for the QAP researcher because P_{rc} will suggest a shift of RC towards liberality, whereas in fact RC has not moved. On the other hand, SDT indices, measured with respect to the evidence distributions and not the confidence ratings, will not be adversely affected by the shift in the confidence criteria. Because the HR and FAR are the same at RC, SDT bias measures will not differ between the liberal and conservative cases. Measured with respect to the confidence criteria, on the other hand, both HRs and FARs will be lower in the conservative case than in the liberal case, meaning that SDT bias indices will be greater (i.e., more conservative). Again, the SDT indices provide accurate information.

The distinction depicted in Figures 9.4 and 9.5 is of critical importance to researchers wanting accurate measurement of bias. If one is using P_{rc} to assess criterion shifts, it must be ensured that the confidence criteria do not shift (with or without RC) otherwise P_{rc} will provide the wrong information. For the particular case under scrutiny, P_{rc} indicated a shift to RC when no such shift has occurred, but P_{rc} will give misleading information (regarding the amount of subjective confidence associated with the RC) in *any* experimental circumstance in which the confidence-criteria shift. Under what circumstances might this happen? The answer to this question lies mostly with methodology. Koriat and Goldsmith (1996, Experiment 1), for example, gathered a single confidence rating in an initial forced-report stage of the experiment. Participants were then given the test again only this time with a report option (i.e., the option to pass) and high- versus low-incentive instructions. With this methodology, the confidence scale is fixed, and the only way that incentives can vary RC in phase 2 is if different levels of confidence are chosen to serve as RC. Thus, RC (and P_{rc}) shifts whereas the confidence criteria do not.

However, there are myriad other methodologies that can be used to investigate the strategic regulation of accuracy that do not ensure a fixed confidence scale between experimental conditions. For example, Higham (2007, Experiment 2) had participants answer questions from a Scholastic Aptitude Test. Participants were required to answer all questions, but were given a report option, which they could designate by selecting either *Go for Points* or *Guess* on the computer screen. The *Go for Points* Decision indicated willingness for the answer to be potentially subject to a penalty. For example, in the low-incentive groups, a correct answer with a *Go for Points* decision would earn a point, but an incorrect one would lose 0.25 points. Answers assigned *Guess* neither gained nor lost points, regardless of their accuracy. Feedback was manipulated such that for the feedback group, the points gained/lost on the previous question and the cumulative point total were displayed throughout the experiment, whereas for the no-feedback group, no points were shown. Participants also rated confidence in the accuracy of their answer on a six-point scale (6 = *high*; 1 = *low*). Because both confidence and report (*Go for Points*) decisions were manipulated between subjects, it is conceivable that the confidence criteria shifted between the groups with or without RC.

To assist in determining the actual scenario, ROC analysis is beneficial, the results of which are shown in Figure 9.2 for the feedback and no-feedback groups (that received low incentives). The different points on the curve represent criterion 'slices' of the evidence distributions. The points for reported answers given a confidence level of 6, 5 or higher, 4 or higher, and 3 or higher are labelled in the figure along with RC for both the feedback and no-feedback groups.

First, note that the two curves overlap indicating that feedback is having no effect on resolution. More important, also note that RCs for the two groups are located in approximately the same place on the ROC, suggesting that feedback is not having an effect on RC either. In support of this conclusion, C and lnß (both SDT indices of bias) did not vary as a function of feedback if measured at RC, both ps > .05. Conversely, note a different pattern for the confidence criteria. Compared to the no-feedback group, the feedback group's ROC points for reported answers assigned confidence levels 3–6 are further to the bottom-left (i.e., more conservative region) of the ROC. Consistent with this observation, SDT bias indices C and lnß were both significantly more conservative at the 3, 4, and 5 confidence criteria in the feedback group compared to the no-feedback group, all ps < .05.

Together, these analyses of the ROC suggest that RC remained static, but there was a shift towards conservativeness for the confidence criteria as a function of feedback, a scenario similar to that depicted in Figure 9.5 (although see below for a more complex case that is ambiguous with this interpretation). If instead the confidence scale had not shifted, then there would be no separation of the ROC points between the feedback conditions. Most critically, what happens to P_{rc} in this scenario? Does it mistakenly suggest a RC shift towards liberality as predicted by Figure 9.5? The answer is yes: there was a highly significant decrease to P_{rc} in the feedback group (2.26) compared to the no-feedback group (2.94), $p < .001$, despite the fact that RCs corresponding to each group are located in almost exactly the same place on the ROC.

To eliminate the possibility of obtaining deceptive values of P_{rc}, researchers using QAP methodology could adhere to the specifics of Koriat and Goldsmith's (1996) design in which confidence-rating assignment cannot vary with conditions. However, this solution is both limiting and denies the possibility of investigating some potentially interesting aspects of accuracy regulation, such as how experimental variables affect setting of the confidence criteria compared to RC.

The critic may argue that SDT bias indices can also be misleading because they sometimes indicate more liberality or conservativeness when, in fact, no criterion shift has occurred. However, this criticism assumes that SDT bias measures are process-pure indices of criterion placement, which most are not. For example, consider a case in which the correct item distribution in Figure 9.1 shifts upward while the other distribution and the criterion remain fixed. The shift will cause the bias index C to indicate more liberality, but why would this be if the criterion has not moved? The answer lies in knowing specifically what C is measuring, which, as noted above, is the distance between the intersection point of the evidence distributions and the criterion: As the correct item distribution shifts upwards, the intersection point too moves, situating the unmoved criterion further to the left of the intersection point (i.e., more liberality). This example demonstrates that if one is specifically concerned with pinpointing criterion shifts but there is a chance that the distributions can move, C is not a good choice. Generally speaking, in the absence of specific knowledge about what a particular bias index is quantifying, it is probably better to investigate criterion shifts using the raw HRs and FARs (as in the examples above) rather than using a bias measure per se.

Calibration curves

Calibration curves are a tool available to QAP researchers that may help identify cases in which a confidence-scaling shift has occurred like that shown in Figure 9.5. In constructing a calibration curve, the actual proportion of correct answers is computed for separate bins of assigned confidence. If one then plots actual proportion correct (y-axis) against confidence levels (x-axis), well-calibrated participants will have a calibration curve that follows the major diagonal. Over- and underconfident participants will have calibration curves that fall under and over the major diagonal, respectively. For a confidence-scaling shift like that depicted in Figure 9.5, calibration curves would show more underconfidence for the conservative condition compared to the liberal condition (i.e., the actual accuracy for any given level of confidence would increase).

To determine whether calibration curves can be used to alert QAP researcher to a confidence-scaling shift, it is necessary to consider all available QAP parameters that would be available under such circumstances. These parameters would be: (a) no change in resolution, (b) lower $P_{rc,}$ and (c) more underconfidence on the calibration curve. These parameters would presumably lead to the erroneous conclusion that there has been a change to both monitoring (based on calibration analysis) and control (based on P_{rc}) between the liberal and conservative conditions when, in reality, there has only been a change to control (criterion setting). On the other hand, the same conservative scaling shift would produce the following type-2 SDT outcomes: (a) no change in resolution (b) the same HR and FAR at RC, and (3) lower HRs and FARs at the confidence criteria.

Together, these should be correctly interpreted as an issue of control, in particular, a change in the subjective confidence associated with the confidence criteria.

Distribution versus criterion shifts

Despite the fact that type-2 SDT analyses are better able to identify confidence-scaling shifts of the sort depicted in Figure 9.5 than QAP analyses, there are cases that are problematic for type-2 SDT as well. However, these cases have less to do with indices being misleading (as with P_{rc}), but more to do with parameter output being ambiguous with other, alternative cases. For example, because HRs and FARs are always measured with respect to the evidence distributions, both distributions shifting downwards would have exactly the same effect as all criteria shifting upwards, assuming, of course, that the size of the shift in the two cases is the same. In particular, both cases would produce for type-2 SDT: (a) no change in resolution, (b) lower HR and FAR at RC, and (c) lower HRs and FARs at the confidence criteria. Does QAP fare any better in discriminating these cases? Unfortunately, it does not. For QAP, both a downward dual-distribution shift versus an upward lockstep RC/confidence criterion shift would reveal an equally ambiguous set of parameters: (a) no change in resolution (b) no change in $P_{rc,}$ and (c) increased underconfidence on calibration curves.

In summary, there are patterns of parameter output for both type-2 SDT and QAP that do not definitively identify the state of the underlying model (e.g., whether a dual-distribution shift has occurred or a shift of the criteria). These ambiguous cases will increase in number if one considers *shift combinations*, such as the distributions shifting in conjunction with the criteria. In this vein, the confidence-criteria shift depicted in Figure 9.5 is ambiguous with a more complex alternative case in which both distributions shifted downwards *and* RC shifted in lockstep with the distributions, whereas the confidence criteria remained static. Naturally, this scenario is less parsimonious than the Figure 9.5 illustration and it would be necessary to consider the specifics of the particular experimental circumstances to decide whether a more complex interpretation is warranted by the data. For the feedback manipulation under consideration (Higham, 2007), it is plausible that it caused participants to adopt higher levels of subjective confidence when assigning confidence ratings (perhaps because feedback on errors made participants more cautious about assigning high confidence: Figure 9.5), but it is not clear why it would decrease the mean subjective confidence of both the correct and incorrect candidate answers (dual-distribution shift) and decrease RC the same amount. Nevertheless, acceptance of more complex interpretations that are ambiguous with the most parsimonious one may be justified under other circumstances.

The second, quite separate, issue pertains to whether the indices associated with each approach provide misleading information. Some type-2 SDT bias indices (e.g., C) can be misinterpreted if they are taken at face value to indicate criterion placement. However, as long as researchers are clear about what the index is measuring and there is careful observation of the HRs and FARs that give rise to bias values, then these problems can be mostly avoided. On the other hand, P_{rc} will be misleading in *any* circumstance that the confidence-criteria shift. Depending on the nature of this shift, P_{rc} can overestimate, underestimate, or not detect at all, actual changes to RC placement, as well as signify changes to RC when it has not been displaced.

Conclusions

In this chapter, I have outlined a powerful approach for investigating monitoring and control processes in metacognitive research. I have shown that the approach is versatile, and easily extendable into new territory. Indeed, the application of type-2 SDT methods to accuracy discrimination is quite straightforward for researchers already familiar with more traditional SDT methods and its valuable analytic tools. These tools, which include ROC analysis, allow the researcher to define the evidence model that is most accurate given the experimental data. Hence, it is no longer necessary to rely on a single measure of resolution such as G, or a single measure of bias such as P_{rc}, to use in all circumstances, which has been the tradition in metacognitive research. Instead, researchers can use the evidence model to identify the best measures of resolution and bias for that particular data set.

The future of type-2 SDT in metacognitive cognitive research may hinge on the extent to which it can be applied to wide range of problems, not just the strategic regulation of accuracy. Recent extensions to JOL data where there is no report option are promising, but potentially, other metacognitive judgements such as FOKs could also be analyzed with SDT as long as accuracy discrimination is involved.

Notes

1. Anyone familiar with Bruce's extensive body of work will recognize NOBAL and FURIG as prototypes in Bruce's early work on categorization (Whittlesea, 1987). HENSION and STOFWUS, on the other hand, formed the basis of discrepancy-attribution hypothesis (e.g., Whittlesea & Williams, 1998).
2. I describe their SDT analyses as type-2 because, fundamentally, the JOL task is to discriminate between correct (later remembered) and incorrect (later forgotten) responses, and hence involves accuracy discrimination. Originally, type-2 SDT was proposed in conjunction with type-1 stimulus-based discrimination, explained in more detail below. However, I do not believe that there is anything to limit the type-2 accuracy-discrimination task to being only about the accuracy of a type-1 decision; it can pertain to any response. Another difference between

JOL research and more traditional SDT tasks is that JOLs are prospective, whereas RCRs are typically used in SDT. Nonetheless, in my view, as long as both RCRs and JOLs are used to measure accuracy discrimination, the tasks that incorporate them are necessarily type 2.

3. These candidate responses might be generated after an interrogation of long-term memory in response to a probe such as a general knowledge question or a cue in a cued-recall task. Alternatively, they may be defined by the task, as in the case of old/new recognition memory or multiple-choice testing.
4. The arguments are only partially correct with respect to type-1 SDT for the following reasons. The confidence ratings that participants typically make in old-new recognition tasks are about the *oldness* of the item, not about the *accuracy* of the response. For example, participants might rate the oldness of the item on a scale ranging from 'certainly new' to 'certainly old.' To rate the accuracy of the response, participants might rate the correctness of the old–new decision on a scale ranging from 'guessing' to 'certainly correct.' The two scales differ in that the former is stimulus-contingent (i.e., the ratings are about stimuli), and hence type-1, whereas the latter is response-contingent (i.e., the ratings are about the accuracy of responses), and hence type-2. As a result, the confidence ratings have very different meanings in the two cases, just as the discrimination measures do (see Higham et al., 2009, for discussion). The extent to which (type-1) confidence ratings discriminate between old versus new items in old–new recognition tasks is not actually a measure of resolution, but is a measure of bias-free recognition accuracy.
5. There is nothing about type-2 SDT that necessitates an assumption of equal-variance Gaussian evidence distributions. However, I will stick with the equal-variance case throughout this chapter simply because there is evidence that such a model sometimes fits type-2 data. In addition to Higham's (2007) RCR data, the type-2 ROCs based on JOL *recall* data published in Masson and Rotello (2009; based on Weaver & Kelemen, 2003) and Benjamin and Diaz (2008; based on Diaz & Benjamin, 2008) were not skewed and curvilinear. Further support for the equal-variance Gaussian model is found in Masson and Rotello's table 4: unlike *G*, *d′*, which assumes equal-variance, rank-ordered the recall conditions the same as d_a, which is more general. Similarly, in Benjamin and Diaz's figure 4, the linear z-ROCs had slopes of almost exactly 1.0. Little is yet known about type-2 ROCs, but it may be that the assumption of equal-variance will not be violated as often as with type-1 recognition, meaning that *d′* may often be suitable as an index of resolution.
6. Bias is sometimes used to refer to overall *calibration*, the degree to which rated confidence corresponds to actual accuracy. However, the term used in this current context refers to the bias that accompanies the measurement of resolution in the type-2 discrimination task.
7. Graphically, the HR and FAR for RC (or for any of the confidence criteria) can be estimated by examining the proportion of the correct (HR) and incorrect (FAR) candidate distributions that fall at or above the criterion on the subjective confidence scale.

10

Quantity-Accuracy Profiles or Type-2 Signal Detection Measures? Similar Methods towards a Common Goal

Morris Goldsmith

Introduction

In the previous chapter, Phil Higham follows up on previous work showing how methods based on type-2 signal detection theory (SDT) can be used to study the strategic regulation of memory performance, and compares some of the advantages and disadvantages of this approach to Koriat and Goldsmith's (1996b) Quantity-Accuracy Profile (QAP) methodology. I am glad to have the opportunity to comment on some of the points made in that chapter. I am also glad to be able to participate in this volume honouring Bruce Whittlesea, who among his many significant contributions to the study of memory, has done much to emphasize the critical role played by post-retrieval evaluation and decision processes in remembering (e.g., Whittlesea, 2002; Whittlesea & Williams, 2001a, 2001b).

That emphasis is shared by the type-2 SDT framework described by Higham in his chapter, and by the metacognitive framework that Asher Koriat and I originally developed (Koriat & Goldsmith, 1996b) and subsequently extended (Goldsmith & Koriat, 2008) for studying the strategic regulation of memory reporting, and more generally, for conceptualizing and assessing the contributions of metacognitive monitoring and control processes to memory accuracy and quantity performance. Thus, before addressing some of the specific differences between the two approaches, I would first like to stress the common view of memory that is fundamental to both frameworks. The basic assumption is that in the process of remembering, people do not simply spill out all of the information that comes to mind. Rather, between the retrieval of information on the one hand, and overt memory performance on the other, lie metacognitive monitoring and control processes that are used to strategically regulate the accuracy and quantity of the information that is reported. Hence, memory performance under free-report conditions – conditions typical of real-life remembering, in which one has the option to respond 'don't know' – depends not only on the ability to access and retrieve the solicited information, but also on the ability to effectively monitor the likely correctness of that information and choose an appropriate control policy (report criterion) based on competing incentives for accuracy and informativeness. These metacognitive contributions to free-report memory performance have been demonstrated in simulation analyses (Higham, this volume; Koriat & Goldsmith, 1996b) and in empirical studies examining the effects of various manipulations and group (e.g., developmental) differences, and how these effects and differences are mediated (see Goldsmith & Koriat, 2008 for review).

Once the potential contributions of post-retrieval monitoring and control processes to memory performance are acknowledged, it is crucial to have a way of isolating and examining these contributions. The QAP method that we proposed (Koriat & Goldsmith, 1996a, 1996b) is based on a two-stage procedure including both forced and free responding: Participants are required to provide a best-candidate answer to each recall cue/question or recognition probe (forced report), together with a confidence rating reflecting the assessed probability that the answer is correct, but are also allowed to decide whether or not they want to volunteer the answer (free report), typically under an incentive structure in which points are gained for each correct volunteered answer, a penalty is paid for each incorrect volunteered answer, and there is no penalty – but also no gain – for withheld answers. Using this procedure, the various components contributing to free-report quantity and accuracy performance can be isolated (see Goldsmith & Koriat, 2008 for more details):

1. *Retrieval/retention* is evaluated in terms of forced-report proportion correct.
2. *Monitoring effectiveness* is evaluated in terms of:
   a. *calibration bias* (under- or overconfidence) – the difference between mean assessed-probability-correct and actual forced-report proportion correct;
   b. *monitoring resolution* – the extent to which the confidence judgements distinguish between correct and incorrect best-candidate answers. This relationship that can be indexed by various measures, including but not limited to the within-participant Goodman-Kruskal gamma correlation (recommended by Nelson, 1984), the adjusted normalized discrimination index (ANDI; recommended by Yaniv et al., 1991), and d' or d_a (recommended by Higham, 2002; Masson & Rotello, 2009).
3. *Control policy* (report criterion) is estimated in terms of P_{rc} (report criterion probability) – the assessed probability that best reflects the level above which answers are reported/volunteered, and below which they are withheld.

4. *Control sensitivity* – the extent to which the report/withhold decisions are in fact based on the monitoring judgements, indexed in terms of the within-participant (gamma) correlation between confidence and the volunteer/withhold decisions.

With respect to this list of component measures, I now address some points of agreement and contention between the QAP and type-2 SDT methodologies for assessing the memory and metamemory components contributing to free-report memory performance.

Retrieval/retention

Let us begin with a point of agreement: Both approaches use the forced-free paradigm just described, and then use forced-report percent correct to index the contribution of 'memory' per se, untainted by the contributions of post-retrieval monitoring and report control. This use, however, deviates from the standard (type-1) SDT approach, in which two basic parameters are derived, one (d', A', etc.) reflecting 'true memory' and the other (β, B''_D, etc.) reflecting the contribution of 'extraneous' decision processes. Because three measures (at least) rather than two are needed to isolate the components contributing to free-report performance, the application of type-2 SDT in this context requires that some adjustments be made, both conceptually and methodologically, in line with the QAP approach.

Monitoring effectiveness

With regard to the effectiveness of monitoring, the QAP approach distinguishes between the absolute aspect, captured in terms of calibration bias (over- or underconfidence), and the relative aspect, captured in terms of resolution or discrimination accuracy. This distinction has a long history within the literature on judgement and decision making (e.g., Lichtenstein et al., 1982; Yates, 1982), and has been adopted in the metamemory literature as well (e.g., Nelson, 1996b; Schraw, 2009). In contrast, the type-2 SDT approach conceptualizes and measures monitoring effectiveness solely in terms of its relative aspect – resolution or discrimination accuracy. With regard to this aspect, there has been a great deal of discussion recently regarding the advantages and disadvantages of different measures, some associated with the SDT framework (e.g., Benjamin & Diaz, 2008; Masson & Rotello, 2009; Rotello et al., 2008), adding to previous discussions of similar issues (e.g., Liberman & Tversky, 1993; Nelson, 1984, 1996b; Schraw, 1995; Yaniv et al., 1991; Yates, 1982). In the context of comparing the QAP and type-2 SDT approaches, this appears to be a relatively minor issue. As noted above, the QAP method is not tied to any particular measure, and there is nothing to preclude the use of SDT-based measures to supplement or replace other measures (e.g., gamma), if indeed these turn out to be the better choice.

The different treatment of the absolute aspect of monitoring, however, appears to constitute a fundamental difference between the two approaches. In fact, by the type-2 SDT approach, over- or underconfidence is not treated as an aspect of monitoring, but rather, of 'control,' captured by differences in the confidence criteria used to assign explicit numeric or linguistic scale values to specific levels of subjective evidence ('true' confidence). Clearly, however, differences in elicited subjective probability values, and hence in over- or underconfidence, are not merely a matter of scaling differences (Wallsten & Budescu, 1983; Wallsten et al., 1993). The source of the problem appears to be that in type-2 SDT, much of what is conceived as 'monitoring' in the metacognition literature is hidden away from view in the presumed mapping function that translates psychophysical evidence (or 'cues,' e.g., retrieval latency and fluency, cue familiarity, accessibility of supporting and competing information, and so forth; see Brewer & Sampaio, 2006; Koriat, 2008; Wixted & Stretch, 2004) into subjective confidence. Thus, theoretically interesting differences in absolute monitoring are expressed as 'distribution shifts,' in which there is an overall increase or decrease in the true subjective confidence levels attached to incorrect and/or correct answers. Yet, distribution shifts are notoriously difficult to distinguish from actual changes in response criteria (e.g., Starns et al., 2007; Wixted & Stretch, 2000), as both of these are picked up as changes in response bias parameters, and therefore, differences in monitoring may mistakenly be attributed to changes in control (see further discussion below). Indeed, as of yet, the type-2 SDT framework offers no standard way of examining or measuring differences in absolute monitoring accuracy between conditions or populations (e.g., age differences). This limits the usefulness of the framework for addressing situations in which subjective confidence and actual performance dissociate (e.g., Busey et al., 2000; Kelley & Lindsay, 1993; Weingardt et al., 1994).

Control policy (report criterion)

Another point of contention, which is the one emphasized in Higham's chapter, concerns the way in which report criterion is conceived and measured in the two frameworks. Higham rightly points out that the QAP measure of report criterion, $P_{rc,}$ is calculated with respect to the numeric values of the subjective probability ratings elicited from the participants, and that this makes the P_{rc} susceptible to scaling issues. However, it is precisely because the various type-2 SDT measures of report criterion are *not* tied to scale values of subjective confidence that these measures become susceptible to the misinterpretation of evidence/confidence distribution shifts, in which the overall level of true subjective confidence relative to the actual correctness of the answers changes between conditions (reflecting differences in over- or underconfidence), without any change in the true report criterion. Because P_{rc} is calculated relative to the elicited confidence values, its value is unaffected by confidence-distribution shifts, whereas all type-2 SDT measures of report criterion are affected.

Fortunately, both QAP and type-2 SDT offer additional diagnostics that can be used to help gauge whether confidence-scaling or -distribution shifts have occurred. First with regard to the risk that a confidence-scaling shift may be influencing the QAP measure of report-criterion, P_{rc} (Higham's Figure 9.5), the threat of a pure scaling shift is signalled whenever a change in P_{rc} (e.g., towards liberality) is accompanied by a parallel and opposite change in calibration bias (e.g., towards underconfidence). Any change in P_{rc} that is not accompanied by a change in calibration bias, or which is accompanied by a change in the same direction (e.g., more liberal P_{rc} accompanied by greater overconfidence), cannot be due to a confidence-scaling shift alone. However, even in the first case it is possible that a true shift in report criterion, rather than a scaling shift, has occurred. As Higham points out, the situation depicted in his Figure 9.5 would also result from a downward shift in the subjective confidence distributions, together with a parallel downward shift in the report criterion. Such a scenario would be reflected correctly in the QAP measures, as a change in calibration bias (less overconfidence) and in P_{rc} (more liberal report criterion), but wrongly in the type-2 SDT measures, as a shift in confidence criteria (confidence-scaling shift) with no change in report criterion.

Which underlying scenario is more plausible: a pure scaling shift or a joint distribution and report-criterion shift? This would seem to be a judgement call, and must take into account not only 'parsimony,' but also the specific memory variables involved, plausible theoretical arguments, and common sense. In his chapter (Chapter 9), Higham discusses a pattern of results taken from Higham (2007), comparing a group of participants given feedback about the correctness of their answers (to SAT-type questions) to a no-feedback group, suggesting that a confidence-scaling shift occurred as a result of the feedback, and that this was wrongly picked up by P_{rc} as a shift in report criterion. Acknowledging that those results are ambiguous with regard to whether they reflect a confidence-scaling shift or a joint distribution and report-criterion shift, Higham nevertheless concludes that the former scenario is more plausible than the latter. But is it? The scaling interpretation assumes that feedback changed the way in which the participants assigned numbers to their true levels of subjective confidence, without actually changing their subjective confidence (or the report criterion). Why should that occur? In contrast, the distribution-shift interpretation holds that feedback about the correctness of their answers actually made the participants less (over-)confident about the correctness of their answers (a confidence-distribution shift), and that in order to continue to volunteer, rather than omit, a reasonable number of answers, the report criterion was relaxed by a corresponding amount. This interpretation seems quite plausible to me, though because the QAP approach was not used in that study, some of the data needed to evaluate it (e.g., mean confidence and overconfidence scores) are not available.

Turning now to the diagnostics that the type-2 SDT approach offers to signal and potentially avoid the threat of a confidence-distribution shift, these too involve examining the overall pattern of indices, and not just the measure of report criterion per se. Thus, although one might naturally misinterpret a change in a type-2 SDT bias measure as reflecting a shift in report criterion when actually it reflects a confidence distribution shift, an examination of the specific pattern of differences in the hit and false alarm rates, and in the confidence criteria, may signal that a distribution shift has occurred, and that the report criterion measure should not be taken at face value. Here too, however, the overall pattern of effects may be complex, and difficult to interpret unambiguously.[1]

In this regard, I disagree with Higham's claim that the SDT and QAP methods are equally threatened by the problem of distribution shifts. When a distribution shift has occurred, the QAP measures, taken at face value, will reflect the correct underlying scenario: a change in calibration bias (overconfidence), reflecting a change in the relationship between subjective confidence and the actual correctness of one's answers, with no change in report criterion. In contrast, the type-2 SDT measures, taken at face value, will reflect the wrong scenario: a change in the report criterion, suggesting that participants were more (or less) willing to risk providing wrong answers, and a change in the confidence criteria, suggesting that there was also a confidence-scaling shift. As Higham correctly observes, it is once again the possible scenario of a confidence-scaling shift – now tied to a lockstep shift in report criterion in the same direction – that threatens the QAP measures, which taken at face value, would wrongly indicate the occurrence of a distribution shift (i.e., an overall shift in subjective confidence that is not a mere scaling effect).

So, it seems that in weighing the advantages and disadvantages of QAP versus type-2 SDT, one really has no choice but to consider the relative risks of confidence-scaling shifts versus confidence-distribution shifts. Of course this is not an easy task, and it may be necessary to gather further empirical data regarding the conditions that are likely to produce such shifts. Here I will just note that psychometric issues concerning the measurement of subjective confidence have been studied and discussed extensively in the judgement and decision-making literature, including issues of reliability and validity that are essentially the same as with any other self-report measure (e.g., Wallsten & Budescu, 1983; Wallsten et al., 1993). In particular, there has been no suggestion that comparisons of mean confidence or mean overconfidence between groups or conditions be avoided because of the risk that these may merely reflect differences in usage of the confidence scale, a suggestion which would essentially preclude the use of confidence judgements (and all other subjective-report measures) in psychological research. Thus, Higham's suggestion that the use of P_{rc} in the QAP methodology should be restricted to situations in which the variables of interest are manipulated only after the confidence judgements have been elicited seems a bit odd. In fact, by the same token, one would have to recommend that the use of the type-2 SDT measures be similarly restricted, since that is also the only way in which a confidence-distribution shift can be completely precluded.

In this vein, I should note that although it is rather difficult to think of common memory manipulations that are likely to systematically affect the way in which subjective confidence is translated into numbers (i.e., a pure scaling shift), it is quite easy to think of those that are likely to increase subjective confidence in one's answers relative to the actual likelihood that they are correct (i.e., a single- or double confidence-distribution shift). This in fact is what most of the memory manipulations commonly used to elicit memory errors are designed to do, including the DRM paradigm (Roediger & McDermott, 1995), the misinformation paradigm (Loftus et al., 1978), imagination inflation (Goff & Roediger, 1998), and so forth. In each of these paradigms, a primary indication that the manipulation has 'succeeded' is that wrong answers are now held with high confidence (see, e.g., Weingardt et al., 1994). Unless there is a corresponding reduction in the confidence with which correct answers are held in these paradigms (and to my knowledge there is not), a confidence-distribution shift is implied (for discussion of this issue in the context of the DRM paradigm, see Starns et al., 2007; Wixted & Stretch, 2000).

Control sensitivity

Finally, turning to the issue of 'control sensitivity' (the extent to which the report control decisions are in fact based on subjective confidence), whereas this variable is explicitly included and measured within the QAP framework, so far it has essentially been ignored in the application of the type-2 SDT methodology to memory reporting. This omission is perhaps not arbitrary, as the signal detection framework is founded on the assumption that control (e.g., report) decisions are made by placing response criteria on distributions of subjective evidence. In contrast, a great deal of work in metacognition has emphasized the potential rift between monitoring and control – the possibility that one may have metacognitive knowledge or information that is not actually used in controlling one's cognitive behaviour, or that is perhaps overridden by other considerations (e.g., Ackerman & Goldsmith, 2008; Dunlosky & Connor, 1997; Schneider & Pressley, 1997). Along these lines, among university undergraduate participants, control sensitivity is generally very high, with within-participant gamma correlations between confidence and volunteering often averaging .95 or higher (Koriat & Goldsmith, 1996b)! Yet, when turning to special populations, a reduction in control sensitivity is sometimes observed, offering insights into the nature of metacognitive control deficits ensuing from old age (Pansky et al., 2009), mental illness (Danion et al., 2001; Koren et al., 2006), and psychoactive drugs (Massin-Krauss et al., 2002).

Conclusions

To sum up, Koriat and Goldsmith's QAP framework and Higham's type-2 SDT framework both share the common goal of studying and assessing the contributions of post-retrieval monitoring and control processes to free-report memory performance, and both have adopted very similar approaches to achieving that goal. Yet, there are disagreements about specific measures, and also about the ways in which some of the underlying theoretical components are conceptualized. Although it is true that the SDT framework has a long history of significant contribution to memory research, offering a rich arsenal of 'tried and true' tools that can be drawn upon, these tools were established in the application of type-1 SDT to old/new recognition memory, in which 'don't know' answers are not allowed. In contrast, the application of type-2 SDT in memory research has been relatively rare and fraught with conceptual and methodological confusions (Healy & Jones, 1973). In his work, Phil Higham has done an outstanding job of clarifying the interpretation of type-2 SDT measures in memory, and adapting them to study the strategic regulation of memory performance in free-report situations. Nevertheless, along with the advantages of adopting a relatively familiar and well established all-purpose framework, there appear to be some disadvantages in being tied to concepts and assumptions that take on different meanings than in the more dominant, type-1 framework, and which may not capture the full complexity of the topic under investigation, in particular, the nature of subjective confidence (e.g., Busey et al., 2000; van Zandt, 2000).

The metacognitive framework and methodology developed by Koriat and Goldsmith (1996b) for studying the strategic regulation of memory reporting is still very much a work in progress. Although originally developed to examine the mechanisms and performance consequences of report option – the option to volunteer or withhold individual items of information – the framework has since been extended to encompass an additional type of control – control over the precision or coarseness of the information that is reported, control over 'grain size' (Goldsmith et al., 2002; Goldsmith et al., 2005). The mechanisms underlying the control over grain size were found to be similar to those involved in the control of report option, and in fact point to the possibility of a common integrated model that can account for the joint use of both types of control in memory reporting (see Ackerman &

Goldsmith, 2008; Goldsmith & Koriat, 2008). In addition, an attempt is now being made to extend the framework and the QAP methodology to examine monitoring and control processes involved in the retrieval of information from memory, in addition to those involved in the evaluation and reporting of that information (Goldsmith et al., 2009; Koriat et al., 2008). It is not clear whether the type-2 SDT framework is flexible enough to support such extensions as well.

It is now the case that researchers interested in studying the strategic regulation of memory performance have alternative guiding frameworks and tools to choose from – each with its own specific advantages and disadvantages. Hopefully, the discussion embodied in this chapter and Phil Higham's chapter (Chapter 9) can offer some guidance to researchers who may choose to join the endeavour, as to which set of concepts and tools is most suited to their personal inclinations and research goals.

Note

1. I should note that I do not accept Higham's distinction between measures that are potentially 'misleading' and those that are merely 'ambiguous.' By this distinction, $P_{rc,}$ which purports to be an unambiguous measure of report criterion is 'misleading,' whereas the type-2 SDT measures such as β, C and B''_D, which do not purport to be unambiguous measures of report criterion, are merely ambiguous. I do not think it should be necessary to scour the literature in order to gauge how often SDT bias measures have been reported and taken at face value as measures of response criterion, regardless of the fact that they are indeed ambiguous when used in this way. Thus, the common use of SDT bias measures as an index of response/report criterion is 'potentially misleading.' Once one is aware of this problem, one can and probably should take Higham's advice and not use the type-2 SDT bias measures to index report criterion (or confidence criteria, for that matter). But this solution essentially leaves the type-2 SDT approach without a measure of report criterion, because inferring criterion changes from differences in the patterns of hit and false alarm rates, or by mathematical modelling (e.g., Starns et al., 2007) is unwieldy, and does not serve well as a dependent variable, for example, in examining the interaction between age and accuracy incentives on the report criterion setting (cf. Kelley & Sahakyan, 2003).

11

A Search for Influences of Feedback on Recognition of Music, Poetry, and Art

D. Stephen Lindsay and Justin Kantner

Introduction

The stream of consciousness has many tributaries. At any given moment some are dry beds, others gently burbling brooks, yet others gushing torrents. The main current is not always mindful of the sources of its flow, such that, for example, the waters of the wellspring of memory may mingle with the freshet of insight. But, to push the metaphor past the breaking point, we'd be at sea if we were completely unable to distinguish observation from expectation, reality from wish or fear, inherent ease from familiarity, etc. Thus the mind/brain needs mechanisms for monitoring (albeit imperfectly and at varying levels of specificity) the sources of influence on its own productions.

Bruce Whittlesea, whose work this volume honours, developed a theoretical perspective called *Selective Construction and Preservation of Experience* (SCAPE). SCAPE includes several theoretical constructs and can be applied to a number of domains (e.g., aesthetic judgements), but we focus on Whittlesea's 'fluency discrepancy hypothesis' in the context of recognition memory. This hypothesis holds that mental productions (thoughts, images, etc.) in response to recognition probes in particular contexts are generated in accord with the transfer-appropriate processing principle (Morris et al., 1977) and that those products are monitored and evaluated on the fly. If the fluency (i.e., ease and speed) with which a test probe is processed differs from expectations for that item in that context, the monitoring process attributes that discrepancy to some source. In Whittlesea's (2005) words, 'people attribute their self-evaluation (of the fluency of their mental productions) to some source that makes sense, given those aspects of the stimuli that are salient to them given the task and context and their intuitive causal theories.' Thus, for example, subjects are prone to false-alarm to new words on a memory test if those words are processed unexpectedly easily (e.g., as in the case of orthographically regular pseudowords such as 'hension'; Whittlesea & Williams, 1998, 2000).

SCAPE can be described as a major elaboration of Jacoby's attribution-making approach to memory (Jacoby & Dallas, 1981; Jacoby et al., 1989; Kelley & Rhodes, 2002). In a representative experiment, Jacoby and Whitehouse (1989) found that priming recognition test words created a bias to judge those words as old if the primes were presented so briefly that subjects were unaware of them. Primes presented for a longer duration had the opposite effect, creating a bias to reject the primed test word (presumably because their fluency was over-attributed to the prime rather than to oldness).

The source-monitoring (SM) framework is yet another approach to the general issue of how the mind/brain identifies the sources of its own productions (Johnson et al., 1993; Lindsay, 2008). Most work inspired by the SM framework has addressed the issue of how people differentiate between memories from different sources (such as remembering when and where a past experience occurred, its medium of presentation and modality of perception, the actors involved in it, etc.). The core idea of the SMF is that aspects of the sources of thoughts and images are inferred from the perceptual, semantic, and affective content of those thoughts and images. For example, memory information that comes to mind from a past conversation with your friend Lee might include information about the meaning of Lee's utterance, the sound of his voice, his appearance and that of the surrounding context, your thoughts and feelings at the time, etc., all of which provide clues to various dimensions of the source of this recollection.

Jacoby's attribution-making approach, Whittlesea's SCAPE, and the SM framework all hold that source inferences are usually made quickly and without conscious deliberation via heuristic processes. People occasionally experience uncertainty as to the provenance of a mental event ('Did I lock the door when I left or did I only think about locking the door?') and they may then use conscious strategies in an effort to identify source. Such deliberative processes may or may not yield a feeling of having successfully identified aspects of source, and if they do it may or may not be justified (i.e., we can identify an aspect of source reflectively and nonetheless be wrong). But most of the time we are no more aware of these monitoring processes than we are of the inferential processes involved in, say, seeing a three-dimensional world around us. The inferences generally unfold quickly and automatically in the flow of ongoing mental processing.

Inferential attribution-making processes have been the focus of most of the research inspired by the attribution-making, SCAPE, and SM approaches. But the processes that lead mental contents to come to mind in response to cues in contexts are also crucial to memory performance and experience. All three of the approaches emphasized here assume that transfer-appropriate processing (Morris et al., 1977) drives the generation or production of mental events. Thus the way subjects think about recognition probes at test influences the mental events that are generated in response to those probes, which in turn affects inferences about their sources.

The attribution-making approach, SCAPE, and the SM framework share the key notion that the mind/brain evaluates its own products and makes inferences about their sources. A casual reading might create the impression that these theoretical perspectives posit two discrete (non-overlapping) stages: First a mental product is generated and then later it is evaluated. But production and evaluation are always going on concurrently, and the outcome of monitoring processes affects generation processes, just as the characteristics of mental productions affect the outcome of monitoring processes.

In the interest of theory development, research inspired by these ideas has typically focused on errors in the attribution process; when a variable is shown to modulate the likelihood of false attributions that finding supports the general claim that there are attribution-making processes and sheds light on the specifics of their operation. But most attributions in everyday life are accurate. Mental products that have the properties of memories usually are memories and usually are experienced as memories, those with the characteristics of perception usually are based (primarily) on perception, etc.

The attribution-making, SCAPE, and SM approaches hold that people use various heuristics to attribute mental events to source. For example, Jacoby and Whittlesea both hold that in the context of a recognition memory test people tend to be biased to attribute unexpectedly fluent processing to the use of memory and hence tend to judge such items as old, whereas in another context they might attribute fluency to well-formedness. In research on another sort of source-attribution bias, Johnson, Raye, Foley, and Foley (1981) had subjects read some words aloud and listen to another person say other words aloud; later, if subjects false alarmed to non-studied test items they were biased to attribute that item to the other person rather than themselves (the 'It had to be you' effect). Presumably, this reflects a (perhaps unconscious) grasp of the fact that memory tends to be better for self-generated and enacted events than for passively perceived events (Cohen, 1989; Slamecka & Graf, 1978).

How do people come by such source-attribution heuristics and biases? Are these built-in tendencies fully available from an early age, or do they develop gradually during childhood? We do not know of any studies of children's memory informed by Jacoby's attributional approach to memory, nor by Whittlesea's SCAPE model, and these seem to us to be rich fields for future inquiry. The SM framework has inspired a fair number of child development studies (see, e.g., an edited volume by Roberts & Blades, 2000). Simplifying greatly, this research indicates that young preschoolers can do as well as adults at discriminating the sources of memories under some conditions (including, interestingly, conditions that lead to performance that is below ceiling in all age groups), but that they tend to do much more poorly than older children or adults when discriminations are particularly difficult (e.g., when candidate sources share many perceptual and/or conceptual characteristics or when delays are long).

The existence of developmental changes in SM performance suggests that experience may play a role in shaping the biases and heuristics that guide the attribution of mental events to particular sources. Indeed, such a claim is more or less explicit in Jacoby and co-authors' exposition of the attribution-making approach: It is because of the fact that use of memory facilitates processing that the system comes to be biased to interpret unexpected fluency as diagnostic of oldness. As direct support of this idea, Unkelbach (2006) demonstrated that a training experience in which experimentally manipulated high fluency was always linked to new items and low fluency to old items led to a reversal of the usual fluency effect; that is, subjects developed a bias to respond 'New' to highly fluent items and 'Old' to non-fluent items (see Unkelbach, 2007, for analogous findings with truth judgements).

Attempts to affect recognition with feedback

Inspired by the idea that even the most simple recognition memory task involves subtly nuanced and sophisticated generation and monitoring processes, in the late 1990s our lab began studying the possibility that recognition discrimination could be fine-tuned via accuracy feedback at test. In our most basic procedure, undergraduate subjects studied a long list of words presented one at a time on a computer screen and then later were shown a randomly ordered mix of studied and non-studied words presented one at a time for recognition judgements. In most of these studies, subjects responded on a 6-point scale from 'sure not studied' to 'sure studied.' A randomly selected half of the subjects were given accuracy feedback after each test response (e.g., if the item was new and they rated it on the 'new end' of the scale they were told 'Correct! That item was NOT on the study list'). Note that in this procedure there are no repeated study test cycles (unlike, say, Jennings & Jacoby's [2003] memory training procedure, in which effects of feedback might be mediated by changes in how items are studied). In our basic procedure there is just one study list followed by one test list in which each test item is presented once. Our hypothesis was that subjects could use feedback to fine-tune the way they engaged with probe words and/or the way they evaluated their memorial responses to probe words, leading to a gradual increase in accuracy over the course of the test.

A number of researchers have used feedback procedures in the context of various memory tasks, but only a few have tested the hypothesis that feedback can enhance recognition discrimination without item repetition or repeated study-test cycles (see Kantner & Lindsay, 2010, for review). Titus (1973) used CVCs and test lists in which only 20% (15 of 75) of the items were old. He found that subjects who received accuracy feedback after each test response adopted a more conservative response criterion but were not significantly more accurate than control subjects. More recently, Verde and Rotello (2007) and Han and Dobbins (2008) reported null effects of feedback on recognition sensitivity, but their studies were primarily oriented towards testing hypotheses regarding effects of feedback on response criterion (which they did obtain).

We set out to explore conditions designed to maximize the likelihood of observing effects of feedback on recognition sensitivity. Our initial strategy involved using multiple words from each of several semantic categories (e.g., birds, minerals), with each category represented by studied and nonstudied items, with the aim of making old/new discrimination rather difficult despite relatively deep processing at study. We predicted that subjects would use feedback to adjust the ways in which they engaged with probe words (thereby improving the diagnosticity of the 'echo' generated from memory) and/or to improve their monitoring/evaluation processes (thereby better discriminating between their own internal responses that were predictive of oldness versus newness). Thus we expected that over the course of the test subjects who received feedback would gradually come to be more accurate than control subjects in their old/new judgements.

There was not a whiff of a feedback effect in our initial experiment, which proved to be the first of an extensive collection of null effects (see Kantner & Lindsay, 2010, for a few select specimens). We tried various numbers of items, orienting tasks, delays, ways of giving the feedback, and so forth. Nothing, neither in terms of accuracy nor confidence. Feedback combined with proportion-old manipulations affected response criterion, presumably because feedback acts as a proportion-old instruction, but we found no other effects of feedback on recognition performance.

We then started employing less standard recognition situations. In one set of studies, we presented two study lists but instructed subjects they were to recognize only items from one of those lists even though the test list included some items from the other list, which subjects were to reject (as in Jacoby's opposition procedure; e.g., Jacoby et al., 1989). Performance on this test was unaffected by feedback. We also conducted studies (inspired by Higham & Brooks, 1997) in which, for some subjects, old words were names of large objects and new words were names of small objects (or vice versa for other subjects); compared to a mixed list condition, this confound led to slightly better recognition performance, but that effect was not amplified by feedback. In another set of studies, we used a variant of the Deese (1959) procedure in which subjects study the second through fifth most often generated exemplars of each of a large number of categories (e.g., precious metals: silver, bronze, copper, platinum) and then take a test that includes, as the critical items, the most often generated exemplar of studied (e.g., gold) and nonstudied (e.g., robin) categories (Seamon et al., 2000; Smith et al., 2000). As other researchers have shown, false alarm rates were high for the critical items from studied categories. Subjects who received feedback were just as likely to false alarm to critical items as were other subjects, even though every time they endorsed such an item they were told that they were wrong.

Over the years, while this recalcitrant project simmered on the back burner, we sought the advice of a number of renowned memory researchers. Some were incredulous that we would entertain the idea that recognition memory discrimination could be tuned by feedback at test without item repetition or repeated study/test cycles. From their perspective, recognition memory is so encapsulated and automatized (at least in adults) that it is impervious to such manipulations. Maybe they are right – to anticipate, to this day we have not firmly nailed down a reliable effect of feedback (without item repetition or repeated study/test cycles) on recognition discrimination (but read on for some tantalizing hints). Other scholars offered suggestions for ways of 'shining the light' (to quote social psychologist Lee Ross) so as to reveal such an effect. Chris Herzog, for example, speculated that subjects might benefit if instead of merely giving them accuracy feedback we also told them their response latency for each old/new response (on the ground that, empirically, quick responses are more often accurate than slow ones [e.g., Koppell, 1977]). We tested this idea a couple of times with no hint of an effect. Asher Koriat proposed that if we mixed markedly high and low frequency items, subjects receiving feedback might better tune in on the fact that recognition discrimination is more difficult for high frequency items (and hence at least respond with lower confidence on such items). Once again, we observed no such effect. Mike Masson suggested that we might obtain feedback effects if we tested with 2-alternative forced-choice recognition rather than yes/no recognition, but here too we found null effects of feedback.

Bruce Whittlesea speculated that subjects may already be optimized in their ability to recognize words; years and years of experience with these materials may have tuned the system as much as it could be tuned. Thus, Bruce proposed that we might observe feedback effects if we used stimulus materials that have a rich and complex structure but are unfamiliar to our subjects. The idea here is that such stimuli will give rise to ambiguous mental productions at test; being strange and complicated, the test items will be somewhat difficult to process and subjects' internal responses to them rather confusing (e.g., 'Is this a stimulus I heard before, or is it just that I sort of like it?'). Bruce's idea set us off on a journey that we recount in the remainder of this chapter. The first stop is Korea.

Korean melodies

To the untutored Western ear, traditional Korean music is passing strange. The instrumentation, scales, rhythms, and structures all depart from those of European music (visit http://203.252.231.26/ for samples). There are several major genres, some being orchestral in quality, others sounding folksy, and yet others reminiscent of free jazz. The music is evocative and challenging; listening to it is not a neutral experience. It struck us as the perfect medium with which to explore Bruce's idea that effects of feedback on recognition sensitivity might be observed with complexly structured yet deeply novel materials.

In an initial foray into this domain, we created 48 10 s mp3 files, each sampled from a different piece of Korean music (collected from the internet and the UVic library). We randomly divided them into two sets of 24 files each. Subjects were asked to listen to each of the melodies in one of the sets, with a few seconds between each one, under instructions to remember the music for a subsequent test. The music was played over the computer's built-in speaker. Then, after a 1m filled delay and the test instructions, each of the 48 melodies was played in a fixed, randomly intermixed order, and subjects made yes/no recognition memory judgements, giving their responses aloud. In this and all of the experiments we have conducted on the effects of feedback on recognition memory, the subjects were University of Victoria undergraduates who participated for optional bonus points in psychology courses or for a nominal payment. In each of the studies reported here, a quasi-randomly selected half of the subjects received accuracy feedback after each recognition judgement.

Mean recognition accuracy (d') in the feedback and control conditions is shown in the far left pair of bars in the upper panel of Figure 11.1; mean response bias (c) estimates are presented in the lower panel (negative scores indicate a conservative bias).[1] Error bars represent the standard error of the mean. Breakthrough! Just as Whittlesea had predicted, we obtained a statistically significant effect of feedback on the accuracy with which Korean melodies were recognized (with no effect of feedback on response criterion, which was essentially neutral).

Given the large number of null effects that had preceded this finding, we had little confidence in it, suspecting that it would prove to be a Type I error. We conducted a conceptual replication with a larger N and a number of minor improvements (e.g., stimulus presentation and feedback administration were more carefully controlled via E-Prime, responses were entered into the computer instead of being spoken aloud, and assignment of items to studied versus new status was randomized anew for each subject). Experiment 2 was conducted in a different semester and by a different experimenter than Experiment 1. We were delighted when the effect of feedback on accuracy replicated, as shown in the second pair of bars in the upper panel of Figure 11.1. This time there was also a significant tendency for subjects who received feedback to be more conservative on the test, but the crucial finding from our perspective was that, unlike all of our prior studies with words, feedback significantly improved recognition discrimination for Korean melodies, just as Bruce had predicted.

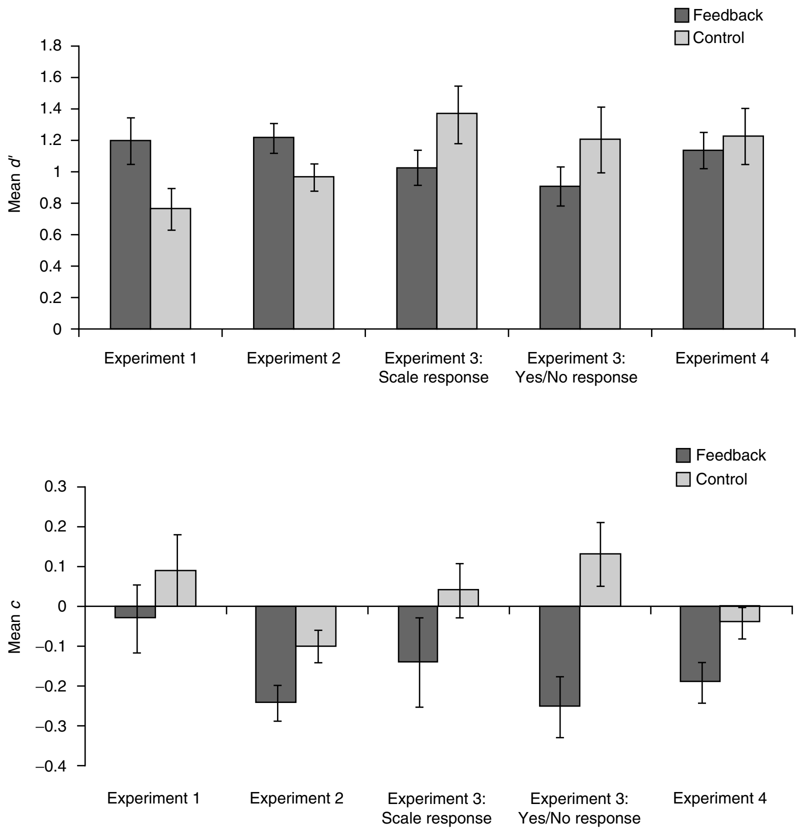

Figure 11.1 Upper panel: Mean values of *d′* (and standard errors of the mean) in the feedback and control conditions of the Korean Melodies experiments. Lower panel: Mean values of *c* (negative is conservative)

Maybe we should have written up this pair of studies and submitted them for publication, but we instead launched several lines of investigation designed to explore for feedback effects in recognition memory for other rich, temporally extended, and/or emotionally evocative stimuli (some of which are described below). For a variety of reasons we subsequently returned to our Korean melody (KM) materials, conducting three additional studies with the assistance of student Danette Dawkin. The first two KM studies, reported above, had used yes/no response alternatives at test, whereas most of our other feedback studies had used a 6-point rating scale. Experiment 3 was essentially the same as Experiment 2, except that we tested 40 subjects and a randomly selected half of them responded on a 6-point rating scale (whereas the remaining subjects made yes/no responses). Also, relative to

Experiments 1 and 2, KM3 was conducted in a different semester and by a different experimenter and subjects in KM3 used headphones to listen to the audio clips at study and test. That same experimenter in that same semester then conducted Experiment 4, a replication of KM3 including only the yes/no response conditions and eliminating the use of headphones on the off chance that they made a difference.

As shown in the three pairs of bars on the right of the upper panel of Figure 11.1, each of these three comparisons yielded a non-significant tendency towards an anti-feedback effect. That is, we failed to replicate the beneficial effect of feedback on accuracy obtained in the first two KM experiments, regardless of test format. The lower panel of Figure 11.1 shows that, as in Experiment 2, both test-form conditions of Experiments 3 and 4 yielded a significant effect of feedback on response criterion: Subjects who received feedback after each test response were more reluctant to judge melodies old than were subjects who did not receive feedback. We defer further discussion of the Korean melodies findings pending report of two other parallel lines of investigation.

Masterwork paintings

A great painting, like a provocative piece of music, can be emotionally moving in complex ways. One may, for example, be enthralled by the technical skill of a painting yet appalled by its subject matter (e.g., Rembrandt's *The Anatomy Lesson of Dr. Nicolaes Tulp*) or one may love the colours and textures of a painting but not really 'get' the forms (e.g., some might feel that way about Rothko's ephemeral paintings). The complexity and evocativeness of great paintings led us to speculate (along the lines of Bruce Whittlesea's suggestion) that we might obtain effects of feedback on recognition of paintings.

Jeffrey Toth (of the University of North Carolina Wilmington) has assembled a large set of standardized digital scans of great paintings for use in a different sort of memory training programme, and he kindly allowed us to use images from that set. For our initial study with these materials, we selected 102 paintings. We avoided super-famous works such as the Mona Lisa, and sampled a wide range of styles, artists, and subject matter. The experiment was conducted by student Kyle Mathewson. In the study phase, 54 paintings were presented one at a time for 1 s each with instructions to 'study each painting carefully and completely.' Immediately after the study phase, the test instructions were given, and then the studied paintings (less 6 primacy and recency buffers) and an equal number of non-studied paintings were presented one at a time in a novel randomized order and subjects used a 6-point scale to make confidence-weighted recognition memory judgements (sure new to sure old). As shown in the leftmost pair of bars in the upper panel of Figure 11.2, feedback had no effect on recognition

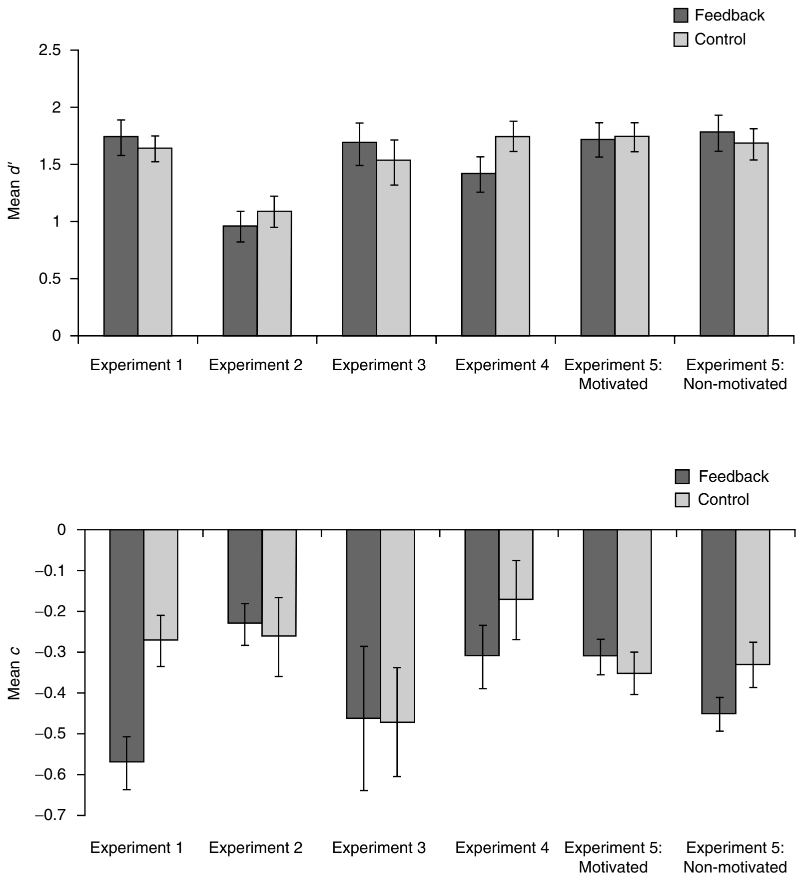

Figure 11.2 Upper panel: Mean values of *d′* (and standard errors of the mean) in the feedback and control conditions of the Paintings experiments. Lower panel: Mean values of *c* (negative is conservative)

sensitivity. But, as in the latter four Korean Melodies experiments, subjects who received feedback displayed a significantly more conservative response criterion than did control subjects.

Recognition accuracy was quite high in Paintings Experiment 1, and we speculated that this may have made it difficult to detect a beneficial effect of feedback on sensitivity. We therefore devised a new version of the experiment, in which all of the paintings were portraits. Relative to Paintings Experiment 1, Paintings Experiment 2 was conducted in a different semester by a different experimenter (student Elaine Blight), and included a larger number of study and test items (75 each, plus 10 primacy and recency buffers). As can be seen in Figure 11.2, the shift to using only portraits and the increase in the number of items apparently had the desired effect of substantially lowering response accuracy, but there was no effect of feedback on accuracy. Moreover, the large effect of feedback on response criterion observed in Paintings Experiment 1 (and in most of the KM experiments) was eliminated. Elaine Blight then conducted an 'exact' replication of Paintings Experiment 1. This Paintings Experiment 3 yielded no effect of feedback on accuracy or response criterion. Then Elaine conducted yet another study, which was identical to Paintings Experiment 2 (i.e., with only portraits) except that a deep orienting task was used during the study phase. Specifically, subjects in Paintings Experiment 4 were asked to rate whether the individual depicted in each portrait looked above or below average in friendliness, with an eye to increasing recognition sensitivity with the homogeneous portrait set. Again, neither sensitivity nor criterion effects of feedback were observed.

In a subsequent semester, new research assistants Brian Buchan and Alison Wegner conducted yet another study using the paintings materials in which we attempted to manipulate how motivated subjects were to perform well on the recognition test. Our speculation was that the experimenter who conducted the first Paintings study may have inspired subjects to try harder on the test, and that this may have led to the pronounced effect of feedback on response criterion. The findings of four of the KM studies and Paintings Experiment 1 suggest that when subjects receive feedback they try especially hard to avoid false alarms (at the cost of increased misses); we speculated that this feedback-driven conservative shift would be particularly strong if subjects were highly motivated to do well. But Brian and Alison found no effect of feedback on recognition test responses in Paintings Experiment 5, regardless of motivation condition.

In Paintings Experiment 5, we included a manipulation check at the end of the procedure, asking subjects to rate how motivated they had been to do well on the recognition test. That measure indicated that our attempt to manipulate motivation was ineffective (on a 7-point scale with 7 = *highly motivated*, means of 5.7 and 5.8 in the high- and low-motivation conditions, respectively). Self-reported motivation was also equivalent in the feedback ($M = 5.8$) and control ($M = 5.7$) conditions. Response bias was marginally more conservative for participants receiving feedback, but only in the low-motivation condition ($p < .07$), and no other remotely significant bias or sensitivity effects were observed. There was a modest and significant correlation between self-reported motivation and d' ($r = .25$, $p < .01$), such that individuals who reported higher motivation tended to be more accurate at discriminating studied from new paintings. But contrary to our speculation about motivation playing a role in the conservative bias, there was no significant correlation between self-reported motivation and c and the miniscule relationship we did observe was in the wrong direction ($r = .10$). At this point we downed brushes.

Poetry

T. S. Eliot wrote, 'Genuine poetry can communicate before it is understood.' Still inspired by Bruce Whittlesea's idea that effects of feedback on recognition might arise with complex, quixotic materials, we assembled a sample of lines of poetry by Rainer Maria Rilke. For example:

> The walls, with their ancient portraits, glide
> away from us, cautiously, as though
> they weren't supposed to hear what we are saying.
> From 'Before Summer Rain'

Our intuition was that intro psych students might find it challenging to read dozens of snippets of Rilke by themselves. We therefore created +/– 10 s audio clips of the lines being read aloud somewhat poetically – not overly so, but with a degree of feeling. During the study phase, each of 54 clips, randomly selected anew for each subject from a set of 96 clips, was played aloud, with a 500 ms pause between clips. Test instructions were presented immediately after the study list, and then each of the studied clips (less 6 primacy and recency buffers) was again played aloud, randomly intermixed with an equal number of non-studied clips. A total of 47 subjects responded to each clip on a 6-point scale from sure new to sure old. Half of the subjects received accuracy feedback following each response.

The mean recognition accuracy scores are shown in the first pair of bars in the upper panel of Figure 11.3. As in Korean Melodies Experiments 1 and 2, subjects who received feedback as they took the test had significantly higher recognition accuracy than did control subjects. As shown in the lower panel of Figure 11.3, we also once again found that subjects who received feedback were significantly more conservative than control subjects.

In two subsequent studies, identical to one another, the same study and test poetry snippets as in Poetry Experiment 1 were presented as text on the computer screen. At study subjects were instructed to read each poetry snippet silently to themselves and to study it for a later memory test. Likewise at test subjects read each snippet silently and responded with the 1–6 rating scale. There were 31 and 34 subjects in Poetry 2 and Poetry 3, respectively. As shown in Figure 11.3, this text-only version of the Poetry procedure yielded mixed effects of feedback. Feedback significantly impaired recognition sensitivity in Poetry 2 but not in Poetry 3. Also, as in many of the studies we have reported, subjects who received feedback were directionally more conservative in their recognition judgements, but in neither of these experiments did that tendency attain statistical significance.

As mentioned above, in Poetry Experiment 1 the Rilke audio clips had been read 'poetically'. We wondered if this emotive expressiveness may have played a role in giving rise to the beneficial effect of feedback on recognition

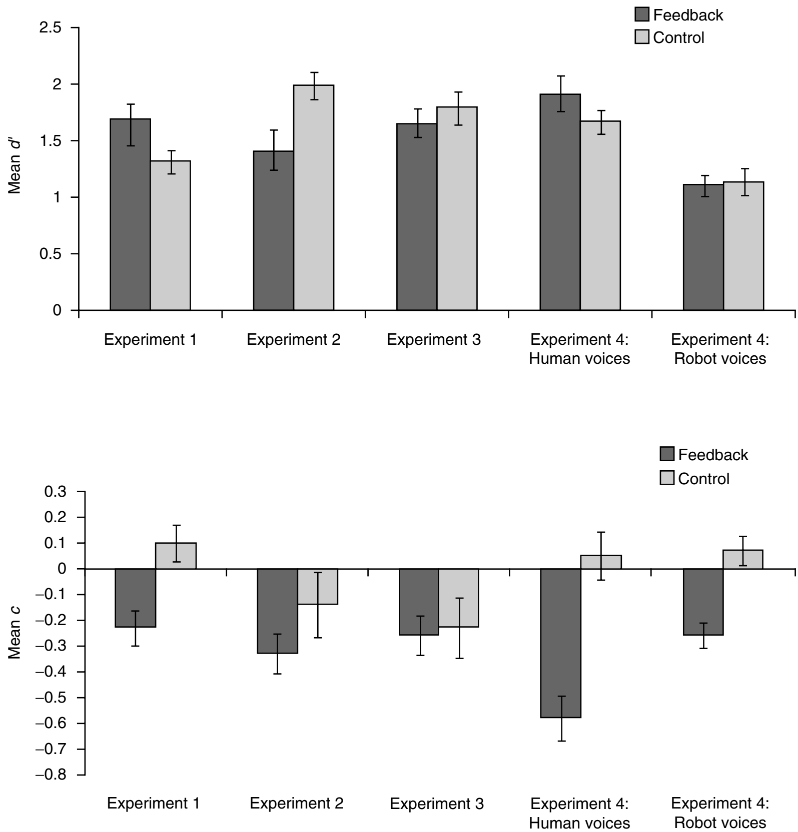

Figure 11.3 Upper panel: Mean values of *d′* (and standard errors of the mean) in the feedback and control conditions of the Poetry experiments. Lower panel: Mean values of *c* (negative is conservative)

accuracy in that experiment. To explore this possibility, we conducted a follow-up experiment in which half of the subjects heard the same audio clips as in Experiment 1 whereas other subjects heard the same poetry snippets rendered by a flat 'robot' voice (using TextAloud software). Experiment 4 Human directionally replicated Experiment 1, but the tendency towards an effect of feedback on *d′* did not approach significance ($p = .19$) and even this estimate exaggerates the tendency towards a feedback effect on sensitivity because *d′* values were skewed due to very low false alarm rates (see Appendix): In terms of hits minus false alarms, there was no difference at all between the conditions. Thus Poetry Experiment 4: Human Voices (which was essentially an exact replication of Poetry Experiment 1) failed to replicate the effect of feedback on sensitivity. In contrast, the feedback-based conservative-shift effect obtained in Poetry Experiment 1 was even larger in Poetry Experiment 4: Human Voices. In Poetry Experiment 4: Robot, there was no hint of an effect of feedback on recognition accuracy, but the conservative-bias effect was comparable to that observed in Poetry Experiment 1.

Summary sans conclusions

In this chapter, we have taken what we suspect is an unusual tack by reporting every experiment we have conducted along these particular sub-lines of research.[2] That is, aside from a few small-*N* pilot studies, this is an exhaustive catalogue of our studies of the effects of feedback on recognition of Korean melodies, poetry, and paintings. By Whittlesean standards the number of studies is modest, but it is sufficient to create a rather complex picture. Indeed, the pattern of results is maddeningly inconsistent. If you are holding your breath waiting for us to provide a coherent theoretical account for this patchwork of null and significant effects, you will soon turn blue. Nonetheless we think our findings make some important points.

Can feedback at test enhance recognition sensitivity? Is such an enhancement effect more likely to be observed with stimuli that are rich, complex, and poorly understood, as Bruce Whittlesea surmised? Maybe a little. Against the backdrop of the forest of null effects obtained in our studies with familiar words as stimuli (Kantner & Lindsay, 2010), the three or four isolated poplars of significant effects with Korean melodies and poetry clips stand out in stark contrast. But even if these effects are real, rather than Type I errors, the effect is far from robust. Indeed, a mega-analysis of the raw data collapsed across the 16 experiments revealed virtually identical mean d' values for subjects who did and did not receive feedback ($M = 1.469$ and 1.468, $F(1, 536) < .001$, $p = .99$, partial eta squared < .001). Also, in the Poetry studies in which there was at least some indication of better discrimination by subjects who received feedback, there was no clear indication that discrimination improved gradually over trials (we judged that the KM studies had too few trials to afford such analyses). Of course, it is always possible that larger and more consistent benefits of feedback on sensitivity would be observed under other conditions, but such conditions have thus far eluded us despite an extended search.[3]

In contrast to the rather murky findings in the analyses of recognition sensitivity, the data on response bias are strikingly consistent in two regards. First, subjects tended to be conservative, preferring misses to false alarms (or, to put it the other way, placing a higher premium on correct rejections than on hits). Of the 32 comparisons reported here, response criterion was at least directionally conservative in 26, and significantly so in 20. Collapsing across experiments and conditions, the tendency for c to be less than zero was significant and substantial ($M = -0.232$, $t(538) = -15.7$, $p < .001$). In all of these experiments, half of the test items had been studied, and there was no explicit incentive to value correct rejections above hits. Although we have not systematically assessed the matter, a casual review of our studies with familiar words indicates no such tendency; lacking relevant manipulations (e.g., of the base rate of old versus new items), in our studies with words response criterion has been near zero. We plan a future experiment designed to test for a hypothesized materials-based conservative-shift effect on recognition memory response bias.

The second clear finding pertaining to response criterion was that receiving feedback substantially increased subjects' conservative bias. This feedback-based conservative-shift effect was directionally present in 14 of the 16 experiments, and statistically significant in 7 (and 'marginal' in one more) of those comparisons. Collapsing across experiments, the difference in c between the feedback ($M = -.32$) and control ($M = -.15$) conditions was significant, $F(1, 536) = 44.17$, $p < .001$, partial eta squared = .08). Here again, this finding contrasts with the results from our studies with familiar words, in which response bias did not consistently differ between subjects who did and did not receive feedback.

Why should Korean melodies, poetry clips, and paintings give rise to a conservative response criterion? Perhaps the evident richness and distinctiveness of the stimuli leads subjects to believe that they should have good memory for them, and hence should be able to recollect a lot about studied items when probed. That is, subjects' expectations about what they should be able to remember about a test item might be exaggerated, leading them to reject a high proportion of test items. Or perhaps subjects believe that to claim to have experienced such a stimulus when one hadn't would be a particularly egregious kind of error. In pondering this issue, we were struck by the parallel to classroom situations in which students are loath to volunteer an answer unless they are very confident of it, perhaps following the proverb 'Even a fool, when he holdeth his peace, is counted wise, and he that shutteth his lips is esteemed a man of understanding' (Proverbs 17:28, King James Bible). Whatever the explanation, to the best of our knowledge no previous publication has reported evidence of a materials-based conservative-shift effect on recognition memory response criterion.

Why should feedback amplify the tendency towards conservative bias on tests of recognition memory for these sorts of stimuli? At first blush, one might have thought that feedback would correct, rather than exaggerate, response bias. After all, feedback informed subjects that they were more often erring by rejecting studied items than by false-alarming to new items. Yet the effect of receiving feedback was to increase that conservative bias. This is arguably consistent with the speculation that subjects were particularly motivated to avoid false alarms.

Whittlesea (2002, p. 112) wrote provocatively about response bias in recognition memory, arguing that although signal detection models often fit data beautifully they do not necessarily adequately capture the underlying processes that modulate participants' attitudes towards their own responses to test probes. More specifically, he proposed that criteria may shift fluidly from trial to trial, 'on the fly,' influenced by interactions between characteristics of the item, the person, and the context. Whittlesea's arguments here were on the cutting edge, as numerous researchers have addressed the issue of criterion flexibility in a variety of contexts in recent years (e.g., Curran et al., 2007; Dobbins & Kroll, 2005; Gallo et al., 2001; Heit et al., 2003; Hockley & Niewiadomski, 2007; McCabe & Balota, 2007; Miller & Wolford, 1999; Rhodes & Jacoby, 2007; for recent reviews of response bias in recognition memory, see Hockley, this volume; Rotello & Macmillan, 2008).

Caveat

The appalling lack of consistency across our studies in the effect of feedback on recognition discrimination provides an important albeit not novel lesson. Consider, for example, the first three of our Korean melodies studies. With respect to recognition sensitivity, KM Experiment 2 nicely replicated KM Experiment 1, but KM Experiment 3: Scale Response yielded an anti-feedback effect. The only planned, systematic difference between the latter two of these experiments was in the test response alternatives. If we (a) indulged in across-experiment comparisons and (b) did not also include in KM Experiment 3 a yes/no condition that produced exactly the same anti-feedback effect as did the scale response condition, we would have been mightily tempted to tell a just-so story in which responding on a scale reverses the beneficial effects of feedback. But KM Experiment 3: Yes/No was a quite close replication of KM Experiments 1 and 2, and we can only speculate as to what caused the reversal in the data pattern. Perhaps mere measurement error in one or both experiments. Similarly, whereas the first of our studies with paintings yielded the most dramatic effect of feedback on conservative bias, a near-exact replication of that study produced no hint of such an effect. Here again, we can only speculate as to why.

The scary thing about this is that it would be very easy for researchers collecting data in a noisy domain such as this to fall into error, knowingly or unknowingly. We could, for example, have made a very compelling case for the reality of effects of feedback on *d′* by omitting mention of the experiments that ‘didn’t work.’ Or we could have gone even further by reporting several studies with an effect of feedback on *d′* and one or two contrasting null-effect studies, interpreting the latter as setting boundary conditions on the effect of feedback on sensitivity. For example, by selecting from these studies one could tell a nice story to the effect that feedback enhances recognition sensitivity with temporally extended auditory stimuli such as Korean melodies and poetry, but not with static stimuli such as paintings. There may even be something to that story, but the fact that we several times failed to observe an effect of feedback on Korean melodies and poetry undermines confidence in it. At this point, despite substantial effort we don’t know what caused the effect to sometimes emerge and other times not.

One implication is that researchers are well advised to determine empirically whether or not the effects they find are replicable, ideally with more than one replication. The fact that a finding is statistically significant does not mean that it is replicable (Miller, 2009). It is not unusual for cognitive psychologists to tweak a procedure until it 'works', which is fine as long as one goes on to determine whether or not it works *consistently*. A related implication is that we must be cautious in making across-experiment comparisons. That point is almost too elementary to warrant mention, but the field is rife with insufficiently qualified across-experiment comparisons.

Last word

Pursuing this line of research has been an intellectual and emotional rollercoaster ride. The significant effect of feedback on recognition sensitivity in the first two Korean melody studies inspired by Bruce Whittlesea's suggestion produced a heady high that lingers on in memory. That this effect again emerged in the first of our poetry studies and (marginally) in the replication of that study (i.e., in 'Poetry Experiment 4: Human Voices') continues to intrigue and encourage us. But our replication failures (which occurred despite our use of highly standardized, computer-run procedures) have at times plunged us to the depths, and at this point we are still far short of understanding the effects of feedback on recognition discrimination.[4]

The fickleness of our discrimination effects was offset by our consistent findings regarding response bias. The suggestion of a materials-based conservative-shift effect (which to date rests solely on across-experiment comparisons!) was an unlooked-for bonus of this line of research. Likewise, the quite strong evidence of a feedback-based conservative-shift effect that emerged in these studies is a serendipitous discovery that may contribute to developments in theorizing about response criterion in recognition memory (Hockley, this volume; Rotello & MacMillan, 2008).

Research sometimes teaches us how little we know. Can feedback enhance recognition discrimination? Is such an effect more likely with certain kinds of materials than others, and if so what are the determining factors? What factors accounted for the inconsistency of the sensitivity effects across experiments? Why is bias more conservative with rich, complex stimuli such as music or poetry than with words? Why does feedback encourage a conservative response bias with such materials? Any ideas, Bruce?

Notes

We thank the many students who helped prepare stimulus materials and test research participants. We also thank Bill Hockley and Zolton Dienes for insightful comments on a draft of this chapter. And thanks to Bruce Whittlesea for being such an inspiration.

1. By the standard calculation of c (Macmillan, 1993), negative values signify a liberal response bias. We follow some authors (e.g., Snodgrass & Corwin, 1988) in removing the negative sign from the formula, reversing the conventional interpretation of c values. Hit and false alarm rates for all experiments are presented in the Appendix. We tried various measures of sensitivity and criterion. For the experiments in which subjects responded on a 6-point scale, we also analysed the data in terms of mean scale response. None of those analyses suggested any different story than did d' and c, so for simplicity we report only those two measures here. Email either author for the mean scale responses or the raw data.

2. We also conducted an experiment with Chinese characters (with the assistance of students Jeffrey Sun, Ben Shiner, and Danika Overmars), a study with pseudowords (with Ben Shiner), a study with 'one-liners' (e.g., 'A drunk man's words are a sober man's thoughts' or '43% of statistics are worthless'; with students Brian Buchan, Kyle Mathewson, and Nicky Schnare), and a study with photos of faces (with Nicky Schnare). None of these studies yielded an effect of feedback on recognition sensitivity or response bias (although there was a nonsignificant tendency towards a feedback-based conservative-shift effect in the study with faces).
3. Lane, Roussel, Villa, and Morita (2007) reported a fascinating experiment using feedback in the eyewitness misinformation domain. Subjects viewed a slide series depicting an event, were exposed to misleading suggestions regarding details in that event and either did or did not receive feedback on the first part of a test on which subjects were to identify the source (slides, narrative, both, or neither) of details from each of those sources. Accurate feedback selectively lowered false attributions of suggested details to the slides, without evidence of a general change in response criterion.
4. Zolton Dienes, in a review of a draft of this chapter, raised the intriguing possibility that fluke variations in item-order effects might have contributed to the inconsistency of our results.

Appendix

	H		FA	
Experiment	**Feedback**	**Control**	**Feedback**	**Control**
Korean Melodies Line				
Experiment 1	.708 (.035)	.675 (.030)	.278 (.039)	.396 (.041)
Experiment 2	.638 (.015)	.646 (.017)	.217 (.022)	.292 (.022)
Experiment 3: Scale	.640 (.028)	.755 (.032)	.280 (.044)	.275 (.039)
Experiment 3: Yes/No	.575 (.038)	.740 (.044)	.250 (.028)	.325 (.027)
Experiment 4	.638 (.030)	.700 (.028)	.231 (.021)	.272 (.030)
Paintings Line				
Experiment 1	.605 (.025)	.696 (.025)	.098 (.013)	.159 (.021)
Experiment 2	.596 (.029)	.602 (.042)	.249 (.027)	.229 (.028)
Experiment 3	.634 (.057)	.607 (.049)	.141 (.041)	.139 (.047)
Experiment 4	.651 (.035)	.740 (.045)	.170 (.031)	.155 (.020)
Experiment 5: Motivated	.689 (.028)	.689 (.021)	.144 (.016)	.131 (.016)
Experiment 5: Non-Motivated	.657 (.025)	.678 (.026)	.113 (.017)	.146 (.019)
Poetry Line				
Experiment 1	.722 (.025)	.766 (.020)	.175 (.028)	.310 (.028)
Experiment 2	.635 (.037)	.781 (.031)	.176 (.029)	.158 (.040)
Experiment 3	.706 (.023)	.730 (.031)	.162 (.025)	.174 (.035)
Experiment 4: Human Voices	.644 (.029)	.801 (.021)	.080 (.019)	.242 (.035)
Experiment 5: Robot Voices	.612 (.029)	.735 (.024)	.213 (.016)	.318 (.029)

12

Criterion Changes: How Flexible Are Recognition Decision Processes?

William E. Hockley

I have known Bruce for a very long time, though I would still be surprised to see him in my kitchen. I was not surprised to see him at conferences. Indeed, I always looked forward to seeing Bruce at meetings. I looked forward to hearing about his latest studies, and the clever experimental manipulations he was carrying out, and to get his advice. Bruce's contributions are far reaching, and his energy, enthusiasm, and passion for cognitive psychology contagious.

Introduction

An assumption of most accounts of recognition memory is that at least some proportion of responses are based on a decision process that is used to assess retrieved information. It is traditional to describe this decision process in the context of signal detection theory (SDT; see Macmillan & Creelman, 1991). In this framework, memory sensitivity is represented by the distance between the means of the old (target) and the new (distractor) distributions that vary along a strength-of-evidence dimension. As these distributions typically overlap, a criterion that bisects the evidence dimension is necessary. A single criterion provides the basis for yes-no recognition decisions such that test probes associated with a strength value above the criterion will be classified as old and those with a value below will be deemed new. Multiple decision criteria provide a means to further partition old and new responses into different degrees of confidence.

SDT has been highly successful in capturing the relationship between subjects' accuracy and confidence, and providing a framework for describing recognition performance. As Dobbins and Han (2008) have recently discussed, however, SDT provides a fairly limited description of recognition memory decisions in that it does not provide any general principles governing how subjects might set or change their decision criterion in different circumstances.

Assumptions concerning how the decision criterion is positioned, or how response bias might be changed, in different situations have important implications for understanding recognition memory. For example, one controversy involving criterion changes concerns the interpretation of responses based on Tulving's (1985) remember-know paradigm. In this procedure, subjects indicate whether old recognition decisions are based on recollection of a prior detail or details of a past experience (remember) or on a sense of familiarity in the absence of recollection (know). Researchers have used dissociations between remember and know responses to support dual-process accounts of recognition memory (see Gardiner et al., 2002). Other investigators have proposed that remember and know responses do not represent different bases of recognition decisions, but rather reflect different levels of confidence (e.g., Donaldson, 1996; Dunn, 2004; Hirshman, 1998; Wixted & Stretch, 2004).

Dunn (2004) demonstrated that a single-process SDT model could account for all of the dissociations between remember and know responses that have been taken as support for the dual-process view. In his fits, Dunn assumed that subjects could adjust the two decision criteria used to distinguish new from old and know from remember responses on the basis of different variables such as auditory versus visual presentation, word versus non-word stimuli, and one versus four study trials. In reply, Diana, Reder, Arndt, and Park (2006) questioned whether subjects are capable of adjusting their decision criterion in each experimental condition in the manner that Dunn assumed.

Although the factors that determine criterion placement, and the degree of flexibility that subjects have in shifting their criterion, are not fully understood, there is growing evidence that criterion changes can occur for a variety of reasons. In this chapter the research that has examined subjects' ability to set their decision criterion in order to try to optimize their recognition performance or to meet task demands is evaluated. Both the issue of how subjects initially establish their decision criterion, and the question of when and on what bases subjects might change their criterion during the course of the test sequence are considered.

Criterion placement: how do subjects set their decision criterion?

Research has shown that criterion placement is influenced by both external influences and subjective assessments of memorability.

Instructions and payoffs

Subjects are able to set their criterion to comply with instructions to be either more liberal or more conservative in their decision making. Postma (1999), for example, asked subjects to be either conservative (very certain of old responses) or lenient (choose old if subjects 'had even only a weak notion that they had studied it previously', pp. 70–71) in a study using the remember-know response procedure. The instructions did not influence discrimination between the two groups, but the lenient group had higher overall hit and false alarm rates. The hit rate for 'remember' responses and the false alarm rate for 'know' responses were also higher for the lenient group, consistent with the view that 'remember' and 'know' responses reflect different degrees of confidence (e.g., Dunn, 2004).

Although there are relatively few studies that have manipulated payoffs in tests of recognition, the available evidence shows that the provision of payoffs that reward one class of responses over another have their desired effect on criterion placement. Van Zandt (2000) varied payoffs across 12 study-test blocks in one experimental session. Prior to each test list, subjects were informed that correct 'new' responses would be worth 0, 1, 2, 3, or 4 points and correct 'old' responses would be correspondingly worth 4, 3, 2, 1, or 0 points (i.e., correct old and new responses always summed to 4). Van Zandt's subjects were able to adjust their decision criterion for each test list to take advantage of the payoff structure.

Proportion of old versus new tests

In the typical recognition test procedure the proportion of old and new items are equal. In situations where the proportions of old and new tests vary, and subjects are informed of the base rates prior to the test phase, they can adaptively set their decision criterion (e.g., Healy & Kubovy, 1978; Ratcliff et al., 1992; Van Zandt, 2000). Studies have also demonstrated changes in response bias when subjects were told that old and new base rates differed, although in actuality they were equal (e.g., Hirshman & Henzler, 1998; Strack & Forster, 1995).

In contrast, uninformed subjects seem quite oblivious of base rates. Recognition performance is largely unaffected when there are no new items in the test phase (Wallace, 1982), and also appears to be unchanged by the absence of targets. Raposo and Dobbins (as described in Dobbins & Han, 2008) presented subjects with a 160-item test list. The first half of the test sequence included both targets and distractors whereas the second half consisted only of new items. Despite this extreme change in base rates, the false alarm rate in each half of the test list was equivalent. In terms of SDT, this result indicates a constant criterion (assuming that the new item distribution remains static).

Healy and Kubovy (1978) compared the effects of payoffs and different base rates (when subjects were correctly informed of the rates in advance) on the magnitude of criterion changes and found that their effects were comparable. It is also relevant to note that changes in criterion placement in response to differences in base rates and payoffs are often smaller than optimal (Healy & Kubovy, 1978; Ratcliff et al., 1992) indicating that while subjects are able to adjust their criterion to be more liberal or more conservative, they do not have fine control over these settings. It may also be the case that criterion changes resulting from correctly or incorrectly informing subjects of base rate differences are in fact instances of the general effect of instructions on response bias. That is, telling subjects that the test includes more new tests than old may be tantamount to instructing subjects to adopt a more conservative criterion.

The above studies demonstrate that subjects are able to adopt a more liberal or more conservative criterion to comply with instructions or payoffs, or when informed of differences in test probabilities. Most typical recognition instructions, however, simply ask subjects to 'respond as accurately as possible' (or 'as accurately and as quickly as possible,' if response time is of concern). How do subjects set their own decision criterion?

Differences in memory strength

When memory strength is varied between lists, the hit rate is higher and the false alarm rate is lower for studied items that were presented more often or more slowly compared to items that were presented less often or at a faster rate (e.g., Cary & Reder, 2003; Hockley & Niewiadomski, 2007; Stretch & Wixted, 1998). Such strength-based mirror effects (so called because changes in false alarm rates mirror changes in hit rates) provide, perhaps, the strongest evidence for strategic criterion placement based on memorability. As Stretch and Wixted (1998) and others have argued, the assumption that subjects adopt a more liberal criterion in the more difficult test condition relative to the less difficult test condition provides a straightforward explanation as to why the false alarm rate is higher in the more difficult condition when the new tests are nominally the same items in each list condition.

Studies of the list strength effect (LSE), the effect of strengthening some list items upon memory for the other items in the list, provide further support for a criterion-based explanation of strength-based mirror effects. In the LSE procedure, recognition performance is compared between items presented in pure strong, pure weak, or mixed lists of strong and weak items. Ratcliff, Clark, and Shiffrin (1990) showed that a LSE is typically observed in tests of recall but not for item recognition. Meta-analyses based on the null LSE in item recognition have indicated differences in response bias between the mixed and pure lists (Chappell & Humphreys, 1994). Hirshman (1995) compared recognition performance for weak items presented in pure and mixed lists and found that criterion placement was more conservative in the mixed-list condition. Hirshman argued that subjects estimate the approximate range of familiarity values of the items in the study list and place their decision criterion within this range. That is, subjects establish their criterion placement based on an assessment of the overall memorability of the study list. Such a strategy for criterion placement would produce differences in false alarm rates between pure strong and pure weak lists, and attenuate differences between mixed and pure lists.

Verde and Rotello (2007) also examined the influence of memory strength on criterion placement. They varied the strength of target items between the first and second blocks of the test list. While the strength of the target items in the first block influenced the subjects' initial criterion placement, changing the strength of the target items midway through the test sequence did not produce a further change or shift in the criterion.

The above studies all support the view that initial criterion placement is based on memory strength, but changes in memory strength during the test list do not lead to any further adjustment in the criterion. Can subjects change their initial criterion placement during the course of the test sequence?

Effects of changes of task difficulty

Brown and Steyvers (2005) reported evidence of systematic changes in the decision criterion over blocks of test trials when they varied the difficulty of lexical decision by changing the nature of the non-words. In a similar vein, Benjamin and Bawa (2004) increased the difficulty of recognition discrimination for exemplars from different taxonomic categories by introducing more categorically related distractors partway through the test list. They found that subjects increased their criterion as a consequence. Interestingly, subjects did not relax their criterion when recognition discrimination was made easier. These studies demonstrate that subjects monitor, at least to some extent, their discrimination performance and that they are able to adjust their decision criterion in response to increases in task difficulty due to changes in the nature of the distractors.

The results of Brown and Steyvers (2005) and Benjamin and Bawa (2004) stand in contrast to those of Verde and Rotello (2007) who did not observe a change in criterion when they manipulated task difficulty by varying the strength of the targets midway through the test list. Verde and Rotello did find, however, that providing immediate feedback on the accuracy of each response did produce a dramatic criterion shift in the second half of the test list. What effect does feedback have on criterion changes?

Role of immediate feedback

Verde and Rotello's (2007) results suggest that feedback is necessary for subjects to adjust to changes in the strength of the targets over the test sequence. Estes and Maddox (1995) and Rhodes and Jacoby (2007) also found that immediate feedback was necessary for subjects to adjust their criterion on the basis of the probability of old versus new tests. Rhodes and Jacoby varied the base rates of target and distractors (33% vs. 67%) presented on the left or right side of the computer screen. They demonstrated that subjects could change their decision criterion in a dynamic item-by-item fashion on the basis of screen location. Both awareness of the difference in base rates with location and response feedback were key factors underlying the shifts in criterion. When feedback was not provided to subjects who were aware of the base rate manipulation, no criterion shifts were observed.

Incorrect feedback can also produce criterion changes. Han and Dobbins (2008) gave subjects biased feedback by informing them either that their misses or that their false alarms were correct. Feedback on the accuracy of the other three classes of responses was accurate. This procedure was designed so that subjects would not notice the incorrect feedback, and was successful in producing criterion shifts in the biased direction.

Why are subjects relatively insensitive to changes in target strength or old versus new base rates in the absence of feedback? What information does feedback provide that subjects do not have? Without feedback, subjects must rely on their own assessment of their performance. Presumably, subjects' confidence in the accuracy of their responses varies from high to low with a substantial proportion of responses in the mid-confidence range as most experiments are designed to avoid floor and ceiling effects. Subjects are sensitive to manipulations that increase the proportion of low confidence responses such as increasing distractor similarity (Bawa & Benjamin, 2004). Subjects appear to be much less sensitive to manipulations that decrease the proportion of low confidence responses, such as decreasing distractor similarity (Bawa & Benjamin, 2004), or manipulations that alter the proportions of high and medium confidence responses by varying the strength of the targets (Verde & Rotello, 2007). The addition of feedback may serve to increase subjects' sensitivity to changes over a greater range of their confidence scale. One way to explore this suggestion is to give subjects the option of receiving informative feedback after each test response and compare the accuracy and confidence of the responses for which they asked for feedback compared to those that they did not. It would also be informative to see if subjects change their requests for feedback when task difficulty is varied.[1]

Rhodes and Jacoby's (2007) results show that subjects can shift their criterion on an item-by-item basis when feedback is provided. Can subjects shift their criterion in such a manner without feedback? This question is the focus of the next section.

Item-by-item criterion shifts

In order to shift criteria from item to item during the course of the test list, subjects must have a discernible basis. Differences in the nature of the recognition task, stimulus differences, and differences in the subjective memorability of the stimuli all provide such a basis.

Recognition task differences

Hockley and Niewiadomski (2007) examined strength-based mirror effects for both item and associative recognition. Associative recognition involves discriminating between pairs of items that were studied together (intact or old pairs) and pairs of items constructed from items that were presented in different pairs at study (rearranged or new pairs). This paradigm provides a relatively pure test of memory for associative information because it controls the familiarity of the individual items. That is, memory for the individual items of the test pairs cannot aid in the discrimination of intact and rearranged test pairs.

Hockley and Niewiamdomski (2007) demonstrated opposing strength-based mirror effects for item and associative recognition by presenting individual words and word pairs at different rates in the same study lists and then testing both item recognition for the individual words and associative recognition for the word pairs in the subsequent test lists. Thus, single items were presented at a slow rate and word pairs at a fast rate in one list condition, and these rates were reversed in the other list condition. Hockley and Niewiandomski described these strength-based mirror effects as 'opposing' because, if strength-based mirror effects are due to criterion changes, then subjects had to move their decision criteria in opposite directions for each item and associative recognition test probe due to the differences in their respective strengths. That is, subjects moved back and forth between a conservative criterion for item recognition decisions and a liberal criterion for associative recognition decisions in the list condition where items were presented at a slow rate and pairs at a fast rate.

Stimulus variables

Salient differences between classes of stimuli also provide a ready basis for criterion changes. One such example is words versus pictures. Hockley and Caron (2007) examined 'opposing' strength-based mirror effects for words and pictures (line drawings). One group of subjects studied a list in which words were presented at a fast rate (1 s/item) and pictures were shown at a slower rate (3 s/item) while another group studied lists where the words were presented at a slow rate (3 s/item) and pictures at a fast rate (500 ms/item). The hit and false alarm rates for each stimulus type and presentation condition are shown in Table 12.1. A picture superiority effect (e.g., Nelson, Reed & Walling, 1976) was found in the form of a mirror effect. More importantly, reliable differences in false alarm rates were observed for both pictures and words between the slow and fast presentation conditions indicating subjects changed their decision criteria for words and pictures in the same test list based on their respective presentation rates during study.[2]

Evidence for criterion changes has also been observed in comparisons between different classes of words. Dougal and Rotello (2007) compared

Table 12.1 Mean hit and false alarm rates and standard errors of the means for pictures and words in the fast-word/slow-picture and slow-word/fast-picture list conditions of Hockley and Caron (2007, Experiment 3)

	Hits				**False alarms**			
	Words		**Pictures**		**Words**		**Pictures**	
	Mean	**SEM**	**Mean**	**SEM**	**Mean**	**SEM**	**Mean**	**SEM**
F-Word S-Pic	.69	.03	.89	.03	.25	.03	.04	.02
S-Word F-Pic	.69	.03	.78	.03	.18	.03	.13	.02

recognition memory for words arousing negative emotions versus neutral words and found that the hit and false alarm rates were greater for the emotional words. Wixted (1992) compared recognition for very rare words (functionally non-words to subjects) with low frequency words and found higher false alarm rates for the rare words. Wixted's subjects also rated the rare words as harder to remember than the more common words. The results of both studies suggest that subjects adopt a more liberal criterion for the class of words they deem less memorable. It is interesting to also note that in both of these studies the criterion changes were inappropriate insofar as there was no difference in discrimination between the different classes of words.

Task difficulty: the revelation effect

The revelation effect refers to the finding that hit and false alarm rates increase when a recognition test probe is preceded by a problem solving task such as anagram solution. Verde and Rotello (2004) showed that when the item revealed in the problem task is the same item as the recognition probe, discrimination is decreased. In contrast, when the problem solving task item is unrelated to the probe item, there is no change in memory sensitivity. An explanation of the revelation effect in the different-task condition is that the problem solving task increases the perceived difficulty of the subsequent recognition task and subjects adopt a more liberal criterion as a consequence (Hicks & Marsh, 1998; Hockley & Niewiadomski, 2001; Verde & Rotello, 2004). This account assumes that subjects shift their criterion from item to item depending on the immediately preceding task.

Support for this interpretation is provided by Major and Hockley (2008) who showed that the same-item version of the revelation effect is observed in forced-choice recognition (see also Hicks & Marsh, 1998), but the different-item version is not. As forced-choice recognition is not a criterion-based decision like yes-no recognition, the failure to find a different-item revelation effect in forced-choice recognition supports the view that the different-item revelation effect seen in yes-no recognition is due to criterion changes.

Differences in memory strength

As discussed previously, there is strong evidence that subjects initially set their decision criterion-based on memory strength. In contrast, a number of studies, such as Verde and Rotello (2007) described earlier, have failed to find any evidence of criterion shifts when memory strength is varied within lists. Stretch and Wixted (1998) provide a dramatic example of such a failure. They differentially strengthened high frequency words relative to low frequency words. To emphasize the difference between the two classes of words, the high and low frequency words were presented in different colours at both study and test. The hit rate was higher for the strengthened high frequency words, but this manipulation did not affect the typical finding of a higher false alarm rate for high frequency words. Thus, even though subjects had an unambiguous basis for distinguishing between strong and weak items, they did not use this information to shift their criterion between strong and weak items. Morrell, Gaitan, and Wixted (2002) also failed to find evidence for criterion changes when they differentially manipulated the strength of items from different taxonomic categories in a within-list design.

Singer and Wixted (2006), however, were able to find evidence that subjects can adjust their decision criterion on an item-by-item basis for tests of varying strength. They had subjects study items from different taxonomic categories. The different categories were studied just prior to a recognition test or after delays of up to two days before. No evidence of criterion differences were observed at the short delays (e.g., 20 or 40 minutes), but the false alarm rate was greater and the estimate of criterion placement was more liberal for categories studied two days before compared to the categories that were studied most recently.

The results of Singer and Wixted (2006) appear to stand in clear contradiction to those of Morrell et al. (2002) who differentially manipulated the strength of words from different taxonomic categories and failed to find any evidence of criterion changes. Singer (2009) resolved this discrepancy by exploring the procedural differences between these studies. Singer found evidence for criterion changes using Morrell et al.'s general procedure combined with the pleasantness rating task that Singer and Wixted's subjects performed at study. Subjects who were given standard learning instructions, as in Morrell et al.'s study, did not show any evidence of criterion changes. It is not clear why subjects would change their criterion following ratings of pleasantness but not after self-directed study. Although overall discrimination was greater in the ratings condition, subjects may have incorrectly judged this condition to be more difficult. Judgements of memorability can influence response bias, as the next section shows.

Subjective memorability

Brown (1976; Brown et al., 1977) proposed that subjects use the characteristics of individual items or events to judge how memorable they are. For example, if I were asked if I ever saw Bruce Whittlesea deliver a paper dressed as Superman, I would use a very conservative criterion in making this judgement based on the belief that such an event would be something I would certainly remember if it had actually happened.[3]

Bruno, Higham, and Perfect (2009) provide evidence that subjects' willingness to shift their criterion is based on their perception of their memory for the study list, a view they termed the *Global Subjective Memorability* hypothesis. These investigators were able to produce a strength-based mirror effect when item strength was varied within lists by presenting items from one semantic category once and items from another category three times. This was only possible, however, by also manipulating the overall subjective memorability of the study list. When the memorability of the study list was substantially reduced by including a large proportion of non-words, or by reducing the presentation duration for each study item to 0.5 s, Bruno et al. proposed that subjects sought alternative sources of decision support and began attending to cues that provided information about item strength. The attention to strength-based cues allowed subjects to shift their criteria on an item-by-item basis between the strong and weak categories.

Whittlesea (Whittlesea & Leboe, 2000; Whittlesea & Williams, 1998, 2001a, 2001b) has presented a detailed and comprehensive view of the role of subjective assessments of performance in his SCAPE framework. In this view it is assumed that people constantly monitor their own performance by comparing their performance against expectations that are based on any of a variety of aspects of the current processing. As a simple example, upon hearing the novel phrase 'He buried the locket under the…,' one is more prepared to encounter 'rock' than 'chair.' Feelings of familiarity do not arise from the fluency associated with an event per se, but rather arise when there is a discrepancy between an expectation and an event.

Whittlesea and Williams (1998) tested the discrepancy-attribution hypothesis by contrasting recognition for three sets of items: common words (e.g., FLOWER), orthographically irregular non-words (e.g., STOFWUS), and orthographically regular non-words (e.g., HENSION). Naming times, a measure of fluency of processing, showed that words were faster than regular non-words, which in turn were faster than irregular non-words. Whittlesea and Williams further found that subjects were much more likely to false alarm to regular non-words (37%) than to words (16%) or irregular non-words (9%). Subjects would expect to process words fluently, as they are well-known items, and they would also expect to process irregular words more slowly, as these items are novel. But subjects would not expect to process regular non-words so easily and would therefore find their fluent processing of such novel stimuli surprising. To resolve this discrepancy, subjects attribute the surprising sense of familiarity of the regular non-words to having seen these items previously. This account assumes that subjects are comparing the actual outcome of processing with what is normatively expected for that item, and that these norms, or criteria, must be computed 'on the fly', or item by item, in light of the particulars of the processing situation.

Summary and conclusions

There is good agreement that subjects can change their decision criterion in response to external influences or demands. Subjects can become more conservative or more liberal when given such instructions (Postma, 1999), or when payoffs favour one type of response over another (Van Zandt, 2000). Subjects will make adaptive criterion changes when informed about the proportion of old versus new test items (Ratcliff et al., 1992), even when this information is incorrect (Hirshman & Henzler, 1998). Subjects will also adjust their criterion when the test becomes more difficult when they are provided with immediate feedback on the accuracy of their performance (Verde & Rotello, 2007).

There is also growing consensus concerning how subjects establish their criterion in the absence of external influences. Initial criterion placement is based on a subjective assessment of the memorability of the study list (e.g., Bruno et al., 2009; Hirshman, 1995) and initial discrimination performance (Verde & Rotello, 2007). Subjects can also change their criterion between different recognition tasks (Hockley & Niewiadomski, 2007), and on the basis of different types of stimuli (Dougal & Rotello, 2007; Hockley & Caron, 2007; Wixted, 1992). Subjects monitor their own performance to a degree and are capable of shifting their criterion on their own accord in response to certain changes in discrimination difficulty (Benjamin & Bawa, 2004), or perceived difficulty of the task (Major & Hockley, 2007; Verde & Rotello, 2004).

The results are more mixed and there is less consensus concerning within-list criterion changes based on memory strength. Several studies have failed to observe such changes (e.g., Morrell et al., 2002; Stretch & Wixted, 1998; Verde & Rotello, 2007) leading Rotello and Macmillan (2008) to conclude that 'in the absence of feedback to the subject, no evidence has been found for criterion shifts that occur in response to a selective strengthening of one class of studied items in memory relative to another' (p. 75). Recent results, however, demonstrate that such criterion shifts can occur, but (and it is an important 'but') the differences in memory strength must be quite dramatic and deemed relevant by the subject (Bruno et al., 2009; Singer, 2009; Singer & Wixted, 2006).

The fact that some manipulations do not produce criterion changes while other manipulations do, and that observed criterion changes appear to be motivated by subjective assessments of memorability and task difficulty, indicates that we will not be able to understand criterion changes using the traditional signal detection theory framework. Rather, we must adopt a more heuristics-based approach, and employ procedures that provide a measure of metacognitive monitoring (e.g., Higham et al., 2009) in future research to understand how subjects control and adjust their recognition decision processes in different situations.

Notes

1. Interpretations of such results, however, would need to also consider the issue of item-selection effects.
2. Morrell, Gaitan, and Wixted (2002) failed to find such a difference in false alarm rates in two experiments where they varied the strength of words and pictures at study. They also did not observe a hit rate advantage for pictures compared to words. Their picture set consisted of drawings of various animals, and the categorical nature of both the drawings (animals) and the words (professions) may have reduced both the actual memorial advantage that is typically seen when pictures are compared with words, and subjects expectations concerning their memorablity.
3. To my knowledge, Bruce has never dressed as Superman to deliver a paper. Nevertheless, he has still fought a never ending battle for truth, justice, and cognitive psychology.

13

Judgements of Learning and Study-Time Allocation: An Illustration from Neuropsychology

Chris J. A. Moulin, Timothy J. Perfect, Shazia Akhtar, Helen L. Williams, and Celine Souchay

Introduction

Metamemory is the experience and knowledge we have about our own cognitive processes and concerns the relationship between monitoring and control of memory processes (e.g., Nelson & Narens, 1990). This is particularly pertinent in memory impairment. For instance, Alzheimer's disease (AD) patients who are more aware of their memory problems benefit more from rehabilitation than those who lack awareness (Clare et al., 2002). The aim of the present chapter is to illustrate the rich theoretical and applied insights that come from studying metacognition in memory impaired groups.

Metamemory has been more extensively researched in AD than in other memory-impaired groups. In a recent overview, Souchay (2007) reported a clear fractionation across metamemory measures in AD; some show impairment (e.g., predictions of memory performance before studying a to-be-remembered list, Moulin et al., 2000c) whereas others show normal performance (e.g., judgements of confidence; Moulin et al., 2003). Even utilizing the same metacognitive measures can reveal dissociations across materials; AD patients are unimpaired for semantic memory feeling of knowing (FOK) tasks, but are impaired on episodic FOK (Souchay et al., 2002).

Both mainstream experimental psychology and neuropsychological research leads to the same conclusion: the accuracy of metacognition is driven by the prevalent cues used by people when making their metacognitive judgements. Such a detailed consideration of the cues used in metacognitive evaluations has a clear Whittlesea flavour. Whittlesea has demystified the processes and feelings people use to gauge their memory performance, demonstrating that a number of heuristics lie behind memory judgements. We echo this approach here and ask whether people with memory impairment are sensitive to different factors on which they may gauge their memory proficiency.

An example of the insights that can come from testing metacognition in neuropsychological groups is provided by our previous work on metacognition in AD. We briefly summarize it here because it introduces two of the measures we discuss in the present chapter – judgements of learning (JOLs) and study-time allocation (STA). We (Moulin et al., 2000a) were interested in the extent to which AD patients are able to monitor and control their learning as a function of study repetition. We examined metacognitive monitoring by asking participants for an explicit evaluation of how well they had learned an item, a JOL. We measured metacognitive control by recording how long people studied an item for before they judged that they had mastered it: study-time allocation (STA: though it has been called recall readiness: e.g., Farrant et al., 1999). The study list contained items studied either once, twice or thrice and AD patients and controls both showed better memory for the repeated items. Control participants gave higher JOLs to repeated items, and studied them for less time, presumably reflecting their mastery. However, for the AD group, JOLs were invariant to item repetition, while STA reduced with repetition as much as controls.

This dissociation echoes the pattern reported by Souchay (2007). It also offers a deeper insight, because it suggests an apparent breakdown in the conventional wisdom about the relationship between control and monitoring (Nelson & Narens, 1990). Thus these data provide an example of the utility of neuropsychological groups for evaluating theoretical models. The control and monitoring framework was developed from data from healthy participants; testing people with memory impairment acts as the exception that proves the rule.[1]

Another aspect of the earlier study is pertinent here. We found (like many others: see Souchay, 2007) that AD patients overestimated their performance – even providing estimates that were higher than controls. Thus, they appeared unaware of their memory deficit. For STA however, people with Alzheimer's elected to study for significantly longer than controls, thus acting as if aware of their poorer memory. Thus there seems to be two independent aspects to each measure pointing to the same conclusion. For JOLs, people with AD predict performance in excess of controls, and fail to vary this prediction as a function of repetition. In contrast, in their STA people with AD appropriately studied for longer than controls, and reduced study time with repetition.

In the current chapter, we present new data looking at the same two measures in a different neuropsychological group – those with Mild Cognitive Impairment (MCI) – to see whether they show the same dissociation. Before we get to these data, we will first provide a quick review of judgements of learning, and STA, and justify a sensitivity-based approach to interpreting these measures.

Judgements of how well material has been learned can be divided into two broad classes: global measures are judgements of how much of a list will be recalled after the entire list has been studied, whilst item-based measures are judgements of the likelihood of recall for each item at the time of study, or shortly afterwards. Our previous work with AD patients has utilized both approaches (e.g., Moulin et al., 2000b, 2000c), but here we concentrate on the most frequently used methodology: item-based JOLs.

Experimental work on JOLs has demonstrated that they are both *sensitive* to known mnemonic factors[2] and *predictive* of subsequent performance. With respect to sensitivity, JOLs have been shown to be higher for semantically related study items than unrelated items (Matvey et al., 2006), higher for distinctive faces than non-distinctive ones (Sommer et al., 1995), higher for normatively easier items than harder ones (Moulin et al., 2000b), and higher for fluently generated items (Matvey et al., 2001). With respect to accuracy, JOLs have been shown to be more accurate than predictions based on normative ratings (Lovelace, 1984) more accurate following intentional than incidental learning (Mazzoni & Nelson, 1995), and more accurate if a brief delay intervenes between study and judgement (e.g., Nelson & Dunlosky, 1991).

JOL sensitivity and accuracy have also been explored in memory-impaired populations or conditions known to produce memory impairment. Normal adult levels of JOL accuracy have been reported in children aged 8 and older (Roebers et al., 2007), in healthy older adults (Connor et al., 1997), and in people with schizophrenia (Bacon et al., 2007) despite all those groups having lower memory performance than healthy controls. Memory is impaired by both nitrous oxide inhalation and midalozam, but the former doesn't impact on JOL accuracy (Dunlosky et al., 1998) whilst the latter impairs JOLs (Merritt et al., 2005).

Our second metamemory measure is study-time allocation. It has well established that participants allocate more time to normatively more difficult materials (for a comprehensive review, see Son & Metcalfe, 2000), suggesting that allocation of STA is sensitive to item differences. STA is certainly strategic and guided by the amount of time for study and the nature of the materials; Son and Metcalfe (2000) demonstrated that STA was modulated by time pressures. With extreme time pressures, people 'gave up' on the difficult-to-remember items. This suggests that a complex set of heuristics is at play in STA; it is not merely that difficult items are studied for longer. In terms of group differences in STA, there is very little on self-allocated study time in memory impaired groups, although our own series of studies show that people with Alzheimer's disease elect to spend at least twice as long as older adult controls. Others have shown that group differences are not apparent – for example, children with autism and learning difficulties do not elect to study items for longer than controls (Farrant et al., 1999).

Rather more research has considered the relationship between study time and subsequent performance (an accuracy analysis). A robust finding is that, even when instructed to master every item, and given an unlimited time, participants do not recall every item. Nelson and Leonesio (1988) suggest that large increases in self-paced study time yield very small increases in recall, naming this the *Labor in Vain Effect*. In memory-impaired groups, it is usual to attempt to equate group differences in recall by allocating extra study. In practice, this does little for performance. AD memory performance is still worse than controls' after twenty times the presentation duration (20s vs. 1s, Heun et al., 1998).

Because there is relatively little work looking at STA in memory-impaired groups, and little on JOLs other than in AD, we present some exploratory data here. Our experiment has three new design features. The first is to escape the circularity inherent in study-time allocation and recall for the same retrieval test. To achieve this, we look at study-time allocation on a second trial, following a retrieval test on which recall has (or has not) occurred. Thus, rather than seeing whether study-time leads to recall, we ask whether knowledge of recall status (recalled/not recalled) for an item leads to changes in subsequent study time.

Our second innovation is to adopt both individual-based and item-based analyses where there are contingencies in the data. Previously, for instance, we have compared mean latency for recalled and non-recalled items, drawing conclusions based on individual-level analyses (where means are calculated for each person for recalled and non-recalled items). However, this approach neglects item differences. Recalled and non-recalled items may vary in many ways other than the JOLs they elicit. An alternative is to calculate, on an item-by-item basis, the mean study time for each item, depending upon whether it had been recalled or not. This analysis removes item differences from the recall status analysis, but introduces a different confound: individual differences. People who recall (or not) the items may vary in ways other than their recall ability – in particular they may study for longer (or shorter). Whilst we believe that neither approach is ideal, we hope that by running both analyses, we can either confirm the robustness of each effect we observe, or we can identify where the conclusions may be open to challenge.

Finally, our previous studies of sensitivity have used of mixed lists of easy and difficult items (Moulin et al., 2000b). We believe that this element of the design may be important, since it may draw attention to item differences and so further complicate the relationship between item difficulty, subjective evaluation and subsequent performance. Here we adopt blocked lists, to see if participants are sensitive in their monitoring and control of a set of similar items. We have demonstrated sensitivity to difficulty using blocked designs but with global judgements (Moulin et al., 2000c) – this is the first attempt to use similar blocked materials but use item-by-item judgements.

We have argued elsewhere that interpretation of metacognitive functioning in memory-impaired groups is best achieved through sensitivity analyses, rather than through accuracy based ones (e.g., see Moulin, 2002). Because we utilize this approach in the present chapter, we wish to recapitulate the argument here.

The starting point of our argument is a no-magic assumption: participants make their predictions on the basis of the current contents of consciousness, not on the basis of an ability to predict the future. The accuracy of any judgement made will be a function of the ability to monitor cues relevant to the items undergoing study, and the relationship between the present state and future test conditions. If the test conditions are unknown, then current predictions may be inaccurate. This uncertainty makes interpretation of absolute predictions of recall problematic. If the participant does not know every last detail of the experiment then there may be valid reasons for predictions to deviate from later performance that would not be indicative of metacognitive failure. [3] Indeed, there is evidence that participants share our uncertainty about predictive accuracy – in the absence of any information about the future test, participants tend to anchor their list-based estimates of recall at about half the list length (Connor et al., 1997).

There are number of practical and theoretical consequences to the idea that metacognitive accuracy is driven by anchoring (predicting an unknown future) and adjustment (monitoring online processing). If one is interested in absolute accuracy of judgement, then the only appropriate methodology is to use a repeated study-judgement-test cycle in which participants experience the test conditions. In this way, participants' knowledge of their test performance can inform subsequent judgements at study. To the extent that test performance on test n is predictive of performance on test n + 1, then participants can make their anchoring meaningful. We used this methodology to demonstrate that the predictions of both AD patients and older adult controls became more accurate with test repetition (Moulin et al., 2000c).

In many cases in the literature, the anchoring issue acts merely as noise, with variations in metacognitive accuracy arising as a function of the monitoring of the contents of consciousness. An example is the delayed-JOL effect (Nelson & Dunlosky 1991). Whether participants make the judgement immediately at study, or after a delay, final test performance remains fixed. The greater accuracy of JOLs after a delay cannot therefore be a function of final test performance. However, when final test performance varies, the issue becomes more problematic.

Evaluating the metacognitive accuracy of memory-impaired groups is less straightforward for two reasons. First, the setting of an inappropriate anchoring point by memory-impaired individuals may swamp any item-by-item adjustments made. Second, recall is likely to be at floor, and this will constrain variance in performance. Thus variations in ratings may not be mirrored by variations in performance because of restriction of range effects. One solution is to move from an accuracy focus, to a sensitivity focus. These difficulties arise when comparing predictions with performance when performance varies (and approaches floor): they do not arise when comparing the sensitivity of judgements independent of ultimate accuracy. For example, a memory-impaired individual may accurately judge the relative difficulty of lists that are normatively easy, moderate, and hard, whether they ultimately recall 6, 4, and 2 items, or 0, 0, and 0 items. That is they can be *sensitive* to the mnemonic cues available at study, whether or not their ratings are ultimately predictive. This is fundamentally different from someone who cannot monitor learning (and later recalls at levels above floor, but at levels uncorrelated with predictions). This insight can be applied at either the list or item-level, with variations in memorability resulting from intrinsic properties of the stimuli (e.g., word frequency) or mnemonic cues (e.g., the nature of the encoding). It can be applied both to measures of monitoring and control, as we show here.

The majority of metacognitive research has been conducted with patients with Alzheimer's disease. Here we extend our research to people with mild cognitive impairment (MCI). MCI is a diagnostic term which applies to non-demented individuals who have objectively assessed memory impairment below expectations for age and education level (Petersen et al., 1999). Primarily, MCI is characterized by poor performance on tests of episodic memory, without any deficit in other cognitive domains and no impact on activities of daily living. It is often conceptualized as being incipient dementia; many people with MCI will later convert to dementia, particularly Alzheimer's disease (Petersen et al., 2001).

An interesting contrast between MCI and AD is that the diagnostic criteria for MCI emphasize subjective awareness of the memory whilst AD patients tend to lack such awareness. Recall that AD patients are inaccurate both in their ability to predict future recall at a gross level, and are insensitive to the effects of repetition in their explicit judgements. If this inability to make appropriate JOLs is associated with lack of awareness of deficit, one might expect the MCI group to show more appropriate JOLs.

This hypothesis is rather speculative at the present time because beyond awareness of memory impairment at the gross level, metacognition in MCI is little-researched. One recent study by Perrotin, Belleville, and Isingrini (2007) investigated metamemory monitoring using an episodic feeling-of-knowing (FOK) procedure, finding that the MCI group were significantly less able than a control group to accurately predict the recognition of unrecallable items. In addition, the MCI group tended to overestimate their performance on this task. In contrast, on a standardized questionnaire measure of memory awareness, the MCI group demonstrated awareness of their memory failures. Whilst this study is suggestive of metacognitive failures in MCI, the previously reported neuropsychological evidence showing dissociations across tasks and materials (Souchay, 2007), means we should hesitate before drawing firm conclusions.

In the study reported here, we adopted a sensitivity approach to explore the ability of participants with MCI to monitor and control their learning, in a multi-trial setting. Participants first studied a list of 10 items at a fixed presentation rate, making JOLs and delayed JOLs for each item. They then attempted a recall test, before engaging in a second time-unlimited study phase and then a final recall test. This procedure was repeated twice, once for a list of normatively easy items, and once for a list of normatively hard items.

Before we describe our predictions for this study, we need to deal with one methodological issue. Previous research on the delayed-JOL effect has explored the effect of judging the likely recall of an item either immediately after study, or after a filled delay, usually of a few study items (e.g., Dunlosky & Nelson 1992). However, previous pilot work with memory-impaired patients found this to be problematic, because many do not recall the prior item after a filled delay, and find the task of judging recallability meaningless and distressing (even though logically an appropriate response would be to predict zero recall). Consequently, we adopted an adjusted procedure in which participants make both immediate JOLs, and delayed JOLs for the same item sequentially, separated by a brief filled interval. To our knowledge this is the first use of such a procedure.

Participants can demonstrate sensitive metacognitive monitoring and control in a number of ways during this procedure. At a gross level, one would expect the MCI group to report lower JOLs, and show longer study times than controls. We would also expect that the average JOL to be higher and study time lower for the easier list than the hard list. Given that delayed JOLs are thought to provide insight into the long-term recallability of items (Nelson & Narens, 1990), JOLs should shift from immediate to delayed in the direction of subsequent recall. Finally, given that the first recall test provides information about the recallability of individual items on the study list, one would expect study time on the final study phase to reflect previous recall status.

There were 16 patients with MCI, and 16 healthy older adults, the characteristics of whom are shown in Table 13.1. The MCI patients were tested as they attended a memory clinic; the older adults (OACs) were volunteers who were community dwelling, and were tested in their own home. The MCI patients were diagnosed by an independent clinician, using Petersen et al.'s (1999) diagnostic criteria. There were no significant differences between groups in the mean ages, $F < 1$, MMSE scores (Mini-Mental State Examination; Folstein et al., 1975), $F(1,31) = 3.62$, $p > .05$ or NART predicted IQ (National Adult Reading Test; Nelson, 1982), $F(1,31) = 2.24$, $p > .05$.

Two sets of ten words were taken from Rubin and Friendly's (1986) Recall Norms and presented item-by-item in a blocked fashion. One set was easy (e.g., grandmother, elephant; mean recall probability = .74) and the other set was difficult (causality, figment; mean recall probability = .24). These were

Table 13.1 Mean and standard deviation for age, MMSE and NART for MCI and OAC groups

	Age	MMSE	NART
MCI	77.43 (6.28)	27.87 (1.2)	37.75 (10.83)
OAC	77.37 (8.09)	28.75 (1.39)	42.31 (5.61)

Notes: MCI = Mild Cognitive Impairment, OAC = Older Adult Controls, MMSE = Mini-Mental State Examination, NART = National Adult Reading Test

the same materials as used in Moulin et al. (2000b). For each set, there were two study-test trials. In the first trial, each word appeared in a random order for 2 s, and an immediate JOL was prompted immediately with the cue: 'How confident are you that in about 10 minutes you will be able to recall the word you have just seen (0% = definitely not, 100% definitely will).' The participant then completed three simple mental arithmetic sums and made a delayed JOL in the same manner. After all words had been studied and JOLs made, participants were instructed to write down all the words they could remember. In the second trial, participants were given unlimited study time (study-time allocation) for the same words presented in a different random order; there were no intervening sums, and free recall was again measured immediately in the same manner.

Before we report the analyses of JOLs and STA, we begin by reporting recall performance, to provide a background context for understanding the metacognitive performance of the MCI group.

Recall

A 2 x 2 x 2 (Group: OAC vs. MCI x List status: Easy vs. Hard x Study Trial: 1st vs. 2nd study-test cycle) ANOVA was performed on the number of words recalled (see Table 13.2). There was a main effect of Group, $F(1,30) = 28.4$, $p < .001$, with poorer recall demonstrated by the MCI group. There was also an overall effect of list status, $F(1,30) = 65.8$, $p < .001$, with higher recall for the easier lists. Recall also improved over the two study-test cycles, $F(1,30) = 51.4$, $p < .001$. However, none of these factors interacted (all *F*s < 2.7, all *p*s > .11). Consequently, it is clear that our two experimental manipulations had their desired effects, and that these were equivalent for both groups.

JOL sensitivity

Given the pattern in the recall data, we anticipated that the MCI and OAC groups' JOLs should be lower for the difficult list and lower in the MCI group. We were also interested in the extent to which JOLs varied from

Table 13.2 Mean correct recall (out of 10) for MCI and OAC groups for easy and hard lists across study trials 1 and 2

	List type							
	Easy				**Difficult**			
	Trial 1		**Trial 2**		**Trial 1**		**Trial 2**	
Group	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
MCI	4.0	2.3	5.7	1.9	2.1	1.3	3.8	2.5
OAC	7.2	1.6	8.6	1.5	4.1	1.7	6.0	1.7

Notes: MCI = Mild Cognitive Impairment, OAC = Older Adult Controls.

Table 13.3 Mean JOL (out of 100) for MCI and OAC groups for easy and hard lists for immediate and delayed judgements

	List type							
	Easy				**Hard**			
	Trial 1		**Trial 2**		**Trial 1**		**Trial 2**	
Group	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
MCI	53.5	15.5	36.5	11.7	48.2	16.5	31.3	21.0
OAC	58.7	11.9	57.0	11.5	51.3	15.0	43.5	11.0

Notes: MCI = Mild Cognitive Impairment, OAC = Older Adult Controls.

immediate to delay. The extent to which the MCI and OAC group show differential sensitivity to list difficulty, study trial and JOL delay was expected to be revealed by interactions with Group in a 2 x 2 x 2 (Group x List status x JOL delay: immediate vs. delayed) ANOVA. The data are shown in Table 13.3.

There were reliable effects of list difficulty, $F(1,30) = 15.9$, $p < .001$, and Group, $F(1,30) = 5.3$, $p < .03$. There was also a main effect of delay, such that delayed JOLs were lower than immediate JOLs, $F(1,30) = 62.9$, $p < .001$. However, the effects of delay were not equivalent for both groups, $F(1,30) = 19.9$, $p < .001$. This interaction was decomposed into simple main effects. For the immediate JOLs there was no difference between the two groups, $F < 1$. However, for the delayed judgement, JOLs were higher for the OAC group than the MCI group, $F(1,30) = 13.5$, $p < .001$. No other interaction was significant (all Fs < 2.1, all ps > .15).

Close inspection of Table 13.3 reveals a number of elements to these data. If one focuses on the immediate JOLs, then the MCI group's predictions resemble those of the OACs. Whilst they are sensitive to overall difficulty of the lists, there is no overall group effect on mean JOL, suggesting perhaps that MCI patients do not take account of their memory impairment when predicting future memory performance. However, for delayed JOLs, the pattern changes and the memory predictions of the MCI group are markedly lower, dropping from a mean of 50.8 to 33.9, whilst OAC's JOLs drop from 55.0 to only 50.3. Contrasting these data with the recall levels for test Trial 1 (From Table 13.2) is informative: the MCI group average 31% recall whilst the OAC group average 57%.

Thus, the delayed JOLs of both groups are appropriate at the group level, and so the MCI group do appear metacognitively sensitive, both to the relative differences between items and to their lower likelihood of future recall. But this pattern only emerges for the delayed JOLs.

We also looked at the JOL data broken down by whether participants later recalled the items. These data are displayed in Table 13.4. We did so with a sense of caution, however, because we retain our concerns that these data can be difficult to interpret in an absolute sense, given variations in level of recall between groups. Further, we are aware that analyses broken down by recall status introduce potential contingency-based confounds into the data: recalled items may differ from non-recalled items in ways other than the JOLs they elicit. Consequently, we analysed these data both by individual, and by item. However, rather than report both analyses in full, we will first report the conventional (individual-based) analysis, and then report any differential outcome produced by the item-based analysis.

With respect to our primary concern about differential levels of performance, we note that comparison of immediate versus delayed JOLs is not confounded. It is a matter of particular interest whether the reduction in mean delayed JOLs in the MCI group represents a downward shift generally (i.e., they merely reduce all ratings equally after a delay) or whether they are meaningfully related to ultimate performance (i.e., they discriminate between items that will later be remembered or forgotten). The mean level of JOLs were calculated for recalled and non-recalled items (on the first retrieval list), and subject to a 2 x 2 x 2 x 2 (Item status: recalled vs. non-recalled x Group x list difficulty x JOL delay) ANOVA. We will not report the analyses that duplicate those above, but will instead focus on the main effect of recall status, and its interactions. Overall, participants gave higher ratings to subsequently recalled items (mean 54.4% vs. 46.4%), although the effect was small, relative to the actual recall status of the two sets of items (i.e., 100% vs. 0% recall), $F(1,26) = 8.0$, $p < .01$. Thus, ability to monitor future recall status is limited. However, this interacted with delay, $F(1,26) = 4.42$, $p < .05$. There was effectively no discrimination in immediate JOLs (recalled: 54%; not recalled: 53.2%), but reliable discrimination in delayed

Table 13.4 Mean and standard deviation of judgements of learning for MCI and OAC groups across easy and difficult tasks for recalled and non-recalled words

	Immediate JOL		**Delayed JOL**	
	Recalled	**Not recalled**	**Recalled**	**Not recalled**
MCI	59.87(17.27)	49.40 (16.21)	53.52 (14.22)	30.07 (15.01)
OAC	59.61 (11.93)	60.71 (14.82)	59.65 (12.07)	56.81 (14.76)

Notes: MCI = Mild Cognitive Impairment, OAC = Older Adult Controls.

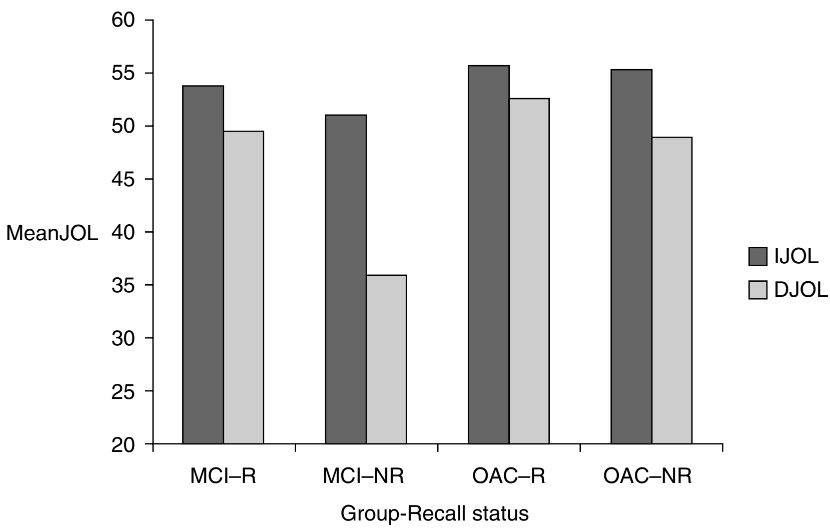

Figure 13.1 Mean immediate and delayed JOLs for recalled and non-recalled words by MCI and OAC participants. MCI = Mild cognitive impairment group, OAC = Older adult control group. R = Recalled items, NR = Non-recalled items. IJOL = Immediate judgements of learning. DJOL = delayed judgements of learning

JOLs (recalled: 51.1%, not recalled 42.4%). This interaction in turn was complicated by a 3-way interaction with group, $F(1,26) = 8.1$, $p < .01$, which is illustrated in Figure 13.1. It is clear that there is no evidence to suggest metacognitive impairment in the MCI group. They are better able to monitor the relative status of recalled and non-recalled items with their delayed JOLs than the OAC group. We return to this point below.

Reanalysis of these data by items did not produce any difference to this pattern of results: the same main effects and interactions were significant, although the magnitude of the effects varied slightly. All other interactions remained non-significant by item-analysis.

Study time on trial 2

We ran two sensitivity analyses of study time for the second study trial. Our first analysis looked at total study time for easy versus difficult lists for the two groups. Overall, the MCI group studied the lists for longer than the OAC group (63.3s vs. 32.6s respectively), $F(1,30) = 13.1$, $p < .001$. Both groups also studied the harder items for longer, $F(1,30) = 13.3$, $p < .001$, but there was no interaction between the two factors, $F(1,30) = 1.8$, $p < .19$. Thus, the MCI group appear to allocate their study time similarly to the controls, devoting additional time to the longer items.

Participants' allocation of study time occurred after a retrieval attempt. Thus, it was possible to determine whether their study-time allocation was sensitive to whether or not the item was recalled previously. To analyse this, we ran a 2 x 2 x 2 (Group x Recall status x List difficulty) ANOVA on study time per item. As before, we found main effects of group and list difficulty, with longer study times for items from hard lists, and longer study times by the MCI group than the OAC group. However, whether or not the items had previously been recalled made no difference to study time, and failed to interact with any of the factors (all *F*s < 1.02, all *p*s > .33).

We also ran an item-based analysis of the same factors. This analysis replicated the main effects of group and list difficulty. However, in contrast to the conventional analysis, this analysis did produce a reliable effect of recall status. Overall, study times were shorter for items that were previously recalled (3.7 s) than items previously not-recalled (5.5 s), $F(1,16) = 5.52$, $p < .05$. No other interaction was significant, although the interaction between group and list difficulty approached significance, $F(1,16) = 4.13$, $p < .06$. Follow up analysis showed that both groups showed significant effects of List difficulty in their study times, but that the MCI group showed a larger increase in study time between easy and hard lists (difference = 2.45 s for MCI group, and 1.17 s for OAC group, both *F*s > 18.7, both *p*s < .001).

We will return to the issue of the different patterns between individual-based and item-based analyses later. For now, we want to focus on the conclusions about metacognitive ability in MCI. Clearly, there is no deficit in study-time allocation in the MCI group: they study harder items for longer than easier items, and non-recalled items longer than recalled ones (by item analysis) either to an equal or greater extent than the OAC group.

What can these data tell us? Our starting point was that AD patients show a contrast between their ability to control their learning through study-time allocation but an apparent inability to monitor their learning through JOLs (at least with regards to repetition). Our previous research had shown AD patients failed to adjust both their overall JOLs (giving predictions that were as high as controls) and their JOLs. We were interested to see whether insensitivity of JOLs would be seen in a group of adults with MCI rather than AD.

In fact, there is evidence to support both normal, and abnormal metacognitive monitoring by the MCI group. On balance (for reasons we will elucidate) we favour the conclusion that there is no metacognitive sensitivity impairment in MCI, but we will start by considering the evidence to the contrary. If one considers the immediate JOLs made by the MCI group, then two patterns are evident. First, although they do give significantly lower ratings to the harder items, the effect is small relative to the large effect of list difficulty on test performance. Their JOLs for hard items are around 5% lower on average than for easy items, but their recall for hard items is around half that for easy items. Thus, their immediate JOLs appear relatively insensitive to item difficulty. The second apparent metacognitive deficit concerns the absolute level of immediate JOLs. Those with MCI make average JOLs that do not differ from normal controls, even though their ultimate recall is almost half that of controls. That is, despite their diagnosis requiring self-awareness of memory impairment, they make immediate JOLs that take no account of memory impairment.

Both these arguments are *straw men* which we can readily knock down. The fact that immediate JOLs are relatively insensitive to item difficulty is not evidence for metacognitive impairment in MCI, since the controls show the same. This confirms that immediate JOLs are less predictive of recall: this is well established, by the existence of the delayed-JOL effect (e.g., Dunlosky & Nelson, 1992). In fact, we believe this is the first study using delayed JOLs in a memory-impaired group. This is due to the adoption of a slightly different single-item JOL procedure with the adoption of a delay filled with sums rather than intervening cue-target pairs, which proved too confusing for pilot work in Alzheimer's disease.[4]

The apparent metacognitive impairment of MCI groups is that their immediate JOLs match those of a group with better memory performance: but given that neither group is able to predict their later recall, it is hard to know what to make of this pattern. The conclusion changes when one looks at the delayed JOLs made by the two groups. Here, there is evidence that the absolute level of JOLs made by MCI patients is lower than older adult controls. Moreover, this shift occurs because the MCI group, but not the OAC group, appear to be sensitive to forgetting over the delay period between immediate and delayed JOLs. Figure 13.1 demonstrates that the OAC group shift their JOLs little over the delay, either for recalled or non-recalled items, and consequently their delayed JOLs remain unpredictive of future recall. In contrast, the MCI group do not shift their JOLs for items they will later recall, but they do shift them dramatically for those they will later forget. Thus, the clear conclusion is that the MCI group show superior, not inferior, metacognitive monitoring.

The second measure we were interested was STA. Once again, these data showed no evidence of a metacognitive deficit in MCI. Like normal controls, they spent longer studying harder items than easier items. They also studied items for longer than the control group, in line with their impairment. Finally, there was no evidence that they allocated study time inappropriately across recalled and non-recalled items. (Even though the evidence on this point depended upon the nature of the analysis conducted.) Perhaps the most striking finding is the additional study time observed in MCI: Overall, MCI patients study twice as long controls on the second trial, but recalled around half as much. One could describe it as a clinical manifestation (at a between group level) of the labour-in-vain effect. However, this doesn't seem have arisen because of a failure to regulate study across items: there were no group interactions with study time across recall status whether measured by individuals or by item.

What do these study time data tell us about potential remediation of memory impairment? In healthy groups STA is predictive of subsequent performance, but large changes in study time do very little for performance (Nelson & Leonesio, 1988). It would be a simple story if the MCI memory impairment was caused by a failure to dedicate appropriate STA whilst materials are being learned. It would be easy for clinical neuropsychologists if memory impairments could merely be resolved by training people to spend longer at study. Unfortunately, neither of these is borne out in our data. In short, STA is driven by factors concerning the to-be-remembered materials and the cognitive status of the individual at work, but there the story ends.

So where does this leave us? First, we think the nature of STA and memory impairment is in some ways better understood through an analysis by items as well as by participants. Item analyses of metacognition may offer new insights into judgements being driven by the nature of the materials as well as the processes undertaken by the participants. In the current study, we illustrate that a clearer pattern emerges through an item analysis, whereby people with MCI were shown to successfully allocate extra study for items not previously recalled. The fact that this finding was emphasized through an item analysis suggests to us that some items afford different levels of study whether or not they are ultimately recalled, and thus such an analysis makes allowances for item-characteristics. As Whittlesea would suggest, one of the cues available to people as they study a word is the ease at which they read it and as such, different items in a list may well be processed more or less fluently. On top of this, idiosyncratic adjustments according to genuine monitoring of memory function might then have some bearing on memory function. (In fact, as we have argued, these adjustments have little value for subsequent recall.)

Second, we are left with an interesting finding that, if anything, people with MCI show superior monitoring, particularly in their response to a delay period. We suggest that the filled delay may produce a faster forgetting rate in the MCI group which gives them a chance to gauge their memory in a way that isn't captured in the OAC group. Perhaps different delays (longer, or filled with different activity) would produce greater sensitivity in OACs. In any case, it seems to suggest that having a poorer memory may – in some cases – actually lead to a greater awareness of it. If you forget particularly quickly, you can more quickly assess what your recall might be over a longer interval.

Notes

1. Whittlesea footnotes tend to be amusing, engaging, and insightful (e.g., see the note about the redintegration of humour in Whittlesea & Williams, 1998). We don't claim to be able to match him but we offer the following: the etymology of 'prove' in 'proving the rule' is 'to test or try,' thus, neuropsychology is the exception to *test* the rule. Note that if you don't understand the etymology, the proverb makes no sense. For the record, our metacognitive feelings of Amusement (FOA), Engagement (FOE) and Insight (FOI) are *low, low* and *medium* on the Whittlesea scales respectively.
2. They have also been shown to be sensitive to illusory factors which *don't* affect memory performance, such as font size (e.g., Rhodes & Castell, 2008).
3. Imagine a participant presented with the first item of a study list and asked to judge the likelihood that they will recall it subsequently. Experimental research would suggest that they would need to appropriately anticipate the effects of the list length, the presentation rate, the categorical structure of the list, item similarity across the list (semantic, orthographic), the number of item repetitions, the length of any delay, the nature of any filler task, and the context of the retrieval test. Given that any or all of these factors may cause the likelihood of recall of an item to rise or fall, in what sense can the accuracy of any prediction be said to be a measure of ability?
4. This departure from the normal delayed-JOL procure is certainly more apt for memory-impaired groups, but we also think it may prove theoretically interesting, but lack of space prevents discussion of these issues. In short, the use of a filler task to clear short-term memory means that the delayed JOLs are based on registration of the item in long-term memory, but the JOLs are not confounded by possible differential practice effects or the effects of interleaving items (see Kimball & Metcalfe, 2003, for an alternative account of the delayed-JOL effect).

14
Agenda-Based Regulation of Study-Time Allocation

John Dunlosky, Robert Ariel, and Keith W. Thiede

Introduction

A great deal of learning occurs in contexts where people can regulate their study and hence can control their success. College students may decide to focus on mastering some class materials and to largely ignore others; on the job, doctors may decide how much time to devote to learning about new advances in their field; and for a hobby like bird watching, an enthusiast can choose how to allocate their time to learning birds' names and their songs. Thus, people's success at learning will be driven in part by how they allocate their study time, which brings us to the main question of this chapter: What drives people's allocation of study time as they are attempting to learn new materials? This question has received much attention since Rose Zacks' seminal research in 1961, so to put our current answer in context, we first briefly describe some of the earliest empirical and theoretical work on study-time allocation.

Most research has examined the degree to which item difficulty is related to study-time allocation. In a typical experiment, participants first study to-be-learned materials (e.g., paired associates) that are presented individually at a fixed rate. They then judge how easy each item would be to learn, which is a subjective judgement of item difficulty. Next, participants select items they want to restudy and/or are given unlimited time to study each one. Eventually, they are tested over the materials. Of central interest is the relationship between participants' judgements of item difficulty and subsequent allocation decisions. In a review of research investigating this relationship, Son and Metcalfe (2000) identified 46 experimental conditions, and 35 of these demonstrated participants' preference to allocate more study time to the more difficult items (versus the less difficult ones) on a list. Given such consistent evidence, it perhaps is not surprising that current theories have been built upon the monitoring-affects-control hypothesis (Nelson & Leonesio, 1988; cf. Miller et al., 1960, TOTE unit). In these theories, people presumably monitor item difficulty to decide how to allocate study time. For instance, according to the discrepancy reduction model (Dunlosky & Hertzog, 1998), people set a goal for learning and then continue studying each item until they have met this goal. Given that the more difficult items (versus less difficult ones) will typically require more time to reach the learning goal, the prediction is consistent with the prevailing evidence; namely, that people will spend more time studying difficult than less difficult items.

Although the discrepancy reduction model is plausible and likely does account for some study-time behaviours, research by Thiede and Dunlosky (1999) revealed its inadequacy at providing a complete account of study-time allocation (see also, Son & Metcalfe, 2000). They used a version of the procedure described above: After a preliminary study and judgement phase, the 30 items were presented in an array for item selection. At this point, some participants were instructed that they had to learn at least 24 items, whereas other participants were instructed that they only needed to learn 6 items. Consistent with the investigations reviewed by Son and Metcalfe (which encouraged participants to learn all the items), participants with a high-performance goal chose to allocate most study to the majority of the most difficult items. That is, their judgements of learning (with higher values indicating an item is easier to learn) were negatively correlated with whether items were selected for study (*M* correlation ≈ -.37, Experiment 1) and they chose about 22 of the 30 items for restudy. By contrast, participants given a low-performance goal focused on the easier items (*M* correlation ≈ +.38) and chose only about 12 items for restudy. In both cases, people used their monitoring of item difficulty to allocate study time, but most important, the different goal instructions – learn most items or only a few – qualitatively changed *how* people used item difficulty.

This change in study-time allocation has been dubbed the *shift-to-easier-materials* (STEM) effect, because task constraints such as having a low-performance goal (or little time to study) encourage people to shift away from studying difficult items and towards studying easier ones (Dunlosky & Thiede, 2004). Such STEM effects cannot readily be explained by the discrepancy reduction model (Son & Metcalfe, 2000; Thiede & Dunlosky, 1999), and most relevant here, they have inspired empirical investigations and more general models of study-time allocation. In the present chapter, we elaborate on a new model of study-time allocation – the agenda-based regulation (ABR) model – that offers a straightforward explanation for why learners demonstrate STEM effects, and as important, why they fail to demonstrate them under conditions in which they should. Next, we describe the ABR model in detail and then consider how it explains people's study-time allocation under a variety of conditions.

Agenda-based regulation model

The model of agenda-based regulation (ABR) emphasizes the role of agendas in the regulation of study. According to this model, 'learners develop an agenda on how to allocate time to various study items and use this agenda when selecting items for study. Like many other theories of regulation (e.g., Benjamin, 2007; Carver & Scheier, 2000; Pintrich, 2000), the ABR model assumes that study regulation is goal oriented' (p. 433; Ariel et al., 2009). When learners develop an agenda, we assume their goal is to maximize the likelihood of obtaining their goals in the most efficient manner (Thiede & Dunlosky, 1999). For instance, when students prepare for an upcoming test, they may attempt to maximize exam performance by directing their study to those materials that they believe are most likely to be tested. The idea here is simply that based on various task constraints (e.g., potential rewards, likelihood items will be tested), learners construct an agenda – or plan – that describes the criteria used to select items for study, and then they chose for restudy items that fit the decision criteria. In the present example, the students' agenda may include the criteria of 'select items for restudy that are likely to be tested,' but learners can adopt a wide variety of criteria depending on their own goals and task constraints.

The ABR framework: top-down and bottom-up processes drive study-time allocation

We purposely called this model *agenda-based* regulation to emphasize the critical role of agenda construction and execution in study-time allocation, because learners' goal-oriented behaviour has received minimal attention in previous theories (but see Dunlosky & Thiede, 2004; Le Ny et al., 1972; Thiede & Dunlosky, 1999; and more recently, Koriat & Nussinson, 2009). Nevertheless, to account for extant data on study-time allocation, the ABR model has been expanded to comprise both top-down processes that are involved in agenda construction and execution as well as bottom-up processes that may also drive study-time allocation. Within this ABR framework, the joint operation of top-down and bottom-up processes within a memory limited system drive study behaviour and influence the degree to which study-time allocation is optimal.

An illustration of the ABR framework is presented in Figure 14.1, which illustrates some aspects of agenda construction. This framework is directly based on Cowan's (1988) information-processing system that describes how selective attention is constrained by the structure of the memory system and the limited processing capacity of a central executive. To help you understand the ABR framework, we first briefly summarize some key aspects of Cowan's model (for in depth discussion, see Cowan, 1995). His model describes the flow of information in the memory system beginning with input of stimuli from the environment into a brief sensory store. From the sensory store, representations of these stimuli in long-term storage can be activated in a short-term store (STS). The central executive can also select a subset of activated memory that will remain in the focus of attention,

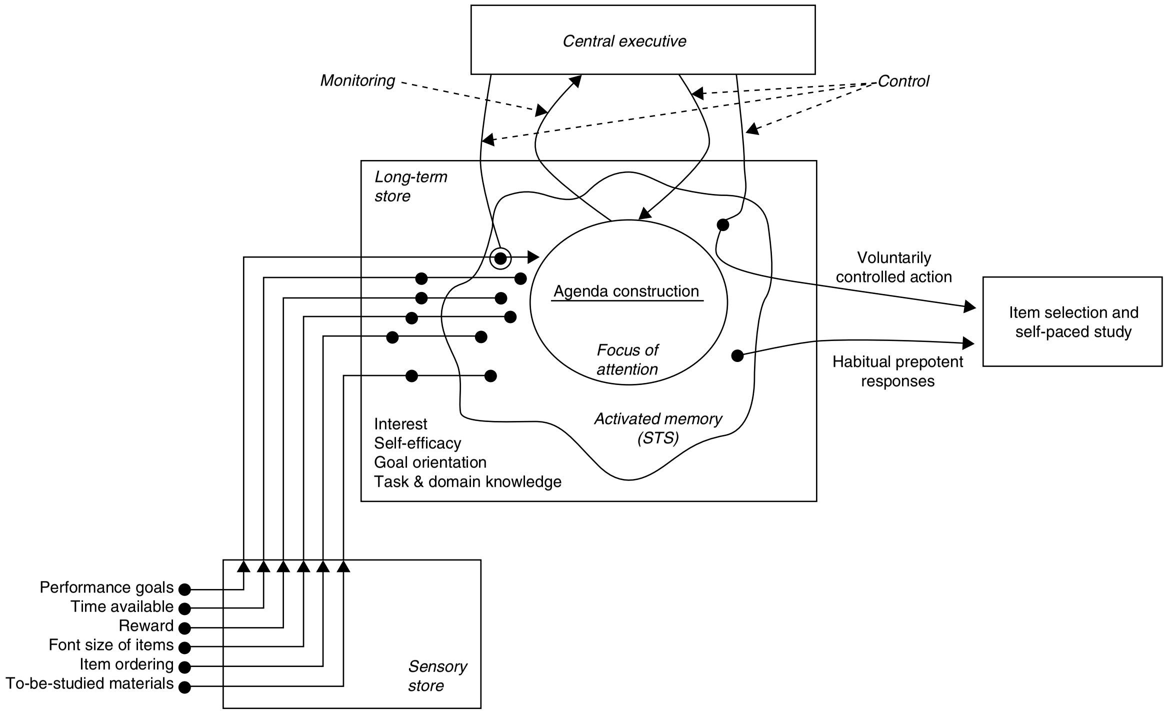

Figure 14.1 Agenda construction portrayed within Cowan's (1988) information-processing system. Learners' construct an agenda to maximize the likelihood of attaining goals using minimal effort. Agenda construction occurs within the focus of attention and involves selectively attending to task constraints within the environment (lower left) as well as individual characteristics (illustrated within the long-term store). See text for details

which is analogous to conscious awareness. The focus of attention is limited in capacity – that is, an individual can only be aware of a limited amount of information. In addition, attention can be captured by goal-irrelevant stimuli (internal or external) without intervention from the central executive. In this case, the central executive must draw attention back to goal-relevant information to complete the current goals. Cowan's model suggests that voluntary control of action can break down for a variety of reasons, including (a) decreases in activation of goal-relevant information or (b) activation of prepotent habitual responses.

By casting study-time allocation within this system, one can readily explain the dynamics of study-time allocation and can develop testable predictions to guide future research. From a bird's eye view, the ABR framework describes how the central executive directs attention to various task constraints (bottom-left of Figure 14.1) to construct an agenda. Agenda construction and execution occur within the focus of attention (Figure 14.1), and agenda execution via the central executive leads to the voluntary control of item selection and self-paced study (not shown). Although an exhaustive description of this system is beyond the scope of this paper, we expand our discussion on some aspects of the framework that are most critical for explaining how people allocate study time.

First, when learners are constructing an agenda prior to studying (Figure 14.1), the central executive monitors the environment for task-relevant information, which is transferred to (or becomes) the focus of attention. In the present illustration, the learner presumably is attending to the current performance and learning goals and is habituating other environmental stimuli, such as the time available for study, the font size of the to-be-studied materials, etcetera. In this case, a learner with a high-performance goal may develop an agenda that involves studying the majority of materials that are likely to be tested. In executing this agenda, the central executive would at least need to maintain activation of the agenda criteria and to voluntarily control action towards selecting the appropriate materials for study.

The relationship between the central executive and the memory system functions as the meta-level and object level (respectively) in Nelson and Naren's (1990) general model of monitoring-control relationships (for further details, see Dunlosky & Metcalfe, 2009; Van Overschelde, 2008). In particular, the central executive receives input from the underlying cognitive system (monitoring) and can use the information to change the state of this system or to redirect attention to the information in the environment or in memory (control). Most relevant here, this aspect of the framework highlights the reciprocal nature of agenda construction and execution. That is, although learners may develop an agenda to allocate time, if feedback from monitoring during agenda execution and study indicates that the agenda is not producing an expected level of progress, then the learner will presumably construct a new agenda. Such feedback from monitoring and its influence on control is a mainstay of general theories of self-regulated learning (Hertzog & Dunlosky, 2004; Winne & Hadwin, 1998).

Second, other information stored in long-term memory may be recruited when a learner constructs and executes agendas. For instance, learners with low memory self-efficacy may develop an agenda to focus on the easier materials that they believe they can learn, and those with a great deal of domain knowledge may develop an agenda that focuses on the most difficult materials (Metcalfe, 2002). Some individual characteristics – such as self-efficacy (Berry & West, 1993; Zimmerman et al., 1994) and interest within a domain (Son & Metcalfe, 2000) – may influence goal setting, which in turn could influence agenda construction. For simplicity, we only illustrated a subset of individual differences that could potentially influence agenda construction and execution, but others (e.g., need for cognition) could be influential as well.

Third, given that agenda construction and execution occur within the focus of attention, capacity limitations can undermine effective agenda use. For instance, individuals who are deficient in central-executive processes (e.g., have limited working-memory spans) may have difficulties constructing an effective agenda when the number of relevant task constraints exceeds their capacity. Likewise, even a relatively simple agenda may be difficult to execute (a) when learners must maintain too many task constraints within the focus of attention or (b) when distractions arise internally (e.g., mind wandering, Smallwood & Schooler, 2006) or in the environment (e.g., goal-irrelevant stimuli that capture attention). When capacity limits have been exceeded, learners may fail to fully execute an agenda because goal-relevant information is not actively maintained. Put differently, when overwhelmed, learners may simply forget their initial agenda and hence habitual responses will gain control of study-time allocation (Dunlosky & Thiede, 2004).

Fourth, prepotent habitual responses can influence study-time allocation. These processes are triggered by the stimulus environment and are not voluntarily controlled; thus, learners may not be aware that these stimuli are influencing their study-time allocation. The proposal here is similar to a recent one by Koriat and Nussinson (2009) that 'study time is data driven rather than goal driven; it is mainly determined ad hoc by the item itself – or more precisely, by the item-learner interaction' (p. 1338). Thus, a normatively difficult item may receive more study time because the item itself takes longer for a learner to process – not because the learner necessarily developed an agenda to voluntarily study difficult items longer than easier ones. Note, however, that the ABR framework does not necessitate that study time is mainly data driven and does predict that under some circumstances it will be largely agenda driven. More important, the current proposal is broader than previous ones in that environmental factors other than the to-be-studied items can trigger prepotent habitual responses that drive allocation. For instance, when studying items in an array (e.g., a textbook displaying a page of Foreign language translation equivalents), learners may first select items for study that are in the top left of the array and then move from left to right. In this case, item order (another task characteristic) may inadvertently drive allocation because a habitual response in Western culture is to read from left to right (Dunlosky & Ariel, 2009).

To summarize, the ABR framework is comprised of the attentional and memory components in Figure 14.1 along with the assumption that learners construct and execute agendas to help them obtain their task goals in an efficient manner. When capacity limits are not exceeded, learners are expected to successfully execute an agenda even when it is in opposition to habitual responding. By contrast, when capacity limitations are exceeded, agenda execution (which occurs within focal attention and hence requires resources) is expected to be disrupted and hence habitual responding triggered by the environment (which does not require resources) will have a larger influence on study-time allocation. To illustrate, learners may develop an agenda that selects those items for study that are most likely to be tested, whereas habitual responding may drive learners to focus on the most difficult items regardless of the likelihood of being tested. In this case, an individual with a low working-memory span who is studying in a noisy environment may have difficulties consistently executing the agenda and hence may be drawn towards focusing on the more difficult – albeit less valued – items.

Why might learners succeed or fail to allocate study time optimally?

A major premise of the ABR framework is that while learners may attempt to develop an agenda to efficiently achieve task goals, they may not achieve optimal learning (i.e., efficiently achieve the task goals) for numerous reasons. One reason is that learners may be misguided by inaccurate monitoring. An agenda criterion may be to 'avoid any item that may already be known,' and if so, learners may not choose to restudy any item that they judge as well learned. Because a learner's monitoring judgements can be inappropriately biased by a variety of factors (for detailed discussion, see Serra & Metcalfe, 2009), he or she may have difficulties appropriately executing agendas that rely on monitoring during study. In the literature on metacognitive monitoring, such biases are often portrayed as arising from an isolated factor, such as people's inferences about the fluency of processing (Alter & Oppenheimer, 2009). Bruce Whittlesea (1993, 2000, 2002) and his colleagues have pioneered analyses of the fluency heuristic in particular and how people implicitly use it to evaluate whether an item has been presented in the past and hence is being remembered now (for detailed discussion, see Arnold, this volume). Whittlesea's groundbreaking theory highlights how people's *expectations* about fluency can bias their self evaluations (Whittlesea & Leboe, 2003); hence, fluency in isolation may not entirely be responsible for illusions of monitoring that can arise during study. Unfortunately, such expectation-fluency interactions have not been investigated with respect to how people evaluate their ongoing learning and deserve systematic scrutiny.

Another reason learners may not achieve optimal learning is that they may fail to construct an agenda prior to studying. If so, their allocation of study may be largely driven by bottom-up processing that may not ensure optimal learning. Alternatively, learners may construct an agenda, but the agenda itself may be flawed. For instance, most learners' agendas will include the following criterion (Ariel et al., 2009): 'If I already know an item, then I shouldn't study it any more.' Consistent with this criterion, prevailing evidence from laboratory-based research indicates that learners typically do not restudy items that they can already recall (Metcalfe & Kornell, 2005; Nelson et al., 1994), and when using flashcards, students drop items from study after they correctly recall them once (Kornell & Bjork, 2008). Unfortunately, such a criterion can lead to suboptimal learning, because criterion test performance is even better when learners practice items after they can recall them once (e.g., Pyc & Rawson, 2009). Even when a learner's agenda would optimize learning, if the agenda is not appropriately executed, then optimality may be undermined. Thus, a learner may set out to study those items that will provide the largest reward with a minimal investment, yet in the face of distraction, habitual processes may override agenda execution and lead to suboptimal allocation of study time.

Some research has explored the degree to which learners regulate study time in a relatively optimal manner (e.g., Atkinson, 1972; Kornell & Metcalfe, 2006; Metcalfe & Kornell, 2003; Nelson et al., 1994). In general, learners do not appear to consistently make *optimal* decisions, but their allocation decisions do appear to benefit their learning, at least when compared to contexts where they are forced to allocate study time in a manner that is inconsistent with their initial decisions (Kornell & Metcalfe, 2006). In the remainder of this chapter, we turn from optimality and consider how the ABR framework accounts for learners' study-time allocation.

Explaining study-time allocation via agenda-based regulation

The previous sections provided a sketch of the ABR framework. Although much of this framework was inspired by previous theories of memory, metacognition, and planning (e.g., Miller et al., 1960; Nelson & Narens, 1990; and Cowan, 1988, respectively), we acknowledge that as a framework of study-time allocation, it is rather speculative and has not yet been systematically evaluated. Nevertheless, as we describe next, some predictions from the framework have received empirical support, and it also provides plausible explanations of many effects described in the literature.

Use of agendas in study-time allocation

Even though the ABR framework comprises more than just agenda construction and execution, if learners never used agendas to regulate their study time then it would be prudent to focus on non-agenda based theories of study-time allocation. We propose that a great deal of study-time allocation reported in the literature reflects learners' construction and execution of relatively simple agendas. For instance, when participants are instructed to learn all the items on a list and are given unlimited time to do so, they tend to not select already recallable items and instead focus their study on just the unlearned ones. By contrast, when learners are told that they will have very limited time for study, they shift towards studying the easiest unlearned items first (Metcalfe, 2002; Thiede & Dunlosky, 1999). This STEM effect is readily explained by agenda-based regulation. In the latter case with limited study time, participants presumably believe that they cannot learn all the items and that their performance would be better if they focused on the easiest items than if they wasted time trying to learn the more difficult ones.

Further evidence for agenda-based regulation arises from investigations in which (a) performance goals are manipulated by instruction and (b) the potential rewards for recalling items are manipulated across them. We previously described preliminary research by Thiede and Dunlosky (1999) that investigated the influence of performance goals. In particular, participants who received a low-performance goal were expected to develop an agenda that includes the criterion to restudy relatively few of the easiest items. This prediction was confirmed by the presence of a STEM effect: From the 30 to-be-learned items, these participants' restudy decisions favoured easier-to-learn over more difficult-to-learn items, and they chose to restudy only about 12 items.

Concerning rewards, Ariel et al. (2009) examined the degree to which various reward structures would influence item selection. In their experiments, participants received a familiarity trial with 30 pairs (e.g., BOOK – HAMMER) and then attempted to recall each one (i.e., BOOK – ?). Items were then presented simultaneously in an array, and participants could choose to restudy 15 of them. Most important, items were assigned either a high (90%) or low (30%) likelihood of being on the upcoming test (Experiment 1) or with a high (5 points) versus a low (1 point) point reward if correctly recalled on the criterion test (Experiment 2). Consistent with predictions from the ABR framework, participants predominantly chose to restudy items assigned to higher values, regardless of whether the higher reward was slated with normatively easy-to-learn or more difficult-to-learn items (see also, Price et al., in press).

Although these effects are in accord with the ABR framework, they do not provide direct evidence that participants are developing agendas to allocate their study time. To more directly scrutinize the contribution of agendas, Dunlosky and Thiede (2004) had participants perform the task described above with the instructions to achieve a low-performance goal (i.e., recall 6 of 30 items). After participants selected items for restudy, they were asked how they decided to select the specific items for restudy. Seventy one percent of the participants reported using criterion to select items that would isolate the easiest items for restudy, which is in accord with an agenda to obtain the task goals using minimal effort. As compared to participants who did not endorse using the agenda, those who did were more likely to focus restudy on easy items and chose reliably fewer items for restudy. In a follow-up study, participants again were given a low-performance goal and asked to choose items for restudy that were presented simultaneously in an array. An experimental group was instructed to use the agenda that was hypothesized as being used when no extra instructions are provided – that is, 'to choose as few of the easiest items that will allow you to achieve the performance goal.' Their study-time allocation was compared to an uninstructed group who were given no extra instructions. Our research strategy here was to use the instructed group as a simulation of the uninstructed group; if they both allocated time in the same manner, then one could infer that both were using the same rudimentary agenda. Both groups allocated time in a strikingly similar manner: they chose to restudy the easier items (both groups *M* correlation between JOLs and item selection was +.44) and chose to restudy around 9 items (Dunlosky & Thiede, 2004).

Such verbal reports about agenda use and instructional equivalence from Dunlosky and Thiede (2004) provide further evidence that some learners use agendas to allocate study time. Given that the contribution of agendas to allocation has received little attention in the field, a major goal will be to develop methods to more directly observe learners' agenda construction and execution. These methods will be especially vital for exploring the joint contribution of top-down and bottom-up processes to study-time allocation.

Resource limits on study-time allocation

Two general classes of outcomes demonstrate the possible contribution of resource limitations to agenda-based regulation – one pertains to the format of presenting items during selection and the other to individual differences in working-memory (WM) span. In these cases, overtaxing resources presumably disrupts learner's agenda execution and leads them to allocate study time in a seemingly non-optimal manner.

Concerning format, Ariel et al. (2009) investigated whether presentation format for selection influenced learners' allocation of study time to items that would be worth either 5 points or 1 point if correctly recalled. After a familiarity trial, the 30 items were represented, and participants were instructed to select 15 for restudy. When items were presented simultaneously for selection (Figure 14.2, left panel), participants were about 50 percent more likely to choose the higher valued items for restudy. Under the sequential format, the point value slated with a given item was also presented above it when the item appeared for the selection decision (Figure 14.2, right panel). Participants here still favoured the more highly valued items, but they were only about 20 percent more likely to choose them over the less valued items!

One explanation for such allocation failures under the sequential format arises from the resource demands of agenda execution. Namely, Thiede and Dunlosky (1999) argued that lesser demands are placed on processing resources when items are selected under the simultaneous than sequential format. When executing an agenda, the simultaneous format allows a learner to easily compare all items, to evaluate which items meet the agenda criteria, and to keep track of which agenda-relevant items have already been chosen. By contrast, the sequential format requires participants to keep extra information active in the focus of attention during agenda execution, such as how many items have been selected, which ones have already been selected, and which still require further study but have not yet been presented for selection. Along with keeping track of this goal-relevant information, learners must also maintain the agenda in the focus of attention. If resources for central-executive processing are taxed, then the agenda criteria would not remain activated in focal attention. In such cases, learners may fail to execute even the most rudimentary of agendas (e.g., focus only on high-valued items) and instead revert to allocating time as they typically would do so (e.g., choose any item that currently is not learned, even if doing so may mean that some high-valued items cannot be studied).

As important, such failures are expected to be magnified for individuals with low working-memory spans who presumably have difficulties with executive control. This prediction has been empirically evaluated in two studies (Ariel et al., 2009; Dunlosky & Thiede, 2004). Consider Ariel et al. (2009, Experiment 4), who had participants select items for restudy under the sequential format: some items were worth 5 points and others were worth 1 point. Each participant also completed the reading span task, which is a standard measure of working-memory performance (Daneman & Carpenter, 1980). Results from this study were presented above, but collapsed across all participants; that is, on average, participants favoured choosing the higher valued items less often under a sequential than simultaneous format. However, an embedded interaction involving span performance was evident, as illustrated in Figure 14.3. As shown in the left-hand panel, both high and low span participants more often chose highly valued items under the simultaneous format. By contrast, under the sequential format

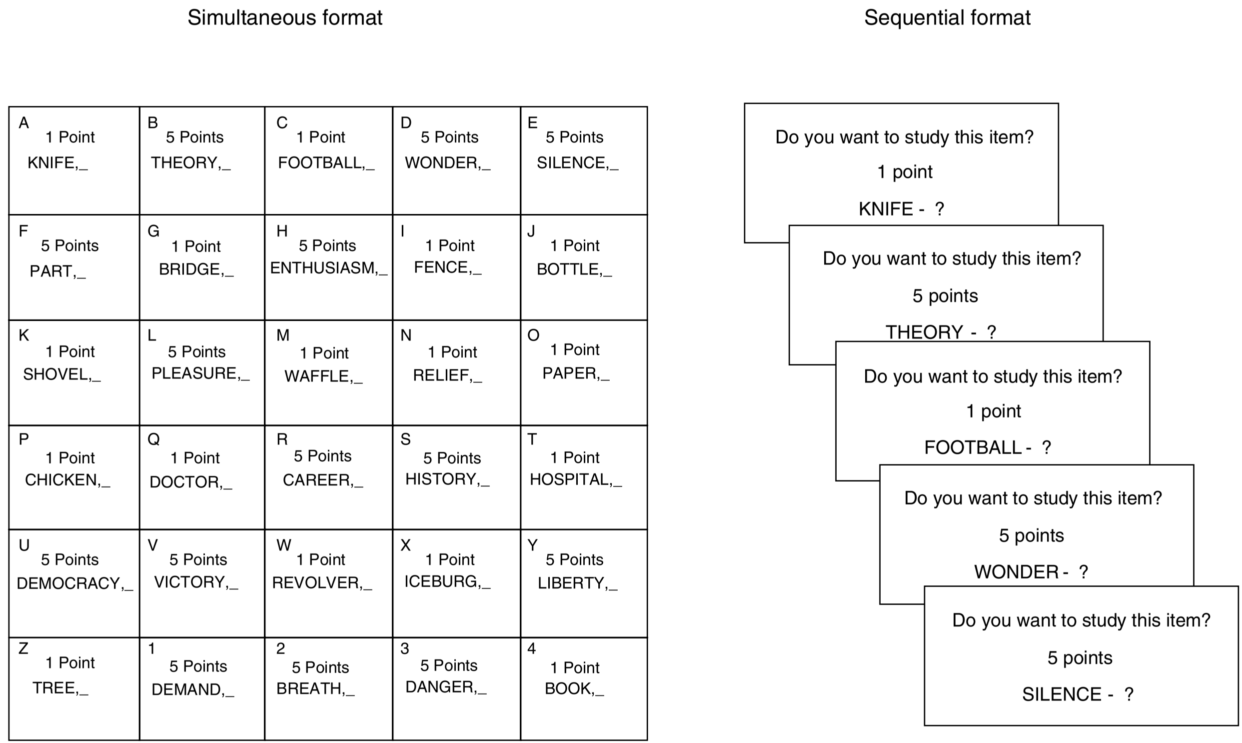

Figure 14.2 Two presentation formats used to display to-be-learned stimuli during item selection. For the simultaneous format (left panel), learners can easily compare across items as they make allocation decisions. For sequential presentation (right panel), learners view one item at a time and decide whether to restudy each one before moving to another item

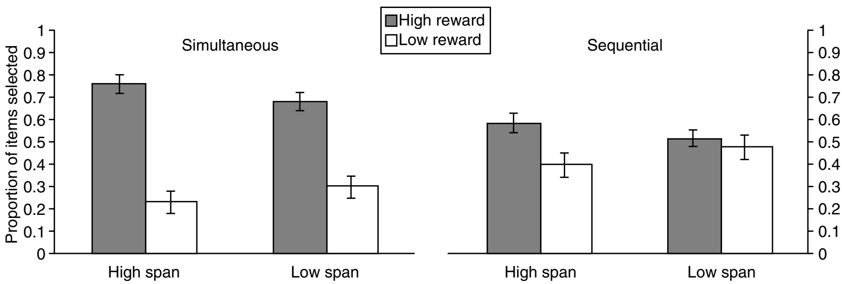

Figure 14.3 Mean proportion of items selected as a function of presentation format (adapted from Ariel et al., 2009). For the simultaneous format (left panel), both high and low span learners focus study time on items slated to receive higher point values if recalled correctly on the criterion test. For the sequential format (right panel), high span learners' still are more likely to select the higher reward items, whereas low span learners are not

(Figure 14.3, right-hand panel), high spans consistently chose to select the more highly valued items for restudy, whereas low spans did not favour high value items during selection. Although this evidence is correlational, the patterns are consistent with the proposal that agenda-based regulation requires resources and can be disrupted when those resources are taxed.

Bottom-up regulation of study

In addition to the top-down voluntary control of study that characterizes agenda-based regulation, certain stimuli in the environment may elicit responding without direct intervention of the central executive. These stimuli may activate a prepotent habitual response that results in a control decision that is neither elicited by the central executive or in the focus of attention. This proposal is new for theories of study-time allocation, but it is by no means novel from a theoretical or applied standpoint. Concerning the latter, the last time you planned to drive from home to somewhere new but found yourself driving to work, you experienced this trade-off between agenda-based and stimulus-driven control. In particular, your agenda was usurped by the habitual response to stimuli that you drive past every day to work.

We have begun to examine the contribution of bottom-up influences on study-time allocation (Dunlosky & Ariel, 2009; cf. Rhodes & Castel, 2009). Our research strategy is to observe the influence of stimuli that are expected to invoke a habitual response, but to do so in contexts where learners are expected to develop an agenda to allocate study time. In one preparation, participants are given a limited amount of time to select items for study from a 3 item array. These items vary in ease of learning (one is easy to learn, one is moderately difficult to learn, and one is very difficult). One might expect participants in such an experiment to develop an agenda to allocate study to the easiest items first, because the most difficult items would be too difficult to learn in a limited time span. This particular agenda – which involves allocating time to one's *region of proximal learning* (RPL) – can benefit learning and is used by learners in a variety of contexts (Kornell & Metcalfe, 2006; Metcalfe, 2002). Concerning habitual responses, individuals in Western cultures habitually read from left to right and this reading order bias may drive some learners away from executing the RPL agenda. Across multiple experiments, Dunlosky and Ariel (2009) found that many participants in fact did not focus restudy on items within their RPL, but began studying by choosing items in the left most position of the array. In such cases, item ordering triggered a prepotent habitual response to choose items on the left, which apparently overrode control by the central executive. Given the preliminary nature of these studies, further investigating the interplay between top-down and bottom-up process in study-time allocation is an important area for future research.

Further issues

The ABR framework proposes that top-down processes (relevant to agenda construction and execution) and bottom-up processes (relevant to habitual prepotent responding) jointly drive people's study-time allocation. Before our closing remarks, we briefly compare our framework to other models of study-time allocation, and we discuss two caveats about our current approach – one concerning the framework and another concerning our current review of the literature.

Are there competing models of study-time allocation?

Several other models can account for study-time allocation under some conditions, such as the discrepancy reduction model (Dunlosky & Thiede, 1998) and the region of proximal learning framework (Metcalfe, 2009). These are not competitors to the ABR framework – they are not mutually exclusive and are largely compatible (for detailed discussion, see Ariel et al., 2009). A major difference lies in the level of generality of the models.

The RPL framework and discrepancy reduction model specify the mechanisms of allocation at a fine-grained level. For instance, the RPL framework describes how learners use one particular agenda (based on allocating time to items in one's region of proximal learning) to allocate study time. Moreover, it assumes learners monitor the rate of learning to make decisions about terminating study. Both components of the RPL framework may be subsumed under the ABR framework, which emphasizes that learner's may use multiple agendas (including RPL) and may use monitoring in multiple ways to allocate study time. This emphasis is critical, however, simply because people do use more than one agenda to regulate learning and can use monitoring in multiple ways to terminate study. Thus, although we do not view the various models as competitors, a comprehensive theory of study-time allocation will need to explain how students construct and execute many kinds of agenda and how optimal regulation can be sidetracked by bottom-up processing.

Caveats concerning the ABR framework and our literature review

The ABR framework

The ABR framework is powerful in that it can be used to develop testable predictions to guide research programmes. This explanatory power, however, should be wielded with caution. Given that this framework often yields multiple hypotheses for any effects on study-time allocation in a post-hoc manner, one must be careful to qualify hypotheses that remain untested. For instance, consider the modal outcome from the first few decades of research on study-time allocation (for a review, see Son & Kornell, 2009): People often spent more time studying difficult-to-learn items than more easy ones. This relationship between difficulty and study time can be easily explained by agenda-based regulation. Namely, learners are instructed to perform well on the entire list, so they develop an agenda to maximize their likelihood of doing so.

Although plausible, an alternative hypothesis is that item difficulty directly drives study-time allocation in a manner that sidesteps voluntary control by the central executive. Other hypotheses can also be developed using the ABR framework. The main point here is that the ABR framework will typically provide multiple post-hoc hypotheses for the effect of any variable on study-time explanation. Thus, confidence in a particular hypothesis should arise from competitive evaluations among the leading ones.

Our literature review

Two caveats concern our literature review. First, we did not exhaustively review the extensive literature on study-time allocation; as discussed above, however, the ABR framework would likely provide several plausible explanations for any effects we did not discuss. Second, and more important, we did not consistently distinguish between item selection and the termination of self-paced study. Most of the research reviewed concerned learners' decisions about item selection, because the STEM effect (which inspired the development of the ABR model) has been exclusively demonstrated in item selection.

An issue arises concerning the degree to which item selection and self-paced study arise from the same mechanisms. As shown in Figure 14.1, we are assuming that both bottom-up and top-down processes can contribute to these two kinds of study-time allocation. This assumption requires further evaluation. For instance, Koriat and Nussinson (2009) proposed that 'study time is data driven rather than goal driven' (p. 1338); the analogous claim in the context of the ABR framework would be that bottom-up processes (which involve an interaction between stimuli and the individual's memory system) are solely responsible for self-paced study. Although possible, some evidence already suggests that self-paced study time can be goal driven. For instance, Dunlosky and Thiede (1998) reported that learner's self-paced study times were moderated by point values (and the likelihood of being tested). More recently, Ariel et al. (2009) found that the same effects of point value on item selection were also evident on self-paced study. Thus, it appears that self-paced study can be goal oriented, although the degree to which agenda-based regulation influences item selection and self-paced study may differ. Given that item selection and self-paced study are the most highly investigated indicators of study-time allocation, a major goal for future research will be to systematically explore the degree to which they are driven by different cognitive and metacognitive processes.

Closing remarks

Since the seminal work was conducted on self-paced study, researchers have developed an extensive body of data that in turn has yielded plausible hypotheses for how people allocate study time. These hypotheses have typically emphasized how learners use their monitoring of item difficulty to make decisions about which items to study and how long to study them. Although mechanisms based on monitoring item difficulty can account for some allocation behaviour, they consistently fall short of providing a complete account because they fail to acknowledge the role of agenda construction and execution in study-time allocation. Accordingly, we proposed the ABR framework (Ariel et al., 2009), which assumes that study-time allocation is a joint product of top-down (agenda-based) and bottom-up (habitual responses) processes that are enacted within a capacity limited memory system (Cowan, 1988). The ABR framework can be used to develop new predictions about when people's study-time allocation will benefit learning and when it may be suboptimal. Further refinement of the framework will certainly be necessary, such as with respect to the basis of self-evaluation during study (vis-à-vis Whittlesea's analyses of fluency and remembering) and how top-down and bottom-up processes interact. We welcome such refinements and are hopeful that the present analysis of the ABR framework will help to foster research in this important field.

Note

Development of this project was supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind and Behavior Collaborative Award. This research was also partially supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B070018 to Keith Thiede, Thomas D. Griffin, and Jennifer Wiley. The opinions expressed are those of the authors and do not represent views of the Institute or U.S. Department of Education.

Part V

SCAPE

15

Surprising Fluency: Bruce Whittlesea's Contributions to Our Understanding of the Role of Fundamental Adaptive Cognitive Processes

Antonia Mantonakis and Reid Hastie

Introduction

The present essay provides a selective review of Bruce Whittlesea's contributions to our understanding of memory and cognition, with a focus on his original and unique insights about heuristic reasoning. We start by outlining Whittlesea's general approach to the study of memory, called *Selective Construction and Preservation of Experiences* (SCAPE; Whittlesea, 1997). Within that larger system, we focus on two profound, but under-appreciated, contributions: Whittlesea's subtle analysis of the metacognitive experience of *fluency*; and his identification of a collection of *heuristic inference processes* that take fluency as a primary input to make judgements about familiarity, classification, and preference. We conclude with an original proposal to add two fundamental heuristics – affect-evaluation and causal-abduction – to Whittlesea's original set of fluency, generation, and resemblance.

Whittlesea often begins his papers with fables from the history of science, especially stories that show how obvious questions and answers are misleading. His home field of research is dominated by the assumption that the fundamental function of our memory system is to take snapshots of experience and store them in various filing systems for later use on tests of our ability to recognize and recall information from the past. He believes this is nonsense and, perhaps a product of the fact that so many memory researchers spend so much of their time lecturing to students and then testing to see how much content remains filed in their heads when final exam time comes at the end of every school term.

Instead, Whittlesea proposes that the human cognitive system has been designed and trained to support adaptive performance in a complex, constantly changing physical and social environment – to help us navigate through the world, making choices that will promote survival and reproduction (Whittlesea, 2003). Whittlesea believes that the primary function of our unitary memory system is to support the construction of a mental model of the current environment. This model is inferred from sensations and it preserves processing experiences, as well as sensory percepts. Processing experiences are determined partly by incoming perception, but also by the person's active goals, recent context, and interactions with prior experiences. Thus, the common distinction (made by other memory researchers) between encoding and retrieval is violated in practice, as encoding inextricably involves retrieval, at least for people more than a few months old. He also believes that many other distinctions, for example between separate memory stores for apparently different kinds of information, are misguided.

The construction of mental models

The activity of constructing and updating a mental model has two components: actual construction of the model and evaluation of the construction experience to assess the validity of that model. Suppose you walk into a retail store and focus your attention on a digital camera. Without deliberately intending to, you are creating a mental model of that object in its context and driven by your current goals. The representation would be different if your goals were different (you were cutting through the store to get to the parking lot), if the context were different (you saw the camera in a friend's apartment),or if your prior experiences were different (you had never seen a digital camera before).

Your first 'production' is to create a mental image of the physical object, probably adding some geometric information based on a lifetime of perceiving 3-dimensional objects, and probably some descriptive information, based on your prior experiences with cameras. (Obviously, the production of a mental representation involves using remembered information from previous experiences with similar objects.) Each production process is immediately accompanied by an automatic evaluation of that experience that tells you how coherent the process was. These collateral evaluations give you feedback on the validity of your initial mental model. If the physical object (or its labels) are confusing (perhaps because the camera looks different from any camera you've seen before) you will have a sense of difficult processing or disfluency; if it was processed easily (perhaps because it looks like a camera you own), it will lead to the metacognitive experience of fluency. The main adaptive function of this evaluative cycle is to signal whether or not you need to allocate more attention to the production process. If the construction feels disfluent, you will allocate more attention to verify and elaborate your initial mental model (e.g., Alter et al., 2007).

A second production process that is also likely to occur involves relating the visual inputs to previously encountered individual cameras, resulting in the *generation* of a category classification summarizing the identity of the camera or bringing to mind previously learned information about the camera. If this *generation process* occurs, it is also immediately evaluated for coherence, based on the speed and completeness of resolution of the search for similar exemplars in memory. A third production process is likely to occur that involves relating the visual information to general knowledge about abstracted 'good examples' of cameras in general. The speed and completeness of this search for a generic identity for the new object is also automatically evaluated resulting in an assessment of the coherence of the results of that *resemblance process*.

Which production processes occur as you encounter a new object will be determined by the context (in the vignette you have just entered a retail store) and your current goals (e.g., what is produced in the construction of a mental model of the object will depend on whether or not you are shopping for a camera). Whittlesea is especially interested in the complex inferences that people make when they rely on the experienced fluency of a production process to make judgements about whether or not the focal stimulus had been encountered before. He rigorously demonstrated that fluency plays a significant role in resemblance and generation processes as well as in perception-identification, supporting his assertion that tasks which can be distinguished operationally – recognition, recall, classification – are actually performed by a unitary cognitive system.

Whittlesea did not limit the constructive production processes to recognition, recall, and classification, although those are the processes most relevant to performance of laboratory memory tasks. Towards the end of this essay, we will propose that two additional processes play major roles in everyday activities: a causal-abduction process that is constantly seeking causal relationships between the events we experience and an affective-valuation process that gives us quick cues as to whether we should approach or avoid objects, people, and events that we encounter in daily life. In the example of the digital camera, the causal-abduction process would add the information to our representation that the camera is probably high-quality because its price is high and that it would be useful at our cousin's upcoming wedding; while the affective-valuation process would cue us that the camera is desirable.

The evaluation of mental models

In Whittlesea's SCAPE framework, the metacognitive evaluation process 'comments' on the validity of the production processes that are operating to construct the current mental model. Evaluations assess the fit between expectations (which are mostly based on mental models of previous experiences in similar situations) and the current situation model. The evaluation process occurs to assure the perceiver that the mental model 'makes sense'. Usually, evaluation gives rise to perceptions of *coherence* and *integrality* when implicit or explicit expectations are confirmed by recent experience and signal that the mental model is valid. If something does not make sense, perceptions of *discrepancy* and *incongruity* occur, signalling that the perceiver needs to allocate more attention to the environment and to reconsider the current mental model. Whittlesea's typology of four evaluations is based on two distinctions: some production/construction activities result in confirmation of expectations and some result in disconfirmations; and some expectations are definite and explicit and some are general and implicit. Table 15.1 summarizes the relationships between the four types of evaluations and the underlying two-dimensional framework.

The most common evaluation during any construction process (perception-identification, generation, resemblance) is *coherence*: 'Everything I'm experiencing right now is sensible and fits with my implicit expectations.' An implicit expectation is an unconscious readiness for an event that is based on prior encounters with similar events (Whittlesea & Williams, 2001b), based on the interaction between memories and current sensory experiences. Coherence occurs when the elements of current experience are expectable, leading to a feeling of comprehension, consistency, or goodness. As we will see, the fact that coherence signals goodness, leads people to reason backwards from a pleasant or fluent comprehension experience to infer that an event has been encountered before. An example of the perception of coherence occurs when a wife sees her spouse in the kitchen. She simply knows this incoming information 'fits' without having an explicit expectation; if there is a subjective feeling it is simply 'nothing special, everything is okay.' In our example of the digital camera, if the mental image of the camera is sensible and consistent with previous encounters with cameras, the perceiver has a low-level feeling of coherence or fluency that does not draw additional attention.

Table 15.1 Evaluations based on nature of an expectation and whether it is validated or violated

	Implicit or Indefinite Expectation	**Explicit or Definite Expectation**
Expectation Validated	**Coherence**: metacognitive feeling of low-level fluency or 'correctness'	**Integrality**: conscious feeling of 'completion'
Expectation Violated	**Discrepancy**: metacognitive feeling of high level of fluency or disfluency	**Incongruity**: conscious feeling that 'something is wrong'

The experience of *integrality* occurs when there is a validation of a definite, explicit expectation. This evaluation is often experienced when situations are frequently experienced, such as when driving to one's best friend's house and consciously thinking, 'There's a sharp curve at the bottom of the next hill.' In the example of the digital camera, the perceiver might explicitly wonder, 'Where's the on/off switch, it should be near the snapshot button?' and experience integrality when the on/off switch is identified where it was expected. Or experience a feeling of integrality after seeing a display of cameras under a 'Cameras' sign in an electronics store. Unlike coherence, which produces a reassuring feeling of low-level goodness, integrality produces a feeling of definiteness, belongingness, or unity, analogous to a mild feeling of insight. It does not usually lead to a judgement of re-encountering a specific previous event, but rather to an attribution of knowledge (Whittlesea, 2002). Like coherence, it does not interrupt the construction process to demand additional attention.

A third perception, *discrepancy*, is like coherence, in that it also results from essentially automatic, implicit, indefinite expectations. While coherence is the common experience that 'things make sense,' the perception of discrepancy occurs when 'things' violate expectations that are indefinite. The experience of discrepancy is complicated and rarely happens in real life. It's the perception of 'it doesn't fit,' accompanied by surprise. Like coherence, because the expectation is implicit, the perception of discrepancy is caused by something for which one cannot easily identify the source. The classic example is when a familiar stimulus is encountered in a novel context; you run into a friend from work, while vacationing in an exotic locale. You implicitly expect the faces you encounter to be unfamiliar and novel in the novel setting, but suddenly one face provides a much too intense experience of perceptual fluency, leading to an inference that 'I must know that person from somewhere,' to explain the unexpected experience of fluency (Whittlesea & Williams, 1998, 2001b). In the digital camera example, maybe you've seen an image of a digital camera in an advertisement that is an exact match to the camera you see in the store; but suppose you fail to recall the earlier ad. One of the cameras would seem to 'pop-out' perceptually producing a feeling of surprise.

The fourth evaluation, *incongruity*, occurs when there is a violation of definite, explicit expectation, and the source of the violation is immediately available. It is a feeling of surprise and the perception of 'wrongness,' whereby the components of a stimulus fit together badly, and where the perceiver can easily identify the source of the surprise. This is the evaluation you might experience if you turned a corner while driving and encountered a new house where there had previously been a vacant lot; or on hearing the phrase 'row, row, row your GOAT.' In the digital camera example, a person would be likely to experience incongruity if she saw a button labelled 'purée' on the camera while looking for the on/off button or if she saw a camera in the middle of display case in a bakery. She would be surprised and immediately know that this function did not fit with a camera. When incongruity is experienced, the source of the surprise is usually correctly attributed to its true source, because the expectation is almost always explicit, partly conscious (Whittlesea & Williams, 2001b).

A key distinction in Whittlesea's typology of evaluations is between explicit, definite expectations and general, implicit evaluations. Across his many experiments, Whittlesea operationalized this distinction in many ways, as well as providing many evocative examples. Because most of his research is embedded in the academic memory and verbal learning tradition, most of the examples and experimental manipulations involve subtle variations in verbal materials. For example, certain linguistic phrases set up explicit, meaning-constraining expectations: *Mary had a little???*. While others create general, open-ended expectations: *Mary entered the barnyard and saw a???*. These expectations can be validated (*Mary had a little LAMB* or *Mary entered the barnyard and saw a LAMB*) or violated (*Mary had a little LION* or *Mary entered the barnyard and saw a LION*). As another example of an implicit expectation that is violated, an orthographically regular letter-string such as HENSION leads to the expectation that you are reading a word, but the during the comprehension process you realize, with a mild feeling of surprise, that it's not a word. In contrast, an orthographically irregular string such as LICTPUB leads you to have the expectation that it is a non-word, and that implicit expectation is fulfilled (Whittlesea & Williams, 1998). (Although these subtle linguistic manipulations are elegant operationalizations of Whittlesea's theoretical distinctions, we fear that the heavy reliance on verbal materials and nuanced manipulations may obscure the truly fundamental and general character of the implications of Whittlesea's empirical demonstrations. This is part of our motivation for introducing the new digital camera example in this essay.)

Whittlesea's best-known experimental contribution is his elegant analysis of the manner in which the violation of indefinite, implicit expectations can produce strong feelings of familiarity. More specifically, he has been fascinated by the perception of discrepancy and its impact on subsequent memory judgements. Using many different experimental conditions he produced surprising comprehension experiences – surprising fluencies and disfluencies – for his participants and in some key treatments, he leads them to explain the surprise by attributing it to an earlier experience or familiarity with the target material. He describes four varieties of discrepancy: surprising consistency, surprising redintegration, surprising coherence, and surprising incongruity (Whittlesea & Williams, 2001b). *Surprising consistency* could be produced, for example, by presenting phonologically plausible non-words (e.g., HENSION), the initial perceptual processes are fast and fluent (as the phonological implications of the letter string are extracted), creating an expectation that a meaning will come to mind. When no meaning is produced as part of the comprehension process, the person is surprised and puzzled, but then realizes that the letter string is a non-word.

In Whittlesea's experiments, the surprising fluency experience leads to a feeling of familiarity and to false alarm responses (to items like HENSION) on recognition memory tests. This led him to spell-out four conditions that were collectively necessary and sufficient to produce the illusion of familiarity and he called these conditions the Discrepancy-Attribution Hypothesis (Whittlesea & Williams, 2000): (1) a general expectation is evoked, (2) the participant is uncertain about whether the expectation had been fulfilled or not, (3) the expectation is confirmed, (4) the feeling of surprising fluency (produced by the first three conditions) is attributed to a past encounter with the stimulus and not to some other source.

Whittlesea also has studied conditions that produce *surprising redintegration*, usually by presenting a letter string that initially seemed difficult to pronounce and meaningless, but which is interpreted with further thought as a meaningful word (e.g., PHRAUG). The emotions evoked in surprising redintegration are typically a startle reaction followed by an 'ah-ha' insight. Finally, there are experiments on *surprising coherence*, in which a semantically satisfying, but not highly constrained sentence completion is presented: 'Mary entered the barnyard and saw a LAMB.' As in the studies of surprising inconsistency, Whittlesea cleverly induced his participants to attribute the feelings of surprising fluency to prior experience, leading them to infer that the presented materials had been seen previously in the experiment, producing false alarm responses on recognition memory tasks. (Whittesea provides examples of *surprising incongruity*, but did not study it under controlled conditions.)

Up to this point we have reviewed two components of Whittlesea's general SCAPE framework for cognition with its emphasis on the adaptive function of creating a useful mental model of the current situation (an overview of the system is provided in Figure 15.1). The constant cycle of production processes (perception-identification, generation, resemblance) and evaluation processes (coherence, integrality, discrepancy, incongruity) which create and update mental models of the current situation. We paid special attention to Whittlesea's analysis of discrepancy, as this rare experience provides some deep insights into everyday cognition and produces some spectacular memory illusions. The third component is a collection of heuristic inference processes that map the mental model onto the cognitive and behavioural actions that perform important adaptive tasks.

Cognitive heuristics

The concept of heuristic processing has been around in Psychology for at least a century. It is probably fair to date its introduction with Hermann von Helmholtz's proposal that *unconscious inferences* were behind the apparently

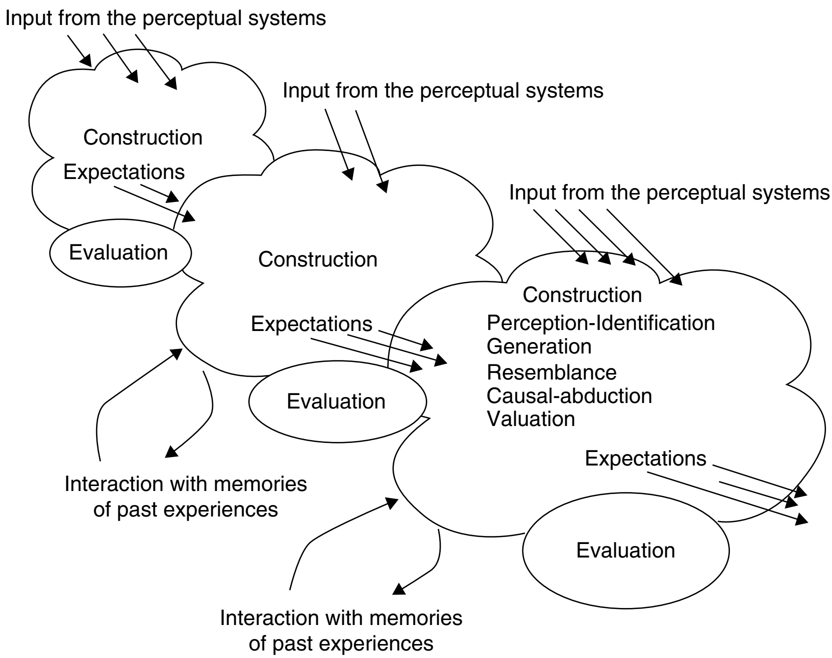

Figure 15.1 An overview of the SCAPE construction-evaluation cycle, summarizing the relationships between construction, evaluation, and expectations; n.b., the construction process has been elaborated to include causal-abduction and valuation processes in addition to Whittlesea's original perception-identification, generation, and resemblance processes (see the concluding section of this essay for the authors' proposal to add these processes to SCAPE)

effortless impressions of distance, location, and colour derived from visual and acoustic sensations (Helmholtz, 1867/1950). Helmholtz realized that our apparently effortless perceptual achievements required a vast number of logic-like inferences that allowed us to form visual impressions of distance, object identity, colour constancy, and analogous achievements with our auditory systems. The term *heuristic* was introduced into modern Psychology by Allen Newell and Herbert Simon (1972) who used it to refer to short-cut strategies that we spontaneously rely on to solve problems and make decisions. Like Helmholtz's inferences, these habits are usually automatic and unconscious and are robust, efficient, and usually provide a practical solution. And like Helmholtz's inferences they can sometimes produce cognitive distortions and illusions. (Newell and Simon probably borrowed the term from George Polya's famous 1945 book on mathematical problem solving, *How to Solve It;* another source is the concept of a 'rule of thumb' that originally referred to wood workers' short-cut use of their thumbs to measure length; see Hoff-Sommers' [1995] citation of Philip Hiscock, Canadian Folklorist).

The term *heuristic* became outrageously popular following the research on judgement under uncertainty by Tversky and Kahneman (1974; see Shah & Oppenheimer, 2008, for a comprehensive review of the concept in Psychology). Among other things, this research programme challenged the assumption that the cognitive system was essentially rational and identified many limitations and biases in heuristic judgements and decisions (Keren & Teigen, 2004).

It is important to distinguish between *procedural heuristics*, exemplified by Simon's Satisficing Heuristic (1955) for making decisions (see also Payne et al., 1993), Gigerenzer's Take-the-Best and Priority Heuristics for making choices (Brandstätter et al., 2006), and Tversky and Kahneman's (1974) Anchor-and-Adjust Heuristic for estimating quantities, versus *substitution heuristics* exemplified by Tversky and Kahnman's (1974) Availability and Representativeness Heuristics, and Gigerenzer's (Gigerenzer et al., 1999; Goldstein & Gigerenzer, 2002) Recognition heuristic (see also Kahneman, 2003). Procedural heuristics involve a series of explicit, partly deliberate inferential steps to 'calculate' an answer from several sources of information; substitution heuristics involve the implicit, mostly automatic and unconscious substitution of one dimension of experience (e.g., fluency) to render a judgement on another dimension (familiarity, probability, frequency, likeability).

Whittlesea uses the term heuristic to refer to the inferences that connect the mental model construction and evaluation processes to decisions (e.g., 'Have I seen this person before?') and actions (e.g., responding appropriately to a person in the role of physician). He has focused his research on the conditions under which people would substitute an impression of fluency for a memory, classification, or liking judgement.

The processes that create a mental model and evaluate it do not lead directly to behaviours that solve the problems posed by the demands of everyday adaptive challenges. In order to understand how specific actions follow from mental models, we need to also understand the demands of the current cognitive task that the person is performing and the subtle heuristic inference processes that map properties of the mental model and the evaluation of its production processes onto appropriate actions. For example, if we want to know if a new experience is a repeat of a previous event, we cannot look the answer up in a memory filing system that has been indexed with tags saying, 'You were here.' Rather, we have by-products of the production and evaluation processes and the contents of the current mental model, that can be used to *infer* whether a new experience is familiar and therefore probably 'old' ('I've met that person before') or to *infer* whether deliberately recalled elements of a prior experience are complete and valid ('I got my last haircut three weeks ago in the salon on Queen Street').

Similarly, when we try to verify whether a factual assertion is true (Toronto is larger than Vancouver) or a prediction is accurate (the Leafs are going to win the Stanley Cup next year) we must rely on indirect cues to infer validity. These inferences are heuristic in the sense that they are not perfectly logical, not certainly true, and are fairly easy to make. Sometimes validation relies on information and implications of information from memory or available in the immediate environment, but often we rely on qualities of our evaluation experiences to assess validity. The aspect of experience that we usually rely on is the apparent coherence or fluency of our cognitive processing of the assertion and information that is associated with it.

The fluency heuristic

When Bruce Whittlesea began to conduct his first original research on conceptual representations and memory in the mid-1970s, the concepts of metacognition and fluency had just been introduced into cognitive psychology by Lila Gleitman (Gleitman et al., 1972) and by Amos Tversky and Daniel Kahneman (1973) respectively. Both terms were usually used to refer to aspects of the memory system; metacognition to tip-of-the-tongue (Brown & McNeil, 1966) and feeling-of-knowing (Hart, 1967) experiences and fluency (originally called 'availability') to the assessment of 'the ease with which the relevant mental operation of retrieval, construction, or association can be carried out' (Tversky & Kahneman, 1973, p. 208) in the context of estimates of frequencies and probabilities. Once introduced, these concepts attracted enormous amounts of attention and became the focus of research on a diverse collection of cognitive tasks in many laboratories (e.g., Alter & Oppenheimer, 2009; Flavell, 1979; Nelson, 1996a; Schwarz, 2004).[1]

One of the most important developments in tidal wave of research on metacognition and fluency was Larry Jacoby's promotion of the idea that recognition memory depended heavily on the fluency of perceptual and retrieval operations (Jacoby & Dallas, 1981). Earlier theorists (Atkinson & Juola, 1974; Jacoby, 1991; Kintsch, 1967; Mandler, 1980) had proposed that familiarity was one of the two primary bases of recognition. Jacoby's contribution was to verify that fluency was the key ingredient in familiarity with a series of dramatic demonstrations that memory illusions could be produced by manipulations of fluency. Thus, for example, manipulations of fluency produce false recognition (Jacoby & Whitehouse, 1989); illusions that a name is famous (Jacoby et al., 1989); and false belief in the veracity of incorrect facts (Bernstein, 2005).

As it happened, Whittlesea was in graduate school at exactly the time and in the same place that Jacoby and his colleagues were developing the theoretical and empirical case for the importance of fluency in many types of adaptive judgements. Whittesea's dissertation (1984) includes extensive discussions of the role of perceptual fluency in memory and categorization judgements (he was especially influenced by Jacoby & Dallas, 1981, and by his advisor Lee Brooks). There are other precursors of the SCAPE framework in his dissertation, including the outlines of an overarching unitary processing system, description of a crucial interaction between cognitive and metacognitive processes, and constant references to the *adaptive* functions of cognition. A fundamental insight in this early work is that the contents of a memory experience (sometimes called 'item information') and the feeling of the experience (metacognitive fluency or familiarity) can have independent effects on judgements and behaviours.

Perhaps the cleanest demonstrations of the dissociation between content and process were provided after Jacoby and Whittlesea's seminal research, in a series of clever experiments by the social psychologist Norbert Schwarz and his colleagues. When some participants in Schwarz's experiments recall six examples of their own assertive behaviour, they rate themselves as more assertive than other participants who recall 12 examples (Schwarz, et al., 1991, Experiment 1). The experience of trying to recall 6 examples is easy, whereas trying to recall 12 examples is difficult; participants misattribute this easy (fluent) or difficult (disfluent) experience to their own personalities. Paradoxically, those who recall fewer examples (i.e., *less* information) judge themselves to be *more* assertive. However, when background music is made salient (which the experimenter suggested could make the recall task difficult), participants in both conditions focus on the amount recalled (information content), and the group who recalls 12 examples rate themselves as more assertive than the group who had to recall 6 examples (Schwarz, et al., 1991, Experiment 3). Schwarz and his colleagues provided many additional demonstrations that misattributions of fluency could perturb judgements as disparate as life-satisfaction, beauty, and riskiness (Reber et al., 2004; Song & Schwarz, 2009).

Jacoby should be credited with introducing the term 'fluency heuristic' and producing some brilliant demonstrations of its effects; but Whittlesea's research provides the definitive detailed analysis of the heuristic process. First, he showed that when fluency has an effect on judgements it is usually *relative* fluency, not absolute fluency that is important. In many situations, the perception of coherence when an object or event is encountered leads to a mild feeling of fluency and that can be enough to support a, 'Yes, I've experienced this event before' judgement. But, the most dramatic effects of fluency occur when it is surprisingly high or low compared to some expectation. Second, Whittlesea showed that these expectations are often specific to the target event and not based exclusively on general, distributional impressions (Whittlesea & Williams, 2001b). The evidence for this claim, that expectations for individual events are created 'on the fly' but could then produce a feeling of surprise associated with the original events, is based on the subtle manner in which context is used to discount high and low fluency experiences. In addition, Whittlesea showed that surprising fluency effects are strongest when there is a brief pause in time between presentation of a stimulus and the request for a judgement of that stimulus; also consistent with the hypothesis that norms are 'computed on the fly' based on the immediate combination of stimulus, context, and goal at the time the stimulus was encountered (cf. Kahneman & Miller, 1986).[2]

Whittlesea's third insight is to recognize the necessity of some kind of logic-like inference that the experience of (differential) fluency is due to the subject of the judgement. If a person experiences surprising fluency when encountering a novel object or event, that experience has to be attributed, for example, to a past encounter (memory), to membership in a category (classification), or to general abstracted knowledge, for it to have an impact on an expressed judgement about the event. If the fluency experience is attributed to something other than the subject of the judgement, for example to something about the person (e.g., something the person ate) or something else in the context (e.g., music playing in the background), then it would not affect the focal judgement.

The generation and resemblance heuristics

Whittlesea contrasts the fluency heuristic (a 'quality of processing heuristic') with 'information heuristics,' which are based on the contents of the mental model that a person has (see discussion of process versus content effects on judgements of assertiveness above). The generation heuristic is based on the coming-to-mind from memory of *specific* information associated with the stimulus from past encounters. If vivid, detailed information about a context in which a stimulus might have been encountered previously comes to mind quickly when the stimulus is encountered (e.g., when you see a face in the crowd in a restaurant you think of seeing the face in a classroom), it supports the inference that the stimulus has been encountered before. Manipulations involving semantic priming (e.g., pairing LION with TIGER) lead to the use of the generation heuristic, and can influence memory recall (Leboe & Whittlesea, 2002).

The resemblance heuristic is based on remembering *general* information associated with the current stimulus. A manipulation involving changing the 'norm' of what is learned about a general class of items, leads to the use of the resemblance heuristic, and influences judgements of classification (Whittlesea & Leboe, 2000). Only when a person can use the entire set of items encountered to make a classification judgement, do people use resemblance. In most situations relying on information that comes to mind in response to a new experience is good evidence for a prior encounter. However, when the stimulus is part of a larger thematic event (e.g., a list of highly associated words, a stylized narrative) contextual information is likely to come to mind because of the theme, and lead to false recognition and even confident false recalls (cf. Bartlett, 1932; Bransford et al., 1972; Roediger & McDermott, 1995)

A modest proposal to expand SCAPE

This essay was written to express our appreciation of Whittlesea's seminal ideas about fluency and heuristic inferences and his elegant experimental dissections of the roles of these operations in memory and classification tasks. We would like to propose a modest revision of his SCAPE framework. We propose that there are some essential highly practiced cognitive capacities that are shared by all normal adults in all human cultures. Some have called these *elementary information operations* or *processes* (Newell & Simon, 1972; Posner & McLeod, 1982). Kahneman (2003) proposed that each major judgement heuristic was based on the substitution of a *natural assessment* for a more difficult evaluation. Whittlesea provides an elegant conceptual and behavioural analysis of the workings of three basic automatic cognitive capacities: recognition, generation (associative retrieval of related specific ideas, called availability by Tversky & Kahneman, 1974), and resemblance (similarity, called representativeness by Tversky & Kahenman, 1974).

We propose that at least two other fundamental cognitive 'operations' are ubiquitous and important in everyday life, although they did not play a central role in the laboratory memory and classification tasks that are the focus of Whittlesea's research. Following Robert Zajonc (1980) and Paul Slovic (Slovic, Finucane et al., 2002) we propose that an affective-valuation process that assesses the personal value or utility of objects in the immediate environment is automatic in many situations and may even precede classification processes. For example, if your current goal were shopping, you will automatically judge the goodness or utility of goal-relevant objects, such as a camera you might purchase. As with other production processes, an impression of coherence will moderate your valuation and surprising fluency would be predicted to exaggerate your evaluative reaction (see Whittlesea's experiments on the 'mere exposure effect,' Whittlesea & Price, 2001).

Finally, there is ample evidence for a fifth production process that identifies physical and social causal affordances for objects in the current environment is also a candidate for a fundamental mental model production process (Gibson, 1977; Michotte, 1963; Scholl & Tremoulet, 2000; Uleman et al., 2008). So, for example, when you encountered the camera in a store display, you would automatically think of how it might be used or which of your current goals it might satisfy.

As with the other production processes identified by Whittlesea, the affective-valuation and causal-affordance production processes would be immediately accompanied by metacognitive evaluations of the ease and completeness of the experiences. Taken together, recognition of objects and events, generation of associated ideas, similarity comparisons, and judgements of valuation and causation compose a toolbox of adaptive capacities that account for most of our automatic cognitive capacities. We further propose that each of these capacities is evaluated according to the ease with which they are executed; the conclusions from the application of each capacity are moderated by subjective fluency. And we predict that Whittlesea's findings concerning discrepancy, surprising (relative) fluency and disfluency, and the effects of a pause will also apply to valuation and causation judgements.

Notes

1. Alter and Oppenheimer (2009) provide a useful review of the wide range of tasks in which fluency has been found to play a role. It seems to be a central variable relating ease of processing in many cognitive tasks (perceptual, linguistic-phonological, linguistic-comprehension, spatial reasoning, reasoning) to confidence, liking, frequency, and several other aspects of judgement. As Alter and Oppenheimer point out, reflecting a Whittlesea's conclusion, it is fluency *plus* an interpretive theory that connects the processing experience to inferential conclusions.
2. Signal Detection models for recognition memory judgements also imply that the difference between experienced and expected familiarity, not absolute familiarity, determines recognition judgements (and confidence in those judgements). According to that model, it is the likelihood *ratio* between the fluency that would be expected from a non-studied item (a sample from the 'noise distribution') and the fluency experienced from the to-be-judged item (sometimes an 'old item' producing a sample from the 'signal + noise distribution'; see Benjamin et al., 1998, for a systematic development of this point). In most applications, the signal detection model assumes that the participant in an experiment has a *general* sense of the distributions of noise-only and signal+noise familiarity experiences. But, Whittlesea favoured the notion of *stimulus-specific* 'norms on the fly' as the source of expectations.

16

Your Effort Is Showing! Pupil Dilation Reveals Memory Heuristics

Megan H. Papesh and Stephen D. Goldinger

Introduction

It is an honour to contribute to a collection of essays celebrating Bruce Whittlesea's career. The research and ideas from Whittlesea and his colleagues have heavily influenced much of the research in our laboratory, particularly our studies of face perception and memory. Although face processing is often considered 'modular' (i.e., highly specialized in neural and computational terms; Haxby et al., 2000), we have consistently observed that judgements of face memory are affected by the evaluative and heuristic processes that Whittlesea has hypothesized (e.g., Whittlesea & Leboe, 2000). In this chapter, we briefly review several prior findings that connect Whittlesea's (1997) SCAPE framework with face memory. We then describe new results, wherein we hypothesize that long-term struggles from the SCAPE evaluation system may inspire a new heuristic (kindly dubbed the *'oh…screw it'* heuristic).

As Whittlesea has argued, in any memory test, people are influenced by multiple sources of information. Consider a study by Jacoby and Whitehouse (1989), wherein people were shown words for memorization. In a later recognition test, words were preceded by identity primes that were either masked (subliminal), or were clearly shown. Given subliminal matching primes, the frequency of 'old' responses, especially false alarms, increased. Conversely, with overt matching primes, 'old' responses decreased. Jacoby and Whitehouse proposed that, with subliminal primes, people interpret enhanced perceptual fluency as familiarity. With overt primes, people discount 'positive' signals (either fluency or true familiarity) as a natural by-product of the priming words.

Building upon such findings, Whittlesea and Leboe (2000; Whittlesea & Williams, 1998; 2001) suggested that recognition entails two stages: First is *production* of mental states, wherein images or ideas are brought to mind. Production may follow perceptual input, such as a face, which the mind immediately elaborates (Neisser, 1967), perhaps with a name. Following production, the second stage is *evaluation*. This is not direct stimulus evaluation, such as deciding whether a recognition target exceeds criterion. Instead, Whittlesea (1997; Leboe & Whittlesea, 2002) suggested that people automatically evaluate their own production functions, keeping a running index of the relative harmony of mind. By its nature, evaluation is based on subjective states of mind. For example, imagine encountering a co-worker, whom you easily recognize. On most days, such an encounter will produce immediate recognition, and the evaluation process will not be unduly aroused, creating no particular feelings. Imagine, however, that you encounter the co-worker after he has shaved his long-standing mustache. The production process easily recognizes your acquaintance, but with an uncomfortable dysfluency – something is different and weird. This momentary processing hitch motivates a careful search, as you try to figure out 'what's different?'

According to SCAPE, such feelings arise from a *discrepancy-attribution* process (Whittlesea & Williams, 2001). Depending upon context, people have different implicit expectations of processing fluency. When those expectations are violated, an evaluation 'flag' is raised, automatically triggering a search for some explanation. Sometimes, the context itself provides a natural attribution – in Jacoby and Whitehouse (1989), inexplicable perceptual fluency evoked feelings of familiarity. In a challenging recognition test, momentary boosts in fluency should evoke the obvious attribution of prior experience. Counter to intuition, when processing expectations are violated, people experience feelings of familiarity. Despite standard usage, truly familiar stimuli generally evoke no feelings of memory. Conversely, mildly familiar stimuli *('what's different about this guy?')* create a strong, nagging sense of familiarity.

In a prior study, Goldinger and Hansen (2005) tested this framework by presenting people with subtle bodily cues that could be experienced as 'familiarity signals.' They fitted a chair with wireless speakers and participants (sitting in that very chair) memorized words, pictures, and faces (in separate blocks). During recognition, half the test items (old and new) were presented with a simultaneous, subliminal buzz (a low-amplitude, 60-Hz sinewave). In a control condition, the buzz was easy to perceive. People made 'old-new' decisions and confidence ratings. To help contrast recollection and familiarity, Goldinger and Hansen presented items that were relatively 'easy' and 'hard' to remember, such as photos of celebrities and medical students, respectively. When faces were hard to recall, there were clear effects – given the subliminal buzz, people were more likely to respond 'old,' increasing hits and false alarms. In the confidence data, when people committed false alarms, the subliminal buzz elicited relatively high confidence. Given hits, the subliminal buzz had the opposite effect, *reducing* confidence. This seemed to reflect the sense of familiarity – when a person really has no memory (for new faces), the buzz created a 'tingle' of familiarity and people responded 'old' with confidence. But, given true memory (for old faces), the buzz created a tingle of doubt. These findings followed predictions from SCAPE, as the same signal created different memorial interpretations, based on context (Whittlesea & Williams, 2001).

Heuristics in face recognition

In SCAPE, the central hypothesis is that memory decisions are often evaluative in nature, as people interpret familiarity cues in context. Lacking absolute criteria for recognition, people rely on memory decision heuristics. Whittlesea and Leboe (2000) described three heuristics, called *generation*, *resemblance*, and *fluency*. In a study of face perception, Kleider and Goldinger (2004) tested the fluency heuristic, as described by Jacoby and Dallas (1981; Jacoby et al., 1989). According to the fluency heuristic, in challenging recognition tasks, people often use the fluency (ease) of perceptual processing as a memory cue. Many data suggest that perceptual processing is enhanced for familiar stimuli: People seem to implicitly assume this relationship, such that 'memory illusions' can be elicited by increasing perceptual fluency. As perception is made more fluent, feelings of familiarity arise, leading to increased 'old' judgements (a liberal criterion shift). Although this effect occurs among old items, it is generally larger for new items, because familiarity is their only available cue.

Although face processing is typically robust (Farah et al., 1998), Kleider and Goldinger (2004) predicted that fluency manipulations in face perception would affect memory, just as Whittlesea and Leboe (2000) observed with verbal materials. In a series of five experiments, people briefly studied clear faces, followed by a distracter task and a test phase. During the recognition tests, new and old faces were embedded in varying levels of visual white noise. In every experiment, 'old' responses increased to clear faces, driven mainly by liberal bias shifts. This pattern occurred regardless of variations in noise levels, warnings to participants that clarity was not related to 'old-new' status, and other factors. In two final experiments, Kleider and Goldinger reversed the pattern: People perceived familiar photos as having greater clarity in noise, and longer presentation durations in a speeded perceptual task. As predicted by SCAPE, increased perceptual fluency created feelings of familiarity, and familiarity created feelings of fluency.

According to Whittlesea and Leboe (2000), memory heuristics fall into two general categories, *quality-of-processing* and *information*. Quality-of-processing heuristics are based on the speed, coherence, and vividness of production – perceptual fluency falls under this domain. Information heuristics are based on contextual or statistical cues. The *generation* heuristic is based on retrieval of episodic details, mentally 'placing' stimuli in decision-relevant contexts. When contextual details are highly available, people feel more confident about memory. The *resemblance* heuristic is based on the 'fit' between a stimulus under consideration and relevant prior experiences, as when people falsely recognize words that are consistent with a prototype (Roediger & McDermott, 1995).

Moving beyond the fluency heuristic, Kleider and Goldinger (2006) investigated the generation and resemblance heuristics, with special attention to the *other-race effect* (ORE) in face recognition. Although people are extraordinary face processors (Bahrick et al., 1975), research has consistently shown that people are better able to distinguish among faces from their own race, relative to members of other races (Meissner & Brigham, 2001). For several reasons (see Whittlesea & Leboe, 2000), the generation and resemblance heuristics are difficult to empirically dissociate. Therefore, Kleider and Goldinger attempted to manipulate the *likely* influence of each heuristic across experiments. Participants were asked to study two series of faces – across conditions, the face sets either combined White and Asian faces, or these subsets were presented in separate lists. In the later recognition task, people were asked to respond 'old' only to faces from the second study list. In mixed conditions, people relied heavily on generation (this was indicated by the relative lack of a 'resemblance pattern,' as people seemed to carefully generate prior list contexts for each face). In the blocked conditions, they relied heavily on resemblance (as evidenced by false-alarm patterns). Once those endpoint conditions were recorded, Kleider and Goldinger explored the middle ground, 'nudging' people back and forth between heuristics by changing the statistical benefits of either strategy.

Eye-tracking and pupillometry: are the eyes a window on memory processes?

As the foregoing review implies, the basic predictions of SCAPE are easily tested (and have been generally verified) in the domain of face memory. As a rule, virtually all such studies have focused on heuristic processes arising during memory *testing*. It seems completely plausible, however, that people might also adopt heuristic information-processing strategies during *learning*, in preparation for upcoming memory tests. Recall that SCAPE is based on two processes: First is the *production* of mental states, such as seeing a face and trying to create a 'strong' memory trace. The second stage is *evaluation* of production efficiency, such as a person feeling confident that she is successfully memorizing the faces. In the case of cross-race faces, however, it seems very likely that self-evaluation will result in decreased confidence, and feelings that future memory will be poor. As such, monitoring behaviour (overt and unconscious) during learning trials should provide a window into the evaluation stage of SCAPE.

We have recently begun to examine cognitive phenomena, and the second stage of SCAPE, via psychophysiological indices, specifically the pupillary reflex. It has long been known that, when people perform challenging tasks that require more cognitive effort, their pupils dilate (Porter et al., 2007), possibly representing a summed index of brain activity associated with such tasks (Beatty & Kahneman, 1966). Beatty (1982) summarized the attractive qualities of the pupil reflex that made it Kahneman's (1973) primary index of mental processing load in his theory of attention allocation. As noted by Beatty (1982), Kahneman discussed three criteria which should be met by physiological indicators of mental processing load:

> It should be sensitive to within-task variations in task demands produced by changes in task parameters; it should reflect between-task differences in processing load elicited by qualitatively different cognitive operations; finally, it should capture between-individual differences in processing load as individuals of different abilities perform a fixed set of cognitive operations. (p. 276)

Beatty (1982) reviewed the evidence and concluded that the pupillary reflex satisfies all three of Kahneman's criteria. In our research, we propose that, when used in concert with the appropriate behavioural indices, pupil dilation can serve as a sensitive indicator of underlying cognitive effort, particularly the evaluation of production efficiency, as in SCAPE. The appeal of this often-overlooked dependent measure lies in its independence – the pupil reflex is not subject to task-specific strategies; it is a bias-free measure of processing load that can estimate mental effort expended in a variety of tasks. Next, we discuss several experiments in which we have successfully applied pupillometry to the study of cognitive processing. To preface, across several domains, including face perception, memory, and word naming, we find tight connections between mental effort and pupillary reflexes.

Before turning to a discussion of our research, we must note that, although pupillometry is a sensitive measure of cognitive effort, pupils also change reflexively, based on visual input. Variations in luminance across stimuli, sudden onsets of stimuli, and variations in colour can all induce pupillary responses (Porter et al., 2007), necessitating tight experimental control. Porter and Troscianko (2003) discussed several methodological approaches that can minimize unwanted pupil reflexes. These include using relatively low stimulus contrast, avoiding coloured stimuli, and using relatively long exposure durations. The use of long exposure durations is particularly germane, as previous research has indicated that task-evoked pupil dilations are a relatively late-arriving index of cognitive effort, beginning several hundred milliseconds following stimulus onset (Kuchinke et al., 2007). In the experiments reported below, we applied a combination of methods to minimize the influence of visually-influenced pupillary reflexes.

Our first investigation examining the nature of the relationship between the pupillary reflex and cognitive effort was another study of the own-race effect (ORE) in face recognition. A number of theories have been proposed to account for the ORE (Levin, 1996; 2000; Maclin & Malpass, 2001; Ng & Lindsay, 1994; Sporer, 2001), but the mechanism by which it occurs is often neglected. We were interested, therefore, in the process by which faces are learned for future recognition. Are there observable encoding differences (across own- and cross-race faces) that could predict future recognition accuracy? That is, can we pinpoint the evaluation stage by examining study trials? We hypothesized that, using eye movements (cf. Henderson et al., 2005) and pupillometry, we could index the amount of effort expended during learning and relate this to behavioural performance on a recognition memory test.

To investigate the relationship between effort during learning and later recognition accuracy, we presented a group of Asian and Caucasian participants with a set standardized, neutral expression photographs from Ekman and Matsumoto (1993). Critically, all faces were presented in greyscale and for relatively long (5 versus 10 seconds) periods of time. In brief, we found that participants (in both racial groups) who demonstrated an ORE in recognition accuracy selectively *withdrew effort* during the encoding of other-race faces, reflected by both decreased eye movements and pupil dilations (see Goldinger et al., 2009, for a full account). Interestingly, we observed a pattern that we compared to 'learned helplessness,' wherein low-scoring participants' physiological indices of effort steadily declined across trials, essentially indicating that they had succumbed to the newly dubbed 'oh ... screw it*'* heuristic.

To investigate, and potentially increase, this effort reduction pattern, we conducted a second experiment, using only Caucasian participants and 10 s exposure durations. Our goal in this experiment was to assess pupil dilation to the exact same faces, but to increase the *perceived difficulty* of the study task by changing the study-list context. Specifically, we added photographs (also from Ekman & Matsumoto, 1993) to the study session, showing the same Asian and Caucasian faces, now with a mixture of different emotional expressions. We expected that, given a longer study phase including emotional faces, people would find the neutral faces progressively less distinctive and interesting. Following SCAPE, we predicted that such perceived difficulty would affect the evaluation stage, such that people would reduce confidence in future memory and selectively withdraw effort from the Asian neutral faces.

Twenty Caucasian students from Arizona State University with normal or corrected-to-normal vision participated for partial course credit. The study stimuli consisted of 52 faces, with equal representation across all variables (sex, emotionality, and race) and no repeated models. All pictures were set to equal mean luminance and were embedded in a black background (1024 x 768) for presentation on a Tobii 1750 17-inch monitor. A chin rest maintained the participants' viewing distance at 60 cm and both eyes were continuously tracked at 50 Hz throughout the experiment. Participants were given 10 s to study each face and were later tested on their recognition memory for the photographs using only neutral expression models.

As in our previous experiment, we observed a robust ORE in recognition accuracy, with higher Pr[1] scores to Caucasian faces, $F(1, 19) = 46.06$, $MSe = .04$, $\eta_p^2 = .45$, and clear differences in eye movement patterns depending on the race of the stimulus face, with participants moving their eyes greater distances across Caucasian faces, $F(1, 19) = 145.8$, $MSe = 44.16$, $\eta_p^2 = .52$. More relevant to SCAPE were the pupil dilation analyses. Pupil diameters were measured between trials (and while the fixation cross was shown) to establish baseline estimates, then during photograph viewing for comparison. We removed missing observations due to blinks or signal loss, filing those gaps by linear interpolation.[2] Another 0.4% observations were replaced, in the same manner, for values falling more than 2.5 standard deviations from their 10 immediate neighbours. For each participant, we selected the 'better' eye (i.e., with fewer corrected observations) for analysis. In an overall analysis of the experiment, we observed significantly greater dilation to Asian faces, relative to Caucasian faces, $F(1, 19) = 185.13$, $MSe = 39.99$, $\eta_p^2 = .67$, suggesting that Asian faces required greater effort during encoding.

Additional analyses on the eye movement data from this experiment demonstrated that, over the course of the experiment, both high- and low-scoring participants exerted diminishing effort to the encoding of Asian faces. We previously found this pattern in only low-scoring participants. Similar patterns were observed for pupil dilations. As shown in Figure 16.1, average dilation to Caucasian faces was statistically equivalent across both high- and low-scorers throughout the entire 13-trial encoding period. By contrast, average dilation to the Asian faces declined across trials, with a greater decline among the low-scoring participants. In an omnibus 2 (Group: high/low) x 2 (Race: Asian/Caucasian) x 13 (Trial Number) ANOVA, we observed a large main effect of Race, $F(1, 19) = 159.6$, $MSe = 42.08$, $\eta_p^2 =.63$, with greater overall dilation to Asian faces, relative to Caucasian faces. In one-way (Trial Number) ANOVAs on the Asian learning trials for each group, we observed a main effect of Trial Number for both the high-scoring group, $F(1, 9) = 8.47$, $MSe = 41.50$, $\eta_p^2 =.10$, and the low-scoring group, $F(1, 9) = 23.83$, $MSe = 47.18$, $\eta_p^2 =.19$. Although the 'oh … screw it' pattern was stronger in the low-scoring group, it was now observed in both groups.

Our data suggest that cross-race faces demand extra encoding effort, and that this extra effort manifests itself in differences in pupil diameter and patterns of eye movements. For the present purposes, we focus on the observed differences in pupil dilation, as we believe that they accurately index underlying cognitive effort and predict future memory. Using pupillometry, we are able to investigate the evaluation stage of Whittlesea's (1997) SCAPE. Whereas the preceding experiment focused on mental effort expended during encoding, and how this effort predicts eventual memory performance, our next experiment was an investigation into the mental effort expended

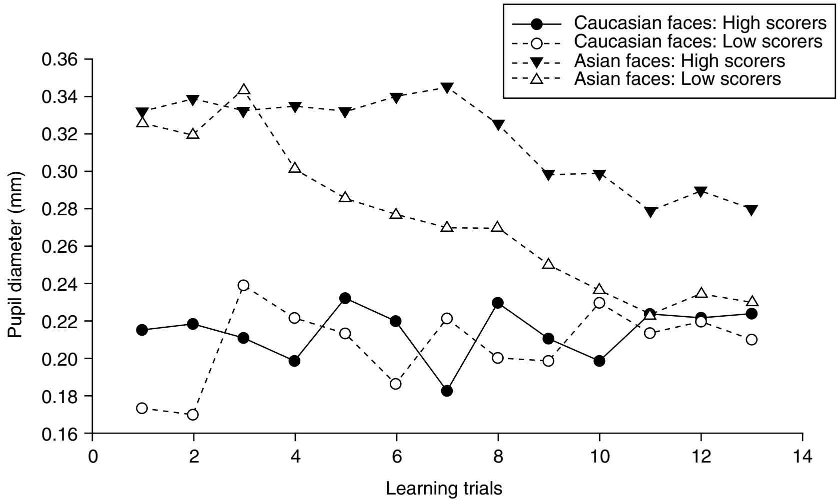

Figure 16.1 Pupil dilation across learning trials. Separate functions are shown for learning of Asian and Caucasian faces, and for participants with relatively high or low memory scores

during recognition as a function of effort expended during encoding. Specifically, we examined recognition memory for spoken words, comparing pupil diameters to the same stimuli across study and test. To preface, we observed that, during accurate recognition trials, pupil diameters increased, relative to average diameters during encoding, and that this difference was greatest for more difficult stimuli (i.e., low frequency words).

As in our ORE experiment, we hypothesized that memory could be predicted or reflected in pupil dilations, and that pupil diameters would differ depending on stimulus difficulty. To investigate this, we used words as stimuli, so that we could more precisely control inherent stimulus qualities (e.g., syllables, length, word frequency, etc.). To examine the pupillary reflex in response to words, we first examined basic word frequency effects in an experiment using a modified delayed naming procedure (see Papesh & Goldinger, 2009).

Thirty native English speakers from Arizona State University, with normal or corrected-to-normal vision, participated in exchange for partial course credit. Stimuli consisted of 150 common/uncommon word pairs (i.e., high frequency, HF/low frequency, LF), which were matched for visual features (e.g., *few/pew*). Words were presented onscreen for 500 ms each (black font on gray background), followed by a variable delay period of 250–2000 ms before a response tone. In the majority of trials (240), the response tone indicated to participants that they were to speak the word. In a minority of trials, however, the response tone signalled that participants were to abandon the speech plan and say 'blah' instead. This secondary task allowed us to examine the effects of word frequency when production was equated. In a previous experiment of this type, Goldinger, Azuma, Abramson, and Jain (1997) observed that participants were faster to read HF words aloud, relative to LF words. We did not observe this effect. Instead, we observed that, at all trial stages following word perception, participants' pupils were more dilated to LF words, relative to HF words. That is, pupil dilations were a more sensitive reflection of underlying cognitive effort, relative to the behavioural index (naming latency).

To extend this finding, and to examine the pupillary reflex during a recognition memory task, we recently compared memory for HF and LF *spoken* words, eliminating the visual input entirely. The participants ($N = 20$) were all native speakers of English with no known hearing disabilities,[3] all of whom participated in exchange for partial course credit. After being familiarized with the task, participants were asked to focus their gaze anywhere on the 17-inch Tobii 1750 monitor (with a solid background of gray, at a constant RGB of 150), so that continuous measurements could be taken of their pupil diameters. Before the study task, baseline diameter estimates were obtained for every participant by instructing them to passively listen to a series of 12 words and sentences. The programme automatically calculated a 'baseline range' for each person, defined as any diameter falling within 2.5 standard deviations of their baseline stage diameter. During the study and test trials that followed, participants' pupil diameters were required to return to this range within 6 s, or the succeeding trial was dropped from analysis.

During study trials, participants passively listened to a series of 32 HF and LF words, with the timing of trials controlled by the speed with which each person's pupil diameters returned to the baseline range. Following a 3-minute distraction task (playing a computer game), participants listened to 64 spoken words, half of which were old, and indicated whether each word was 'old' or 'new.' The voice of the speaker was constant across study and test.

Pupil dilations were trimmed for blinks and sorted into 'trial events' corresponding to the fixation cross, the spoken word, and the blank screen (wherein the participants' pupils returned to the baseline range). We analysed the data in a 2 (Word Frequency: HF/LF) x 2 (Presentation: first/second) x 2 (Accuracy: hit/miss) x 3 (Trial Event: fixation/word/blank) RM ANOVA. Although the Accuracy x Presentation interaction was not statistically significant ($F < 1$), a priori pair-wise comparisons indicated that, during accurate recognition trials, pupils dilated to a greater extent during the second presentation (4.39 mm), relative to the first (4.24 mm), $F(1, 16) = 4.85$, $\eta_p^2 = .23$. Although this trend was evident for both HF and LF words, further pair-wise comparisons indicated that this effect was driven primarily by LF words. When participants correctly recognized LF words, their pupils dilated to a greater extent during the second presentation (4.38 mm), relative to the first (4.23 mm), $F(1, 16) = 6.02$, $\eta_p^2 = .27$. Thus, as in our previous experiments, pupil dilation revealed underlying memory processing. In this particular case, people expended greater effort when making correct recognition decisions, relative to errors. In ongoing research, we plan to compare experimental situations that invite more or less 'heuristic processing,' expecting that pupil dilation will be a sensitive predictor of memory performance, especially when people must generate prior contexts for items.

Conclusion: production and evaluation as 'stream of consciousness'

Across all our studies, the heuristic processes predicted by SCAPE have been observed. Extending prior studies, our new results suggest that people engage in production and evaluation continuously, rather than engaging such processes only during a memory test. Moreover, our results show that these processes are accurately indexed by psychophysiological measures. The pupillary reflex reflects effort during both learning and recognition, which has promise for differentiating among theories of memory and assigns a concrete, dependent measure to the hypothesized stages of SCAPE. We believe the pupil reflex is sensitive enough to reflect fine gradations in the quality of memory and can indicate the use, or non-use, of various heuristic processes.

Notes

1. Pr is a common accuracy score, representing the difference between hits and false alarms (see Feenan & Snodgrass, 1990).
2. This resulted in less than 4.5% of data repair for all participants.
3. This was inferred from accurate performance on a tone identification task prior to the experiment.

17
Remembering Under the Influence of Unconscious Expectations

Bruce W. A. Whittlesea

Introduction

The editors of this volume asked all us contributors to say something nice about the honouree. Okay: he writes long, convoluted articles packed with experiments (probably some kind of compensation for his shortness); he does have a nice 'conference' voice (resulting from liberal application of scotch to his tonsils); and he may be nearly as smart as he thinks he is. Best I can do.

The most obvious fact about remembering is that exact repetition of a novel stimulus compound, such as 'GREEN – TIGER', often enables people to later recognize that compound, or to regenerate one part ('TIGER') when cued by the other ('GREEN'). One way to think about this commonplace phenomenon is in terms of the notion of re-activation: the attention paid to the original event causes the encoding of an effective representation, and the features of the repeated event serve as cues to retrieve that representation, thus enabling the person to become aware of its previous occurrence. By such an account, the re-activation of the earlier representation is coextensive with the act of remembering: partial or weak re-activation of that representation will cause a feeling of familiarity, and more effective activation will cause actual recall of the event.

However, there are other acts of remembrance that cannot be easily explained in those terms. One occurs when the recognition targets consist of a sentence stem plus a sensible terminal word, with high- versus low-constraint contexts serving as the stem. High-constraint stems (e.g., 'After the accident he was covered in') can be sensibly completed only with a small number of words (e.g., BLOOD, GLASS, perhaps PAINT or SHAME), whereas low-constraint stems (e.g., 'On the corner of the table there was a bit of') can be sensibly completed with many more words. Whittlesea (2002, 2004) presented sentences with high- and low-constraint stems in a study phase; in a later phase, subjects were shown the same sentences or alternatively stems completed with words taken from a different sentence, but which made as much sense as the original word (e.g., substituting BLOOD for GLASS after the subject had studied 'After the accident he was covered in GLASS'). This procedure produces a highly reliable pattern of effect, consisting of approximately equal hits for targets presented with either kind of stem, but lower false alarms for the low-constraint stems.

This asymmetric pattern is difficult to explain through any account of memory that is based only on assumptions about encoding and retrieval. For example, an SDT account of memory would have to assume different criteria and different underlying distributions of familiarity to explain the effects caused by high- versus low-constraint of stems; however, it would have difficulty explaining why the two kinds of stem cause such differences in underlying distributions and why those distributions are so arranged as to cause differences only in false alarms. That difficulty is compounded by observations that the same pattern of effect is caused by high-constraint stems terminated by congruous versus incongruous words (e.g., 'After the accident he was covered in BLOOD/GEESE'; Whittlesea, 2004).

An alternate way to think of how people become aware of their personal past is provided by accounts of memory based on the ideas of construction and attribution, such as the SCAPE framework (e.g., Whittlesea, 2002). By that account, the act of remembering consists of two quasi-independent components: the occurrence of a mental content, controlled by the interaction between current cues and representations of prior experiences, and the occurrence of a subjective feeling of remembering, resulting from a process of evaluation and inference that takes into account those aspects of the stimulus, the context and the quality of current processing that are salient to the person. This decision process results in the person adopting an attitude towards their current mental experience of a stimulus, variously regarding its integrity, meaning, significance, affective quality, or source: which of these attitudes the person adopts depends not just on the objective historical status of the stimulus, but also on the person's current intentions in dealing with the stimulus in the current context. The subjective feeling of remembrance occurs when the person is led by circumstances to make an attribution to a source in the past. That is, awareness of the relationship between current and past events is constructed, not apprehended.

The idea that awareness of the past is the result of a constructive process implies that it is open to systematic errors of omission and commission, caused by providing or drawing attention to aspects of processing which normally would be valid indicators of prior experience, but which are misleading in the context. A great deal of research has been focused on one potential basis of decision, the fluency heuristic (Jacoby & Dallas, 1981). However, while it is clear that people can and do use the fluency heuristic under some circumstances, some systematic patterns of illusory and accurate remembering appear to require a more complex explanation, invoking the concepts of expectation and resolution in addition to evaluation and attribution (Whittlesea & Leboe, 2003). There are a number of observations suggesting that people often do not base recognition decisions on any single processing characteristic, but instead compare the quality or content of their performance against an expectation aroused by concurrent or foregoing processing (e.g., Whittlesea & Williams, 2001). In those studies, validation or violation of these expectations by further processing of the stimulus apparently caused people to adopt an attitude towards their processing, resulting in a subjective experience of remembering or non-remembering.

Whittlesea (2002, 2004) used the concepts of expectation, resolution and attribution to interpret the asymmetric data pattern reported earlier (approximately equal hits for targets presented with either high- or low-constraint stem, but lower false alarms for the low-constraint stems). I proposed that experience of the sentences in training created representations that could serve as expectations when the sentences were seen again in test. The idea was that the pairing of particular stems with particular termination words in training would create an overall thematic understanding, or schema, of the sentence. When encountered again in test, the stems would arouse this schema, creating an expectation about the terminal word that should complete those stems. The difference in effect of high versus low constraint of stems occurs because of the differential impact of such stems in constructing an overall understanding of the sentences on their first presentation. The meaning of a sentence containing a high-constraint stem is given primarily by the stem. For example, the sentence 'After the accident he was covered in glass' has much the same overall meaning as 'After the accident he was covered in blood.' The common theme is the severity of the accident. When the stem is repeated in test, it arouses the schema formed on the earlier occasion, which now serves as an expectation about how that stem should be completed. However, because this theme is not specifically defined by the terminal word, that word could sometimes be replaced in test (as in 'BLOOD' for 'GLASS' in the example above) without changing the perceived meaning of the overall sentence, resulting in a false claim of recognition. In contrast, the meaning of a sentence containing a low-constraint stem is given primarily by its terminal word. For example, 'On the corner of the table there was a bit of glass' suggests a minor household accident, whereas the same stem completed with 'blood' suggests something more sinister, perhaps murder. In this case, the overall meaning of the sentence is given by the terminal word. In consequence, changing the last word of such sentences in test is more likely to cause the subject to experience a violation of expectation, thereby enabling easy detection of the substitution and producing fewer false alarms. Either type of expectation (relatively general or specific) would be validated by re-presentation of the original terminal word, resulting in approximately equal hits in the two cases.

This understanding of the effects of high versus low constraint of stem is supported by the results of parallel tests of remembering using recall instead of recognition. Whittlesea (2002, Experiment 5) presented both whole sentences and single words in a training phase; in this phase, one member of each critical pair of words (e.g., BLOOD and GLASS) was presented in isolation, the other as the terminal word of either a low- or high-constraint sentence. In the test phase, the stems that had been presented earlier were presented as retrieval cues; the subjects were required to complete the stems with a sensible word, using the word that had been presented with that stem during the training phase if possible. One major result of this study was that high-constraint stems caused the subjects to report the actual word presented with a stem on about 50% of trials, whereas low-constraint stems supported report of that word on only 25% of trials. The second major result was that high-constraint stems caused the subjects to report the alternate word (the other equally sensible termination that was presented in training without a stem) on 23% of trials, whereas low-constraint stems caused the subjects to report the alternate word on only 3% of trials. That is, high-constraint stems caused more recall of words originally presented with those stems, but also caused recall of words not presented with that stem, whereas low-constraint stems caused less recall of training words, but recall that was much more specific to the original encoding episodes.

Interpreting these recall results within the same assumptions as the recognition data, I suggested that presentation of the high-constraint stems in test often engaged the schema formed for those stems in the training phase, resulting in the person completing the stem with a word consistent with the meaning of that schema. However, because schemas based on high-constraint stems are relatively general, the subjects often substituted the alternate word from a pair. Reporting that word was a false alarm, because it had been seen in isolation in training, not with that stem. However, that word fits the global meaning of the schema as well as the word that should have been reported. In contrast, recall of terminal words for low-constraint stems presented in test was much more specific: if the subjects reported a word from the training phase, it was almost always the word actually presented with that stem earlier, rather than the equally sensible alternate word. That again supports the idea that the schemas created in training for sentences with low-constraint stems are relatively specific with respect to the identity of the terminal word. However, the much lower rate of recalling a word from the training phase when cued with a low-constraint stem compared to a high-constraint stem stands in stark contrast to the equality of hits observed in those conditions in recognition, as described earlier. The major difference in procedure leading to this difference in pattern of response is, of course, that candidate terminal words were actually presented in test in the recognition studies, but not in the recall studies. That leads to a further interesting speculation: that both recognition and recall were performed using expectations acquired in training, but that such expectations are not as often re-aroused in test by low-constraint stems when only the stem is presented as a cue. That implies that the expectations formed with low-constraint stems, which are just as useful as expectations created for high-constraint stems for producing hits in recognition and even more useful for rejecting novel terminal words, remain latent when the terminal word is not presented as a candidate. That is, the terminal word of low-constraint sentences is not only responsible for determining the overall meaning of the sentence during the initial exposure, but also appears to be responsible for re-engaging that meaning and expectation on a subsequent occasion. That idea has broad implications for the ability of people to remember different kinds of detail, as discussed later.

I have argued previously that the surprising resolution of *indefinite expectations* causes a perception of discrepancy.[1] That perception itself is unconscious; people lack metaknowledge of both the expectation and the occurrence of resolution. However, the perception of discrepancy motivates an attribution process that constructs conscious feelings about the stimulus (pleasantness or goodness), the self (mood or skill) or the person's past history (feelings of familiarity). The current project was designed to investigate instead *definite, general* and *definite, specific* expectations: whether people have conscious access to the validation or violation of such expectations, or whether those processes also occur unconsciously, causing the person to experience feelings of rightness or wrongness, satisfaction or unease, and familiarity or unfamiliarity without knowing why. To get at this issue, I asked subjects to describe their metaknowledge before or after attempting a recognition test, using the question: 'Will it be/was it hard or easy to perform the recognition judgement on that item?' The idea was that if people have conscious access to the source of their remembering phenomenology, then factors that promote accurate recognition should also enable the person to decide that the item was easy to judge. Specifically, recognition decisions performed on low-constraint novel test sentences should be accompanied by more claims of 'Easy' than high-constraint novel test sentences. In contrast, if people have to infer their metaknowledge from observation of their behaviour, as an outside observer would have to do, then there might not be a strong relationship between factors influencing accuracy of performance and feelings of ease.

General method

Twenty-six Simon Fraser University students were tested in Experiment 1 and 27 in Experiment 2. The experiments used the 60 frames presented in the Appendix of Whittlesea (2002). Each frame consists of a pair of stems (e.g., 'The prisoner screamed insults at the…' and 'He didn't really need to have another…') and a pair of terminal words (e.g., JUDGE/GUARD), either of which can sensibly complete either stem. In each pair, one of the stems could be sensibly completed only by a small number of words (high-constraint stem) and one could be completed with many words (low-constraint stem).

In the training phase of both studies, subjects were shown all 120 stems (60 constraining and 60 unconstraining), each completed (at random) by one or the other of the two terminal words for that set (the other word being used to complete the other stem). These sentences were presented in a random order, re-randomized for each subject, under the constraint that the two sentences from a frame were not presented successively. Subjects read each sentence aloud, striking a key when finished to cue the next sentence.

At test, half of the sentences containing constraining stems and half of the sentences containing unconstraining stems were presented unchanged. These sentences were selected in pairs, at random, such that if the sentence completed by GUARD was not changed then neither was the sentence containing JUDGE. The remaining sentences were altered, such that the terminal word was interchanged between members of a pair (e.g., interchanging GUARD and JUDGE). The resulting sentences were still sensible, but literally new (e.g., 'The prisoner screamed insults at the guard' changed to 'The prisoner screamed insults at the judge' and 'He didn't really need to have another judge' changed to 'He didn't really need to have another guard'). Sentences from all four conditions (constrained/unconstrained X new/old) were presented in a random order, re-randomized for each subject, again under the constraint that two sentences from a pair were not presented successively. Subjects read each sentence aloud and then made a recognition decision about the sentence as a whole.

The subjects were introduced to the ease-of-remembering judgement with several examples, using pairs of sentences with high- and low-constraint stems (e. g., 'The stormy seas tossed the BOAT' and 'She saved her money and bought a YACHT'). They were informed that test sentences would either be identical to training sentences, or altered by re-pairing terminal words with stems; they were shown examples of how that would produce novel but still sensible sentences (e.g., 'The stormy seas tossed the YACHT' and 'She saved her money and bought a BOAT').

In Experiment 1, the ease-of-remembering judgement preceded the recognition judgement. On each test trial, a stem was exposed, without the terminal word; the subjects were asked: 'How hard will it be to remember the exact sentence?' In making this judgement, they were provided with two keys, marked 'HARD' and 'EASY.' Subjects in Experiment 2 made the same decision, but did so after being shown both the stem and terminal word and making their recognition decision for that sentence. The question was modified to reflect the change in sequence of tasks, and now read 'How hard was it to remember the exact sentence?'

Recognition results

I will describe the results of the recognition tests in both studies now, and then go on to the more interesting ease-of-remembering data. As can be seen in the upper panels of Tables 17.1 and 17.2, the asymmetric 'signature' pattern described earlier (approximately equal hits but more false alarms for high-constraint stems) was observed in both studies, regardless of when the ease-of-remembering question was asked. Constraint of the stem had no reliable effect on decisions about old sentences in either study, $F(1,25) = 0.60$, $MSe = .01$, $p = .446$, $\text{eta}^2 = .06$ and $F(1,26) = 0.97$, $MSe = .01$, $p = .334$, $\text{eta}^2 = .03$ for Experiments 1 and 2, respectively. In contrast, the subjects correctly rejected re-paired sentences about 10% more often when the stem was unconstraining than when it was, $F(1,25) = 17.31$, $MSe = .01$, $p < .001$, $\text{eta}^2 = .41$ and $F(1,26) = 17.37$, $MSe = .01$, $p < .001$, $\text{eta}^2 = .40$ for Experiments 1 and 2, respectively. These results closely replicated the pattern observed for these stimuli by Whittlesea (2002, 2004) and demonstrate the robustness of the effect. Again, I interpret these data to suggest that, during the study phase, the subjects acquired the two kinds of expectation (*definite, general* versus *definite, specific)* described earlier, and that those expectations controlled their recognition decisions. The next question was whether the subjects were, or could learn to be, aware of this basis of their recognition decisions.

Experiment 1: ease-of-remembering judgements prior to recognition

The subjects judged sentences with high-constraint stems to be 'easy to remember later' 22% more often than sentences with low-constraint stems, $F(1,25) = 106.89$, $MSe = .02$, $p < .001$, $\text{eta}^2 = .81$ (Table 17.1, lower panel).

That discrimination indicates that they could clearly feel the difference between the two types of sentence, whether or not they were aware of what the difference was, or even aware that there were two types. However, their predictions about how they would do in the recognition test were backwards: they thought they would do better on high-constraint sentences (however

Table 17.1 Recognition and ease-of-remembering data for Experiment 1

Recognition: Hits and correct rejections		
Constraint of stem	**Original Sentences**	**Re-paired Sentences**
High	.73	.73
Low	.75	.84
Confidence of remembering		
Constraint of stem	**p (claim 'easy')**	
High	.67	
Low	.45	

they thought of them), whereas, as just reported, they actually did better on low-constraint items. Instead, the pattern of the prediction data matches what is observed in a pure recall test, as reported earlier. It seems likely that subjects in the current study, given only a stem at the beginning of each trial, performed recall as a way to guess what memorial resources they had that would assist them in recognition a moment later. However, unlike the subjects in that earlier study, on every trial these subjects were then presented with the entire sentence and required to perform recognition; and on that portion of the trial, they received feedback on their prediction in the form of an opportunity to experience the ease of performing the recognition decision. The dissociation between ease-of-remembering judgements and recognition decisions thus means that in making the former decision, they were wholly insensitive to the differential effect that the two kinds of sentence had had on their remembering judgements on previous trials. Thus I concluded that the expectation which would control the success of recognition a moment later remained latent, inaccessible to consciousness even in the form of a subjective feeling, before the terminal word was presented. The remaining question was whether the subjects could become aware of the expectation when it was resolved (either validated or violated) by presentation of the terminal word.

Experiment 2: ease-of-remembering judgements following recognition

In this study, the subjects did not have to deduce the ease of the recognition decision from other characteristics of the trial: they were asked to make the ease-of-remembering judgements immediately after the recognition decision. In contrast to the results of Experiment 1, the subjects' ease-of-remembering judgements were almost as high for sentences with low-constraint stems as those with high-constraint stems (Table 17.2, lower

Table 17.2 Recognition and ease-of-remembering data for Experiment 2

Recognition: Hits and correct rejections		
Constraint of stem	**Original Sentences**	**Re-paired Sentences**
High	.76	.69
Low	.73	.80
Confidence of remembering (p [claim 'easy'])		
Constraint of stem	**Original Sentences**	**Re-paired Sentences**
High	.65	.60
Low	.63	.58

panel): judgements of 'easy' were made only 2% more often for sentences with high-constraint stems than low-constraint stems, the difference not being reliable, $F(1,26) = 1.38$, $MSe = .01$, $p = .251$, $\text{eta}^2 = .04$, although still in the wrong direction.

Apparently, the two types of sentence simply did not feel different to these subjects. The only distinction they made was on the basis of the novelty of the sentences, judging re-paired sentences 5% harder to classify in the recognition task compared to sentences presented in their original form, $F(1,26) = 4.00$, $MSe = .01$, $p = .056$, $\text{eta}^2 = .14$. Thus, although the subjects in this study did not make the mistake of claiming high-constraint items to be easier to judge for recognition, they failed to detect that they performed considerably better in rejecting novel sentences with low-constraint stems compared to those with high-constraint stems. According to my earlier interpretation of the effect, this difference comes about through the differential experience of violation of definite but general expectations versus violation of definite, specific expectations. The data demonstrate that the subjects could not consciously detect the difference between these types of violation of expectation, even though their recognition performance was strongly influenced by it.

General discussion

On the basis of evidence discussed earlier, I believe that remembering in this paradigm can only be understood through the idea that people's prior experiences serve as expectations about future interactions with the world. The present evidence suggests that validation or violation of those expectations causes conscious, valid feelings of familiarity or unfamiliarity; but neither the expectations nor the process of resolution is accessible to consciousness.

The idea that people carry around consciously inaccessible expectations is useful in understanding two kinds of memory errors: commissions and omissions. An example might be taking a holiday in Las Vegas. Before ever visiting one of its casinos, one has general expectations about what kinds of events might transpire. These general expectations, based on many prior second-hand experiences (books, films, etc.), prepare one to incorporate specific details of one's own personal experience, such as the degree of luck one might have in an evening at the tables, the behaviour of other gamblers, the general layout of the place and the kinds of games available to be played. They do not prepare one to incorporate other details, such as becoming violently ill through ptomaine poisoning in a restaurant. These differential outcomes will have predictable consequences for later remembering. The former experience will enable one later to remember the general nature of the experience (I won a lot or lost a lot); but because the detail is absorbed into the general meaning of the event, one can easily later generate or accept false suggestions of what transpired (e.g., misremembering that one won a bit in blackjack and lost a bit at vingt-et-un, when in fact the opposite occurred). That is, encoding detail under the influence of a strong, general expectation can result in easy later remembering of much of the event, but also in easy misremembering of consistent detail. In contrast, the latter experience (poisoning) changes the entire meaning of the event ('the vacation from hell'). The evidence suggests that such specific expectations, defined by a detail, can produce similar hits and even better rejection of lures than more general expectations. It also suggests a rather counter-intuitive outcome of such an experience: that despite the intensity of the experience, it may not be recallable when cued by the initial theme. If, some years after that holiday, one is asked if one has ever been to Las Vegas (without any cueing of illness), one may experience some sense of familiarity with the idea, but fail to retrieve any specifically corroborating evidence, and conclude that one has never been there. But cued by the idea of 'illness when traveling,' one may suddenly retrieve a particularly vivid (and accurate) series of images or sensations, and reverse the remembering claim: 'Of course I've been to Vegas! Nearly died there.'

One might interpret the pattern of recognition seen in the present experiments (equal hits, less false alarms for low-constraint stems) to mean merely that information that is consistent with an established schema is difficult to retrieve distinctively, but that inconsistent information is more elaborately encoded and more selectively retrieved. Those ideas are well established in the schema literature (e.g., Maki, 1990; Neuschatz et al., 2002; Woll & Graesser, 1982). But the present evidence is not just a re-statement of that principle. First, in a typical schema study, the subject is given only one script or story, or at most three or four. These themes are usually about broad and socially important concepts, such as 'restaurant,' 'political speech,' or 'the crime event,' with a wealth of detail that is more or less centrally associated with the major theme. In contrast, in the present studies, the subjects were presented with 120 themes within a test session, each consisting of a single statement about some simple and unimportant action or fact ('The gardener spent the day trimming the HEDGE;' 'The pond was alive with ducks and GEESE'). The schemas that controlled the subjects' recognition performance were thus coextensive with single simple events, rather than abstract, general prototypes. That schemas can be developed in single experiences, unrelated to any other events, is not an idea that has received much attention in the schema literature.

Second, the details that define the meaning of low-constraint sentences in the present studies are not inconsistent with the foregoing context (e.g., Her husband thought it would be nice to have a HEDGE; He spent a pleasant day watching the GEESE). Instead, they provide the first definite and specific piece of information to organize an otherwise fairly meaningless sentence. Thus their effect in creating more effective recognition than does similar detail following high-constraint stems is not modulated by violating expectation, but instead by creating it. Importantly, exactly the same pattern of effect in recognition (equal hits, asymmetric false alarms) is observed when high-constraint stems are followed by consistent words ('They sat in the garden on an old oak BENCH') versus incongruous words ('They sat in the garden on an old oak BLOOD'). Thus the challenge of these data for standard schema theory is to explain why terminal words that are simply consistent with low-constraint stems and terminal words that are incongruous with high-constraint stems produce similar recognition patterns. Whittlesea (2004) argued that high-constraint incongruous sentences have the same effect as low-constraint congruous sentences for the same reason: Both create an expectation about the terminal word that is more specific than that created by high-constraint congruous sentences. This explanation does not conflict with schema theory; instead, it goes beyond standard assumptions of that theory to examine the implications of expectations acquired 'on the fly.'

Third, the evidence from the current experiments, that expectations based on training experiences control recognition decisions but not judgements of ease-of-remembering, at least adds a dimension of complexity to the schema idea. Rather than thinking of schemas as robust, consolidated summaries of experience that inevitably bias performance, such evidence instead encourages thinking of them as potentially conditional and sensitive to interactions with details of the persons' current processing. The important part of the schema idea is that expectations control behaviour; but as illustrated by the current studies, expectations can come in a variety of flavours that can control performance in different ways, depending on the specific way that they are induced. The critical points demonstrated by the current experiments are that highly specific expectations that control behaviour may not actually become operative until the moment that the behaviour is executed; and that even then, they may control one aspect of the person's phenomenology (feelings of remembrance) without affecting others (feelings of ease of remembrance). That is, unlike the usual view of schemas as grand cognitive structures, editing and organizing all related stimulus inputs, the highly specific schemas demonstrated in these studies have strong but conditional influence on performance, occurring only when cued by specific combinations of stimulus input.

Afterword

I have loved my study of psychology. I have loved working with my grad students, trying to crack Nature's shell (I'm convinced She has no bones, no neat subdivisions); and with those researchers who really are trying to Figure It All Out, and who, in consequence, are deeply puzzled, not quite sure what to think. More than anything else, I have loved and been most impressed by the power of mind to overcome, with crushing simplicity and indifferent ease, the perverse, the indeterminate, the essentially ambiguous nature of our environment. *I* haven't figured it all out, far from it; but I've enjoyed the ride. So I say to you all, Goodbye, and thanks for the ghusc (or, as GBS put it, the photi). It's been (as certain Ladies have been known to say) a business doing pleasure with you.

Cheers, Bruce.

Notes

This research was supported by a grant from NSERC.

1. My operational definition of an indefinite expectation is to provide a sentence-stem context, in test only, prior to the presentation of a recognition target word, which was presented in isolation during the training phase (or is novel). My operational definition of a definite expectation is to provide an entire sentence during the training phase (using the same frames as before), thereby establishing a pre-existing relationship between stems and terminal words before the test.

Epilogue

Philip A. Higham and Jason P. Leboe

This volume is entitled *Constructions of Remembering and Metacognition*. That title is intended to reflect our belief that progress in understanding the human mind does not consist of discovery, but is rather an act of invention. As demonstrated by so many of the essays in this volume, people's awareness of their world is constructed, not apprehended: As Thomas Hobbes (1651/1904) explained more than three hundred years ago, we experience our reactions to the world, not the world itself. In consequence, our knowledge of the world is inevitably indirect, coloured, and contaminated by the inferential and attributive processes through which it is created. In addition to Hobbes, we owe this framework for understanding human psychology to a number of thinkers across the past few centuries, such as Immanual Kant (*Critique of Pure Reason*, 1781/1932), Jeremy Bentham (*Theory of Fictions*, Bentham & Ogden, 1814/1932), Hans Vaihinger (*Philosophy of As If*, 1911/1924), Alfred Adler (1927), and Sir Frederick Bartlett (1932). This constructive nature of mind applies just as much to the act of theorizing about mental functions as to any other activity. In our study of the human mind, we have only one tool: the human mind. Thus our job as cognitive psychologists is to be storytellers, crafting imaginative fictions about what mind might be like if we were ever able to examine it directly. Of course, some stories are better than others.

This volume is meant as a parting tribute to the Whittlesea approach to psychology. The approach is unusual, starting as it does with the assumption that the most obvious distinctions that can be drawn about human behaviour, such as remembering versus knowing, having general concepts versus specific experiences, knowing *how* versus knowing *what*, and being aware versus unaware of the source of one's performance, are merely nature's way of confusing and misleading scientists from the simplicity and generality of the underlying principles. Indeed, the title of the book was particularly meant to draw attention to common processes underlying judgements about previous experience (Ansons & Leboe; Arnold; Dienes, Scott, & Wan; Evans & Benjamin; Hockley; Kurilla & Westerman; Lindsay & Kantner; Lloyd & Miller; Mantonakis & Hastie; Mazzoni & Hanczakowski; Papesh & Goldinger) and the judgements people make about their own progress in learning, the strategies they employ to guide the learning process, and the processes they engage in during efforts to maximize their performance when the quality of their learning is being evaluated (Dunlosky, Ariel, & Thiede; Goldsmith; Higham; Moulin, Perfect, Akhtar, Williams, & Souchay). The role of attribution and construction emerges despite the range of cognitive tasks that form the basis of the studies described across the chapters of this volume and despite the conventional classification of these studies as investigations of remembering on one hand and metacognition on the other.

Rather than revealing divisions between functionally distinct subsystems of minds, such psychological dichotomies and even finer distinctions are imposed on mind by the investigator, in an attempt to break the apparent complexity of mind into manageable chunks. But according to Whittlesea, that breaking is an act of violence that fractures the very organization that the investigator is seeking to comprehend. Simply because the scientist is human, any attempt at understanding is a creative, constructive activity that imposes an organization on the thing to be understood that is different from what the thing really is. And that is the paradox within which cognitive scientists must conduct their business; it is the necessary starting point for any useful theoretical advance. We encourage our readers to avoid the reification error; of assuming that what seems obvious must be true. As a topic of investigation, the human mind is like the Delphic oracle, giving ambiguous messages that the credulous take as unequivocal support for their biases.

Across the chapters of this volume, there is a diversity of topics discussed and the authors provide a number of important, distinctive insights into the nature of human cognition. Hopefully, the unique aspects of each chapter will be useful and provocative, but our greater hope is that readers will use their powers of construction to adopt a more holistic appreciation of this volume. If you find yourself at this final stage of the book swimming in distinctions and nuance, why not abandon your *sharpener* ways, at least temporarily, and adopt a *leveller's* perspective. If it is not an approach you are used to, it might be a fun change and it will give you a taste of what it's like to be schooled in the Whittlesea style. To paraphrase John Lennon, perhaps one day you will even join us.

References

Abdi, H. (2002). What can cognitive psychology and sensory evaluation learn from each other? *Food Quality and Preference, 13*, 445–451.

Ackerman, R. & Goldsmith, M. (2008). Control over grain size in memory reporting – with and without satisficing knowledge. *Journal of Experimental Psychology: Human Learning and Memory, 34*, 1224–1245.

Adler, A. (1927). *Understanding human nature*. Oxford, UK: Greenberg.

Allison, R. I. & Uhl, K. P. (1964). Influence of beer brand identification on taste perception. *Journal of Marketing Research, 1*, 36–39.

Alter, A. L. & Oppenheimer, D. M. (2006). Predicting short-term stock fluctuations by using processing fluency. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 9369–9372.

Alter, A. L. & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*, 219–235.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136*, 569–576.

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General, 133*, 432–447.

Arnold, M. M. & Lindsay, D. S. (2002). Remembering remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 521–529.

Arnold, M. M. & Lindsay, D. S. (2007). 'I remember/know/guess that I knew it all along!' Subjective experience versus objective measures of the knew-it-all-along effect. *Memory & Cognition, 35*, 1854–1868.

Atkinson, R. C. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology, 96*, 124–129.

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 developments in mathematical psychology: Learning, memory, and thinking* (Vol. 1, pp. 184–216). San Francisco: Freeman.

Bacon, E., Izaute, M., & Danion, J. M. (2007). Preserved memory monitoring but impaired memory control during episodic encoding in patients with schizophrenia. *Journal of the International Neuropsychological Society, 13*, 219–227.

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experiment Psychology: General, 104*, 54–75.

Balcomb, F. & Gerken, L. (2008). Three-year-old children can implicitly monitor memory to guide responses on a visual matching task. *Developmental Science, 11*, 750–760.

Barnes, A., Dunlosky, J., Mazzoni, G., Narens, L., & Nelson, T. O. (1998). An integrative system of metamemory components involved in retrieval. In D. Gopher & A. Koriat (Eds), *Attention and performance, Cognitive regulation of performance: Interaction of theory and application* (Vol. 17). Cambridge, MA: MIT Press

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge, England: Cambridge University Press.
Bartoshuk, L. M., Fast, K., & Snyder, D. J. (2005). Differences in our sensory worlds: Invalid comparisons with labeled scales. *Current Directions in Psychological Science*, *14*, 122–125.
Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*, 276–292.
Beatty, J. & Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science*, *5*, 371–372.
Begg, I. & Armour, V. (1991). Repetition and the ring of truth: Biasing comments. *Canadian Journal of Behavioral Science*, *23*, 195–213.
Berger, J. & Fitzsimmons, G. (2008). Dogs on the street, pumas on your feet: How cues in the environment influence product evaluation and choice. *Journal of Marketing Research*, *45*, 1–14.
Bergström, Z. M., de Fockert, J., & Richardson-Klavehn, A. (2009). Event-related potential evidence that automatic recollection can be voluntarily avoided. *Journal of Cognitive Neuroscience*, *21*, 1280–1301.
Benjamin, A. S. (2008). Memory is more than just remembering: Strategic control of encoding, accessing memory, and making decisions. In A. S. Benjamin & B. H. Ross (Eds), *The psychology of learning and motivation: Skill and strategy in memory* (Vol. 48, pp. 175–223). London, UK: Academic Press.
Benjamin, A. S. & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language, 51,* 159–172.
Benjamin, A. S. & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. Reder (Ed.), *Implicit memory and metacognition*. Mahwah, NJ: Erlbaum.
Benjamin, A. S., Bjork, R. A., & Hirshman, E. (1998). Predicting the future and reconstructing the past: A Bayesian characterization of the utility of subjective fluency. *Acta Psychologica*, *98*, 267–290.
Benjamin, A. S. & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds), *Handbook of memory and metamemory* (pp. 73–94). New York: Psychology Press.
Benjamin, A. S., Diaz, M. L., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review*, *116*, 84–115.
Bernstein, D. M. (2005). Making sense of memory. *Canadian Journal of Experimental Psychology, 59*(3), 199–208.
Bernstein, D. M., Laney, C., Morris, E. K., & Loftus, E. F. (2005a). False beliefs about fattening foods can have healthy consequences. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 13724–13731.
Bernstein, D. M., Laney, C., Morris, E. K., & Loftus, E. F. (2005b). False memories about food can lead to food avoidance. *Social Cognition*, *23*, 11–34.
Bernstein, D. M. & Loftus, E. F. (2009). The consequences of false memories for food preferences and choices. *Perspectives on Psychological Science*, *4*, 135–139.
Bentham, J. & Ogden, C. K. (1814/1932). *Theory of Fictions*. London, UK: Paul, Trench, Trubner.
Berry, C. J., Shanks, D. R., & Henson, R. N. A. (2008). A single-system account of the relationship between priming, recognition, and fluency. *Journal of Experimental Psychology: Learning Memory and Cognition, 34*, 97–111.
Berry, D. C. & Dienes, Z. (1993). *Implicit learning: Theoretical and empirical issues*. Hove: Lawrence Erlbaum.

Berry, J. & West, R. L. (1993). Cognitive self-efficacy in relation to personal mastery and goal setting across the life span. *International Journal of Behavioral Development*, *16*, 351–379.

Bjorklund, D. F., Dukes, C., & Brown, R. D. (2009). The Development of Memory Strategies. In M. L. Courage & N. Cowan, N. (Eds), *Development of memory in infancy and childhood* (pp. 145–176). East Sussex: Psychology Press.

Blank, H., Nestler, S., von Collani, G., & Fisher, V. (2008). How many hindsight biases are there? *Cognition*, *106*, 1408–1440.

Bodner, G. E. & Lindsay, D. S. (2003). Remembering and knowing in context. *Journal of Memory and Language*, *48*, 563–580.

Bodner, G. E. & Mulji, R. (2010). Prime proportion affects masked priming of fixed and free-choice responses. *Experimental Psychology, 57,* 360–366.

Bodner, G. E. & Richardson-Champion, D. D. L. (2007). Remembering is in the details: Effects of test-list context on memory for an event. *Memory*, *15*, 718–729.

Bornstein, R. F. & D'Agostino, P. R. (1994). The attribution and discounting of perceptual fluency: Preliminary tests of a perceptual fluency/attributional model of the mere exposure effect. *Social Cognition*, *12*, 103–128.

Bornstein, R. F., Kale, A. R., & Cornell, K. R. (1990). Boredom as a limiting condition on the mere exposure effect. *Journal of Personality and Social Psychology*, *58*, 791–800.

Boucher, L. & Dienes, Z. (2003). Two ways of learning associations. *Cognitive Science, 27*, 807–842.

Brainerd, C. J. & Reyna, V. F. (2005). *The science of false memory.* New York, NY, US: Oxford University Press.

Brainerd, C. J., Reyna, V. F., & Ceci, S. J. (2008). Developmental reversals in false memory: A review of data and theory. *Psychological Bulletin*, *134*, 343–382.

Brainerd, C. J., Wright, R., & Reyna, V. F. (2001). Conjoint recognition and phantom recollection. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 307–327.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, *113*, 409–432.

Bransford, J. D., Barclay, J. R., & Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology*, *3*, 193–209.

Brewer, W. F. & Sampaio, C. (2006). Processes leading to confidence and accuracy in sentence recognition: A metamemory approach. *Memory*, *14*, 540–552.

Briñol, P., Petty, R. E., & Tormala, Z. L. (2006). The malleable meaning of subjective ease. *Psychological Science*, *17*, 200–206.

Brooks, L. R. (1978). Non-analytic concept formation and memory for instances. In E. Rosch & B. Lloyd (Eds), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Erlbaum.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin*, *109*, 204–223.

Brown, A. S. (2003). A review of the deja vu experience. *Psychological Bulletin*, *129*, 394–413.

Brown, J. (1976). An analysis of recognition and recall and of problems in their comparison. In J. Brown (Ed.), *Recall and recognition* (pp. 1–35). New York: Wiley.

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *Quarterly Journal of Experimental Psychology*, *29*, 461–473.

Brown, R. & McNeil, D. (1966). The 'tip of the tongue' phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *5*, 325–337.

Brown, S. & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 587–599.

Bruno, D., Higham, P. A., & Perfect, T. J. (2009). Global subjective memorability and the strength-based mirror effect in recognition memory. *Memory & Cognition*, *37*, 807–819.

Buchanan, B. & Henderson, P. W. (1992). Assessing the bias of preference, detection and identification measures of discrimination ability in product design. *Marketing Science*, *11*, 64–75.

Buchner, A. (1994). Indirect effects of synthetic grammar learning in an identification task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 550–566.

Bullinaria, J. A. (1994). Modeling reaction times. In L. S. Smith & P. J. B. Hancock (Eds), *Neural computation and psychology – Proceedings of the 3rd neural computation and psychology workshop, Stirling, 31st August 1994* (pp. 34–48). New York: Springer.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, *7*, 26–48.

Carruthers, P. (2007). Higher order theories of consciousness. In Velmans, M., & Schneider, S. (Eds), *The Blackwell Companion to Consciousness* (pp. 277–286). New York: Blackwell.

Carver, C. S. & Scheier, M. F. (2000). On the structure of behavioral self-regulation. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds), *Handbook of self-regulation* (pp. 41–84). New York: Academic Press.

Cary, M. & Reder, L. M. (2002). Metacognition in strategy selection: Giving consciousness too much credit. In P. Chambres, M. Izaute, & P. J. Marescaux (Eds), *Metacognition: Process, function, and use* (pp. 63–78). New York, NY: Kluwer.

Cary, M. & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language*, *49*, 231–248.

Ceci, S. J., Papierno, P. B., & Kulkofksy, S. (2007). Representational constraints on children's suggestibility, *Psychological Science*, *18*, 503–509.

Chalfonte, B. L. & Johnson, M. K. (1996). Feature memory and binding in young and older adults. *Memory & Cognition*, *24*, 403–416.

Chappel, M. & Humphreys, M. S. (1994). An auto-associative neural network for sparse representations: Analysis and application to models of recognition and cued recall. *Psychological Review*, *101*, 103–128.

Chi, M. T. H. (1978). Knowledge structures and memory development. In R. Siegler (Ed.), *Children's thinking: What develops?* (pp 73–96). Hillsdale, NJ: Erlbaum.

Clare, L., Wilson, B. A., Carter, G., Roth, I., & Hodges, J. R. (2002). Relearning face-name associations in early Alzheimer's disease. *Neuropsychology*, *16*, 538–547.

Clarke, F. R., Birdsall, T. G., & Tanner, W. P., Jr. (1959). Two types of ROC curves and definitions of parameters. *The Journal of the Acoustical Society of America*, *62*, 961–970.

Cleary, A. M., Ryals, A. J., & Nomi, J. S. (2009). Can déjà vu result from similarity to a prior experience? Support for the similarity hypothesis of déjà vu. *Psychonomic Bulletin & Review*, *16*, 1082–1088.

Cleeremans, A. (1993). *Mechanisms of implicit learning: Connectionist models of sequence processing.* Cambridge, MA: MIT Press.

Cleeremans, A. & Dienes, Z. (2008). Computational models of implicit learning. In R. Sun (Ed.), *Handbook of computational cognitive modeling* (pp. 396–421). Cambridge: University Press.

Clore, G. L. (1992). Cognitive phenomenology: Feelings and the construction of judgment. In L. L. Martin & A. Tesser (Eds), *The construction of social judgments* (pp. 133–163). Hillsdale, NJ: Erlbaum.

Cohen, R. L. (1989). Memory for action events: The power of enactment. *Educational Psychology Review*, *1*, 57–80.

Collins, A. M. & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging*, *12*, 50–71.

Conway, M. A. & Dewhurst, S. A. (1995). Remembering, familiarity, and source monitoring. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *48A*, 125–140.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*, 163–191.

Cowan, N. (1995). Attention and memory: An integrated framework. *Oxford Psychology Series, No. 26.* Oxford University Press.

Craik, F. I. M & Jennings, J. M. (1992). Human memory. In F. I. M. Craik & T. A. Salthouse (Eds), *The Handbook of Aging and Cognition* (pp. 51–110). Hilldale, NJ: Erlbaum.

Craik, F. I. M., Morris, L. W., Morris, R. G., & Loewen, E. R. (1990). Relations between source amnesia and frontal lobe functioning in older adults. *Psychology & Aging*, *5*, 148–151.

Curran, T., DeBuse, C., & Leynes, P. A. (2007). Conflict and criterion setting in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 2–17.

Cycowicz, Y. M., Friedman, D., Snodgrass, J. G., & Rothstein, M. (2000). A developmental trajectory in implicit memory is revealed by picture fragment completion. *Memory*, *8*, 19–35.

Daneman, M. & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Danion, J. M., Gokalsing, E., Robert, P., Massin-Krauss, M., & Bacon, E. (2001). Defective relationship between subjective experience and behavior in schizophrenia. *American Journal of Psychiatry*, *158*, 2064–2066.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17–22.

Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin & Review*, *13*, 1–21.

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2008). The effects of unitization on familiarity-based source memory: Testing a behavioral prediction derived from neuroimaging data. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 730–740.

Diaz, M. & Benjamin, A. S. (in press). The effects of proactive interference (PI) and release from PI on judgments of learning. *Memory & Cognition.*

Dienes, Z. (2008). Subjective measures of unconscious knowledge. *Progress in Brain Research*, *168*, 49–64.

Dienes, Z., Altmann, G., Kwan, L., & Goode, A. (1995) Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 1322–1338.
Dienes, Z. & Berry, D. (1997). Implicit synthesis. *Psychonomic Bulletin and Review, 4*, 68–72.
Dienes, Z. & Scott, R. B. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological Research*, *69*, 338–351.
Diller, D. E., Nobel, P. A., & Shiffrin, R. M. (2001). An ARC-REM model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *27*, 414–435.
Dobbins, I. G. & Han, S. (2008). What constitutes a model of item-based memory decisions? In A. S. Benjamin & B. H. Ross (Eds), *The Psychology of Learning and Motivation: Skill and Strategy in Memory Use* (Vol. 48, pp.95–144). London: Academic Press.
Dobbins, I. G. & Kroll, N. E. A. (2005). Distinctiveness and the recognition mirror effect: Evidence for an item-based criterion placement heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1186–1198.
Dodson, C. S. & Johnson, M. K. (1996). Some problems with the process-dissociation approach to memory. *Journal of Experimental Psychology: General*, *125*, 181–194.
Dodson, C. S. & Schacter, D. L. (2001). 'If I had said it I would have remembered it:' Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin and Review*, *8*, 155–161.
Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition*, *24*, 523–533.
Dougal, S. & Rotello, C. M. (2007). 'Remembering' emotional words is based on response bias, not recollection. *Psychological Bulletin & Review*, *14*, 423–429.
Dougal, S. & Schooler, J. W. (2007). Discovery misattribution: When solving is confused with remembering. *Journal of Experimental Psychology: General*, *136*, 577–592.
Drummey, A. B. & Newcombe, N. (1995). Remembering versus knowing the past: Children's explicit and implicit memories for pictures. *Journal of Experimental Child Psychology*, *59*, 549–565.
Drummey, A. B. & Newcombe, N. (2002). Developmental changes in source memory. *Developmental Science*, *5*, 502–513.
Dunlosky, J. & Ariel, R. (2009, November). *The influence of top-down and bottom-up control of self-regulated study*. Paper presented at the 50th Annual Meeting of the Psychonomic Society, Boston, MA.
Dunlosky, J. & Connor, L. T. (1997). Age differences in the allocation of study time account for age differences in memory performance. *Memory and Cognition*, *25*, 691–700.
Dunlosky, J., Domoto, P. K., Wang, M. L., Ishikawa, T., Roberson, I., Nelson, T. O., et al. (1998). Inhalation of 30% nitrous oxide impairs people's learning without impairing people's judgments of what will be remembered. *Experimental and Clinical Psychopharmacology*, *6*, 77–86.
Dunlosky, J. & Hertzog, C. (1998). Training programs to improve learning in later adulthood: Helping older adults educate themselves. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds), *Metacognition in Educational Theory and Practice* (pp. 249–275). Mahwah, NJ: Erlbaum.
Dunlosky, J. & Metcalfe, J. (2009). *Metacognition*. Sage Publications, Inc.
Dunlosky, J. & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*, 374–380.

Dunlosky, J. & Thiede, K. W. (2004). Causes and constraints of the shift-to-easier materials effect in control of study. *Memory and Cognition*, *32*, 779–788.

Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review*, *111*, 524–542.

Ekman, P. & Matsumoto, D. (1993). *Japanese and Caucasian neutral faces (JACNeuF)*. Photographs on CD-Rom.

Elder, R. S. & Krishna, A. (2010). The effects of advertising copy on sensory thoughts and perceived taste. *Journal of Consumer Research*, *36(5)*, 748–756.

Estes, W. K. & Maddox, W. T. (1995). Interactions of stimulus attributes, base rates, and feedback in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1075–1095.

Fagan, J. F. (1970). Memory in the infant. *Journal of Experimental Child Psychology*, *9*, 217–226.

Fang, X., Singh, S., & Ahluwalia, R. (2007). An examination of different explanations for the mere exposure effect. *Journal of Consumer Research*, *34*, 97–103.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is 'special' about face perception? *Psychological Review*, *105*, 482–498.

Farrant, A., Blades, M., & Boucher, J. (1999). Recall readiness in children with autism. *Journal of Autism and Developmental Disorders*, *29*, 359–366

Feenan, K. & Snodgrass, J. G. (1990). The effect of context on discrimination and bias in recognition memory for pictures and words. *Memory & Cognition*, *18*, 515–527.

Feustel, T. C., Shiffrin, R. M., & Salasoo, A. (1983). Episodic and lexical contributions to the repetition effect in word identification. *Journal of Experimental Psychology: General*, *112*, 309–346.

Fischhoff, B. & Beyth, R. (1975). 'I knew it would happen': Remembered probabilities of once-future things. *Organizational Behavior and Human Decision Processes*, *13*, 1–16.

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 349–358.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive developmental inquiry. *American Psychologist*, *34*, 906–911.

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology*, *50*, 21–45.

Flavell, J. H., Beach, D. H., & Chinsky, J. M. (1966). Spontaneous verbal rehearsal in a memory task as a function of age. *Child Development*, *37*, 283–299.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state: A practical method for grading the cognitive state of the patient for the clinician. *Journal of Psychiatric Research*, *12*, 189–198.

Gallo, D. A., Roediger, H. L., III, & McDermott, K. B. (2001). Associative false recognition occurs without strategic criterion shifts. *Psychonomic Bulletin & Review*, *8*, 579–586.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type-2 tasks in the theory of signal detectability. *Psychonomic Bulletin & Review*, *10*, 843–876.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory and Cognition*, *16*, 309–313.

Gardiner, J. M., Gregg, V. H., & Karayianni, I. (2006). Recognition memory awareness: Occurrence of perceptual effects in remembering or in knowing depends on conscious resources at encoding, but not at retrieval. *Memory and Cognition*, *34*, 227–239.

Gardiner, J. M., Kaminska, Z., Dixon, M., & Java, R. I. (1996). Repetition of previously novel melodies sometimes increases both remember and know responses in recognition memory. *Psychonomic Bulletin & Review*, *3*, 366–371.

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (2002). Recognition memory and decision processes: A meta-analysis of remember, know, and guess responses. *Memory*, *10*, 83–98.

Gellatly, A., Banton, P., & Woods, C. (1995). Salience and awareness in the Jacoby-Whitehouse effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 1374–1379.

Gentner, D. & Collins, A. (1981). Studies of inference from lack of knowledge. *Memory & Cognition*, *9*, 434–443.

Geraerts, E., Bernstein, D. M., Merckelbach, H., Linders, C., Raymaekers, L., & Loftus, E. F. (2008). Lasting false beliefs and their behavioral consequences. *Psychological Science*, *19*, 749–753.

Gibson, J. J. (1977). The theory of affordances. In R. E. Shaw & J. Bransford (Eds), *Perceiving, acting, and knowing*. Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Todd, P. M., & The ABC Research Group (1999). *Simple heuristics that make us smart*. New York, NY: Oxford University Press.

Gillund, G. & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 1-67.

Gleitman, L. R., Gleitman, H., & Shipley, E. F. (1972). The emergence of the child as grammarian. *Cognition*, *1*, 137–164.

Glucksberg, S. & McCloskey, M. (1981). Decision about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 311–325.

Goff, L. M. & Roediger, H. L., III. (1998). Imagination inflation for action events: Repeated imaginings lead to illusory recollections. *Memory & Cognition*, *26*, 20–33.

Goldinger, S. D., Azuma, T., Abramson, M., & Jain, P. (1997). Open wide and say 'blah!' Attentional dynamics of delayed naming. *Journal of Memory and Language*, *37*, 190–216.

Goldinger, S. D. & Hansen, W. A. (2005). Recognition by the seat of your pants. *Psychological Science*, *16*, 525–529.

Goldinger, S. D., He, Y., & Papesh, M. H. (2009). Deficits in cross-race face learning: Insights from eye movements and pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1105–1122.

Goldinger, S. D., Kleider, H. M., & Shelley, E. (1999). The marriage of perception and memory: Creating two-way illusions with words and voices. *Memory & Cognition*, *27*, 328–338.

Goldsmith, M., Jacoby, L. L., Halamish, V., & Wahlheim, C. N. (2009). *Metacognitively Guided Retrieval and Report (META-RAR): Quality control processes in recall*. Paper presented at the 50th Annual Meeting of the Psychonomic Society, Boston, MA.

Goldsmith, M. & Koriat, A. (2008). The strategic regulation of memory accuracy and informativeness. In A. Benjamin & B. Ross (Eds), *The psychology of learning and motivation, Vol. 48: Memory used as skilled cognition* (pp. 1–60). San Diego, CA: Elsevier.

Goldsmith, M., Koriat, A., & Pansky, A. (2005). Strategic regulation of grain size in memory reporting over time. *Journal of Memory and Language (Special Issue on Metamemory)*, *52*, 505–525.

Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). The strategic regulation of grain size in memory reporting. *Journal of Experimental Psychology: General*, *131*, 73–95.

Goldstein, D. G. & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75–90.

Goodman, L. A. & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*, 732–764.

Gruppuso, V., Lindsay, D. S., & Kelley, C. M. (1997). The process-dissociation procedure and similarity: Defining and estimating recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 259–278.

Guttentag, R. & Dunn, J. (2003). Judgments of remembering: The revelation effect in children and adults. *Journal of Experimental Child Psychology*, *86*, 153–167.

Han, S. & Dobbins, I. G. (2008). Examining recognition criterion rigidity during testing using a biased-feedback technique: Evidence for adaptive criterion learning. *Memory & Cognition*, *36*, 703–715.

Harmon-Jones, E. & Allen, J. J. B. (2001). The role of affect in the mere exposure effect: Evidence from psychophysiological and individual differences approaches. *Personality and Social Psychology Bulletin*, *27*, 889–898.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. J*ournal of Educational Psychology*, *56*, 208–216.

Hart, J.T., (1967). Memory and the memory-monitoring process. *Journal of Verbal Learning and Verbal Behavior*, *6*, 685–691

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Science*, *4*, 223–233.

Healy, A. F. & Jones, C. (1973). Criterion shifts in recall. *Psychological Bulletin*, *79*, 335–340.

Healy, A.F. & Kubovy, M. (1978). The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Memory & Cognition*, *6*, 544–553.

Heit, E., Brockdorff, N., & Lamberts, K. (2003). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin & Review*, *10*, 718–723.

Helmholtz, H. V. (1867/1950). *Treatise on physiological optics*. New York: Dover.

Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Memory & Cognition, 33,* 98–106.

Hertzog, C. & Dunlosky, J. (2004). Aging, metacognition, and cognitive control. In B. Ross (Ed.), *Psychology of learning and motivation*. (pp. 215–252). Elsevier Academic Press: Sand Diego, CA.

Heun, R., Burkhart, M. Wolf, C., & Benkert, O. (1998). Effect of presentation rate on word list learning in patients with dementia of the Alzheimer type. Dementia and Geriatric Cognitive Disorders, *9*, 214–218.

Hicks, J. L. & Marsh, R. L. (1998). A decrement-to-familiarity interpretations of the revelation effect from forced-choice tests of recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *24*, 1105–1120.

Higham, P. A. (unpublished manuscript). Dumb use of the fluency heuristic: Reliance on perceptual fluency in the absence of perceptual facilitation.

Higham, P. A. (2002). Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition*, *30*, 67–80.

Higham, P. A. (2007). No Special K! A signal detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General*, *136*, 1–22.

Higham, P. A. & Arnold, M. M. (2007). How many questions should I answer? Using bias profiles to estimate optimal bias and maximum score on formula-scored tests. *European Journal of Cognitive Psychology*, *19*, 718–742.

Higham, P. A. & Gerrard, C. (2005). Not all errors are created equal: Metacognition and changing answers on multiple-choice tests. *Canadian Journal of Experimental Psychology*, *59*, 28–34.

Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *35*, 57–80.
Higham, P. A. & Tam, H. (2005). Generation failure: Estimating metacognition in cued recall. *Journal of Memory and Language*, *52*, 595–617.
Higham, P. A. & Tam, H. (2006). Release from generation failure: The role of study-list structure. *Memory & Cognition*, *34*, 148–157.
Higham, P. A. & Vokey, J. R. (2000). Judgment heuristics and recognition memory: Prime identification and target processing fluency. *Memory, & Cognition*, *28*, 574–584.
Higham, P. A. & Vokey, J. R. (2004). Illusory recollection and dual-process models of recognition memory. *The Quarterly Journal of Experimental Psychology*, *57A*, 714–744.
Higham, P. A., Vokey, J. R., & Pritchard, J. (2000). Beyond dissociation logic: Evidence for controlled and automatic influences in artificial grammar learning. *Journal of Experimental Psychology: General*, *129*, 457–470.
Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528–551.
Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 302–313.
Hirshman, E. & Henzler, A. (1998). The role of decision processes in conscious recollection. *Psychological Science*, *9*, 61–65.
Hobbes, T. (1651/1904). *Leviathan, or, the matter, forme & power of a commonwealth, ecclesiasticall and civill*. Cambridge, UK: Cambridge University Press.
Hockley, W. E. & Caron, A. M. (2007). *Opposing strength-based mirror effects for words versus pictures: Evidence for within-list criterion changes*. Paper presented at the *48th Annual Meeting of the Psychonomic Society*, Long Beach, CA.
Hockley, W. E., & Niewiadomski, M. W. (2001). Interrupting recognition memory: Tests of a criterion-change account of the revelation effect. *Memory & Cognition*, *29*, 1176–1184.
Hockley, W. E. & Niewiadomski, M. W. (2007). Strength-based mirror effects in item and associative recognition: Evidence for within-list criterion changes. *Memory & Cognition*, *35*, 679–688.
Hoegg, J. & Alba, J. W. (2007). Taste perception: More than meets the tongue. *Journal of Consumer Research*, *33*, 490–498.
Hoff-Sommers, C. (1995). *Who stole feminism*. New York, NY: Touchstone.
Holliday, R. E. & Weekes, B. S. (2006). Dissociated developmental trajectories for semantic and phonological false memories. *Memory*, *14*, 624–636.
Hsee, C. K., Yang, Y., Gu, Y., & Chen, J. (2009). Specification seeking: How product specifications influence consumer preference. *Journal of Consumer Research*, *35*, 952–966.
Hsee, C. K., Yang, Y., Li, N., & Shen, L. (2009). Wealth, warmth, and well-being: Whether happiness is relative or absolute depends on whether it is about money, acquisition, or consumption. *Journal of Marketing Research*, *46*, 396–409.
Huber, D. E., Clark, T., Curran, T., & Winkielman, P. (2008). Effects of repetition priming on recognition memory: Testing a perceptual fluency-disfluency model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1305–1324.
Huber, D. E. & O'Reilly, R. C. (2003). Persistence and accommodation in short-term priming and other perceptual paradigms: Temporal segregation through synaptic depression. Cognitive Science, *27*, 403–430.

Jacoby, L. L. (1983a). Perceptual enhancement: Persistent effects of an experience. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *1*, 21–38.

Jacoby, L. L. (1983b). Remembering the data: Analyzing interactive processes in reading. *Journal of Verbal Learning and Verbal Behavior*, *22*, 485–508.

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. (1998). Invariance in automatic influences of memory: Toward a user's guide for the process-dissociation procedure. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *24*, 3–26.

Jacoby, L. L., Allan, L. G., Collins, J. C., & Larwill, L. K. (1988). Memory influences subjective experience: Noise judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *14*, 240–247.

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

Jacoby, L. L. & Kelley, C. M. (1987). Unconscious influences of memory for a prior event. *Personality and Social Psychology Bulletin*, *13*, 314–336.

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. In H. L. Roediger & F. I. M. Craik (Eds), *Varieties of memory and consciousness: Essays in honor of Endel Tulving* (pp. 391–422). Hillsdale, NJ: Erlbaum.

Jacoby, L. L. & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General*, *118*, 126–135.

Jacoby, L. L. & Witherspoon, D. (1982). Remembering without awareness. *Canadian Journal of Psychology, 36,* 300–324.

Jacoby, L. L., Woloshyn, V., & Kelley, C. (1989). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. *Journal of Experimental Psychology: General*, *118*, 115–125.

Jacoby, L. L., Yonelinas, A. P., & Jennings, J. (1997). The relation between conscious and unconscious (automatic) influences: A declaration of independence. In J. Cohen & J. W. Schooler (Eds), *Scientific approaches to consciousness* (pp. 13–47). Mahwah, NJ: Erlbaum.

Jamieson, R. K. & Mewhort, D. J. K. (2010). Applying an exemplar model to the artificial-grammar task: String-completion and performance on individual items. *Quarterly Journal of Experimental Psychology, 63,* 1014–1039.

Jennings, J. M. & Jacoby, L. L. (2003). Improving memory in older adults: Training recollection. *Neuropsychological Rehabilitation*, *13*, 417–440.

Johansson, T. (2009). In the fast lane toward structure in implicit learning: Non-analytic processing and fluency in artificial grammar learning. *European Journal of Cognitive Psychology*, *21*, 129–160.

Johnson, M. K., Foley, M. A., Suengas, A. G., & Raye, C. L. (1988). Phenomenal characteristics of memory for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General*, *117*, 371–376.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source Monitoring. *Psychological Bulletin*, *114*, 3–28.

Johnson, M. K., Raye, C. L., Foley, H. J., & Foley, M. A. (1981). Cognitive operations and decision bias in reality monitoring. *American Journal of Psychology*, *94*, 37–64.

Johnston, W., Dark, V., & Jacoby, L. L. (1985). Perceptual fluency and recognition judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *11*, 3–11.

Johnston, W., Hawley, K., & Elliott, J. (1991). Contribution of perceptual fluency to recognition judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *17*, 210–223.

Joordens, S. & Merikle, P. (1992). False recognition and perception without awareness. *Memory, & Cognition*, *20*, 151–159.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697–720.

Kahneman, D. & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*, 136–153.

Kant, I. (1781/1932). *Critique of pure reason*. London, UK: MacMillan.

Kantner, J. & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition, 38,* 389–406.

Karpicke, J. D., McCabe, D. P., & Roediger, H. L. (2008). False memories are not surprising: The subjective experience of an associative memory illusion. *Journal of Memory and Language*, *58*, 1065–1079.

Kelley, C. M., Jacoby, L. L., & Hollingshead, A. (1989). Direct versus indirect tests of memory for source: Judgments of modality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1101–1108.

Kelley, C. M. & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, *32*, 1–24.

Kelley, C. M. & Rhodes, M. G. (2002). Making sense and nonsense of experience: Attributions in memory and judgment. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory, vol. 41* (pp. 293–320). San Diego, CA US: Academic Press.

Kelley, C. M. & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language*, *48*, 704–721.

Keren, G. & Tiegen, K. H. (2004). Yet another look at the heuristics and biases approach. In D. J. Koehler & Nigel Harvey (Eds), *Handbook of judgement and decision making* (pp. 89–109). Oxford, UK: Blackwell.

Kimball, D. R. & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition*, *31*, 918–929.

Kinder, A. & Assmann, A. (2000). Learning Artificial Grammars: No evidence for the acquisition of rules. *Memory and Cognition*, *28*, 1321–1332.

Kinder, A., Shanks, D. R., Cock, J., & Tunney, R. J. (2003). Recollection, fluency, and the explicit/implicit distinction in artificial grammar learning. *Journal of Experimental Psychology: General*, *132*, 551–565.

Kinoshita, S. (1997). Masked target priming effects on feeling-of-knowing and feeling-of-familiarity judgments. *Acta Psychologica*, *97*, 183–199.

Kintsch, W. (1967). Memory and decision aspects of recognition learning. *Psychological Review*, *74*, 496–504.

Kleider, H. M. & Goldinger, S. D. (2004). Illusions of face memory: Clarity breeds familiarity. *Journal of Memory and Language*, *50*, 196–211.

Kleider, H. M. & Goldinger, S. D. (2006). The generation and resemblance heuristics in face recognition: Cooperation and competition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 259–276.

Klin, C. M., Guzman, A. E., & Levine, W. H. (1997). Knowing that you don't know: Metamemory and discourse processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1378–1393.

Klinkenborg, V. (2009, June 3). The familiar place. *The New York Times* (p. A26).

Knowlton, B. J. & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 79–91.

Koppell, S. (1977). Decision latencies in recognition memory: A signal detection theory analysis. *Journal of Experimental Psychology: Human Learning and Memory*, *3*, 445–457.

Koren, D., Seidman, L. J., Goldsmith, M., & Harvey, P. D. (2006). Real-world cognitive – and metacognitive – dysfunction in schizophrenia: a new approach for measuring (and remediating) more 'right stuff'. *Schizophrenia Bulletin*, *32*, 310–326.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *34*, 945–959.

Koriat, A., Ben-Zur, H., & Sheffer, D. (1988). Telling the same story twice: Output monitoring and age. *Journal of Memory and Language*, *27*, 23–29.

Koriat, A. & Goldsmith, M. (1996a). Memory metaphors and the real-life/laboratory controversy: Correspondence versus storehouse conceptions of memory. *Behavioral and Brain Sciences*, *19*, 167–188.

Koriat, A. & Goldsmith, M. (1996b). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490–517.

Koriat, A. Goldsmith, M., & Halamish, V. (2008). Control processes in voluntary remembering. In J. Byrne (Series Ed.) & H. L. Roediger, III (Vol. Ed.), *Cognitive psychology of memory. Vol. 2 of Learning and memory: A comprehensive reference, 4 vols.* (J. Byrne, Editor) (pp. 307–324). Oxford, UK: Elsevier.

Koriat, A. & Lieblich, I. (1977). A study of memory pointers. *Acta Psychologica*, *41*, 151–164.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationship between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*, 36–69.

Koriat, A. & Nussinson, R. (2009). Attributing study effort to data-driven and goal-driven effects: Implications for metacognitive judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1338–1343.

Kornell, N. & Bjork, R. A. (2008). Optimizing self-regulated study: The benefits- and costs- of dropping flashcards. *Memory*, *16*, 125–136.

Kornell, N. & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 609–622.

Köster, E. P. (2003). The psychology of food choice: Some often encountered fallacies. *Food Quality and Preference*, *14*, 359–373.

Kreutzer, M. A., Leonard, C., and Flavell, J. H. (1975). An interview study of children's knowledge about memory. *Monographs of the Society for Research in Child Development*, *159*, 1–58.

Krishna, A. (2006). Interaction of senses: The effect of vision versus touch on the elongation bias. *Journal of Consumer Research*, *32*, 557–566.

Kuchinke, L., Võ, M. L-H., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology*, *65*, 132–140.

Kuhn, G. & Dienes, Z. (2005). Implicit learning of nonlocal musical rules: Implicitly learning more than chunks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1417–1432.

Kuhn, G. & Dienes, Z. (2008). Learning non-local dependencies. *Cognition*, *106*, 184–206.

Kurilla, B. P. & Westerman, D. L. (2008). Processing fluency affects subjective claims of recollection. *Memory and Cognition*, *36*, 82–92.

Labroo, A. A. & Kim, S. (2009). The 'instrumentality' heuristic: Why metacognitive difficulty is desirable during goal pursuit. *Psychological Science*, *20*, 127–134.

Lane, S. M., Roussel, C. C., Villa, D., & Morita, S. K. (2007). Features and feedback: Enhancing metamnemonic knowledge at retrieval reduces source-monitoring errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 1131–1142.

Laney, C., Morris, E. K., Bernstein, D. M., Wakefield, B. M., & Loftus, E. F. (2008). Asparagus, a love story: Healthier eating could be just a false memory away. *Experimental Psychology*, *55*, 291–300.

Le Ny, J. F., Denhière, G., & Le Taillanter, D. (1972). Regulation of study-time and interstimulus similarity in self-paced learning conditions. *Acta Psychologica*, *36*, 280–289.

Leboe, J. P. & Ansons, T. A. (2006). On misattributing good remembering to a happy past: An investigation into the cognitive roots of nostalgia. *Emotion*, *6*, 596–610.

Leboe, J. P. & Whittlesea, B. W. A. (2002). The inferential basis of familiarity and recall: Evidence for a common underlying process. *Journal of Memory and Language*, *46*, 804–829.

Lee, A. Y. (2001). The mere exposure effect: An uncertainty reduction explanation revisited. *Personality and Social Psychology Bulletin*, *27*, 1255–1266.

Lee, A. Y. & Labroo, A. A. (2004). The effect of conceptual and perceptual fluency on brand evaluation. *Journal of Marketing Research*, *41*, 151–165.

Lee, L., Frederick, S., & Ariely, D. (2006). Try it, you'll like it: The influence of expectation, consumption, and revelation on preferences for beer. *Psychological Science*, *17*, 1054–1058.

Levin, D. T. (1996). Classifying faces by race: The structure of face categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1364–1382.

Levin, D. T. (2000). Race as a visual feature: Using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General*, *129*, 559–574.

Leynes, P. A., Landau, J., Walker, J., & Addante, R. J. (2005). Event-related potential evidence for multiple causes of the revelation effect. *Consciousness and Cognition*, *14*, 327–350.

Liberman, V. & Tversky, A. (1993). On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psychological Bulletin*, *114*, 162–173.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, England: Cambridge University Press.

Lie, E. & Newcombe, N. (1999). Elementary school children's explicit and implicit memory for faces of preschool classmates. *Developmental Psychology*, *35,* 102–112.

Lindsay, D. S. (2008). Source monitoring. In J. Byrne (Series Ed.) & H. L. Roediger, III (Vol. Ed.), *Learning and memory: A comprehensive reference. Vol. 2: Cognitive psychology of memory* (pp. 325–348). Oxford, England: Elsevier.

Lindsay, D. S. & Kelley, C. M. (1996). Creating illusions of familiarity in a cued recall remember/know paradigm. *Journal of Memory and Language*, *35*, 197–211.

Lloyd, M. E. & Newcombe, N. S. (2008). Implicit memory in childhood: Reassessing developmental invariance. In M. L. Courage & N. Cowan (Eds), *The development of memory in childhood, 2nd edition*. London: Psychology Press, 93–114.

Lloyd, M. E., Westerman, D. L., & Miller, J. M. (2003). The fluency heuristic in recognition memory: The effect of repetition. *Journal of Memory & Language*, *48*, 603–614.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 19–31.

Loftus, E. F. & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*, 585–589.

Lotz, A. & Kinder, A. (2006). Transfer in artificial grammar learning: The role of repetition information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 707–715.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *10*, 756–766.

Lutz, K. A. & Lutz, R. J. (1977). Effects of interactive imagery on learning: Applications to advertising. *Journal of Applied Psychology*, *62*, 493–498.

Lynch, J. G. J. & Srull, T. K. (1982). Memory and attentional factors in consumer choice: Concepts and research methods. *Journal of Consumer Research*, *9*, 18–33.

MacLin, O. H. & Malpass, R. S. (2001). Racial categorization of faces: The ambiguous race face effect. *Psychology, Public Policy, & Law*, *7*, 98–118.

Macmillan, N. A. (1993). Signal detection theory as data analysis method and psychological decision model. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 21–57). Hillsdale, NJ: Erlbaum.

MacMillan, N. A. & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge, UK: Cambridge University Press.

Major, J. C. & Hockley, W. E. (2007). A test of two different revelation effects using forced-choice recognition. *Psychonomic Bulletin & Review*, *14*, 1096–1100.

Maki, R. H. (1990). Memory for script actions: Effects of relevance and detail expectancy. *Memory & Cognition*, *18*, 5–14.

Malmberg, K. J. (2008). Investigating metacognitive control in a global memory framework. In J. Dunlosky and R. E. Bjork (Eds), *Handbook of memory and metacognition* (pp. 265–283). Hillsdale, NJ: Psychology Press.

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

Mandler, G. (2008). Familiarity breeds attempts: A critical review of dual-process theories of recognition. *Perspectives on Psychological Science*, *3*, 390–399.

Mantonakis, A., Whittlesea, B. W. A., & Yoon, C. (2008). Consumer Memory, Fluency, and Familiarity. In C. P. Haugtvedt, P. Herr, & F. R. Kardes (Eds), *The handbook of consumer psychology* (pp. 77–102). Mahwah, NJ: Erlbaum.

Massin-Krauss M., Bacon E., & Danion J-M. (2002). Effects of the benzodiazepine lorazepam on monitoring and control processes in semantic memory. *Consciousness and Cognition*, *11*, 123–137.

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 509–527.

Matthews, R. C., Buss, R. R., Stanley, W. B., & Blanchard Fields, F. (1989). Role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1083–1100.

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs? *Memory & Cognition*, *29*, 222–233.

Matvey, G., Dunlosky, J., & Schwartz, B. L. (2006). The effects of categorical relatedness on judgements of learning (JOLs). *Memory*, *14*, 253–261.

Matzen, L. E. & Benjamin, A. S. (2009). Remembering words not presented in sentences: How study context changes patterns of false memories. *Memory & Cognition*, *37*, 52–64.

Matzen, L. E., Taylor, E. G., & Benjamin, A. S. (in press). Contributions of familiarity and recollection rejection to recognition: Evidence from the time course of false recognition for semantic and conjunction lures. *Memory*.

Mazzoni, G. (2007). Did you witness demonic possession? A response time analysis of the relationship between event plausibility and autobiographical beliefs. *Psychonomic Bulletin & Review*, *14*, 277–281.

Mazzoni, G. (in preparation). Plausibility as a metacognitive judgment in retrieval.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study time allocation? *Memory & Cognition*, *18*, 196–204.

Mazzoni, G. & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning Memory and Cognition*, *21*, 1263–1274.

Mazzoni, G. & Nelson, T. O. (Eds), (1998). *Metacognition and cognitive neuropsychology*. Mahwah, NJ: Erlbaum.

Mazzoni, G. & Kirsch, I. (2002). Autobiographical memories and beliefs: A preliminary metacognitive model. In T. Perfect, B. Schwartz (Eds.), *Applied Metacognition* (pp. 121–145). Cambridge, U.K.: Cambridge University Press.

McCabe, D. P. & Balota, D. A. (2007). Context effects on remembering and knowing: The expectancy heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 536–549.

Mecklenbrauker, S. Hupbach, A., & Wippich, W. (2003). Age-related improvements in a conceptual implicit memory test. *Memory & Cognition*, *31*, 1208–1217.

Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Meissner, C. A. & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*, 3–35.

Merritt, P., Hirshman, E., Hsu, J., & Berrigan, M. (2005). Metamemory without the memory: are people aware of midazolam-induced amnesia? *Psychopharmacology*, *177*, 336–343.

Metcalfe, J. (1993). Novelty monitoring, metacognition, and control in a composite holograph associative recall model: Implication for Korsakoff Amnesia. *Psychological Review*, *100*, 3–22.

Metcalfe, J. (1996). Metacognitive processes. In E. L. Bjork & R. A. Bjork (Eds), *The handbook of perception and cognition: Vol 10. Memory* (pp. 381–407). San Diego, CA: Academic Press.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General*, *131*, 349–363.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychology Science*, *18*, 159–163.

Metcalfe, J. & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General*, *132*, 530–542.

Metcalfe, J. & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*, 463–477.

Metcalfe, J. M., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 861–861.

Michotte, A. (1963). *The perception of causality.* Oxford, UK: Basic Books.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior.* New York, NY: Holt.

Miller, J. K., Lloyd, M. E., & Westerman, D. L. (2008). When does modality matter? Perceptual versus conceptual fluency-based illusions in recognition memory. *Journal of Memory and Language*, *58*, 1080–1094.

Miller, M. B. & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. Psychological Review, *106*, 398–405.

Mitchell, A. A. & Olson, J. C. (1981). Are product attribute beliefs the only mediator of advertising effects on brand attitude? *Journal of Marketing Research*, *18*, 318–332.

Mitchell, K. J. & Johnson, M. K. (2000). Source monitoring: Attributing mental experiences. In E. Tulving & F. I. M. Craik (Eds), *The Oxford handbook of memory* (pp. 179–195). New York, NY: Oxford University Press.

Morrell, H. E. R., Gaitan, S., & Wixted, J. T. (2002). On the nature of the decision axis in signal-detection-based models of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1095–1110.

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.

Moscovitch, M. (1992). Memory and working with memory: A component process model based on modules and central systems. *Journal of Cognitive Neuroscience*, *4*, 257–267.

Moscovitch, M. (1994). Interference at retrieval from long-term memory: The influences of frontal and temporal lobes. *Neuropsychology*, *4*, 525–534.

Moulin, C. J. A. (2002). Sense and sensitivity: Metacognition in Alzheimer's disease. In T. J. Perfect & B. L. Schwartz (Eds), *Applied metacognition*. Cambridge, UK: Cambridge University Press.

Moulin, C. J. A., Perfect, T. J., & Jones, R. W. (2000a). The effects of repetition on allocation of study time and judgements of learning in Alzheimer's disease. *Neuropsychologia*, *38*, 748–756.

Moulin, C. J. A., Perfect, T. J., & Jones, R. W. (2000b). Evidence for intact memory monitoring in Alzheimer's disease: metamemory sensitivity at encoding. *Neuropsychologia*, *38*, 1242–1250.

Moulin, C. J. A., Perfect, T. J., & Jones, R. W. (2000c). Global predictions of memory in Alzheimer's disease: Evidence for preserved metamemory monitoring. *Aging Neuropsychology and Cognition*, *7*, 230–244.

Moulin, C. J. A., James, N., Perfect, T. J., & Jones, R. W. (2003). Knowing what you cannot recognise: Further evidence for intact metacognition in Alzheimer's disease. *Aging, Neuropsychology & Cognition*, *10*, *74–82*.

Murrell, G. A. & Morton, J. (1974). Word recognition and morphemic structure. *Journal of Experimental Psychology*, *102*, 963–968.

Neisser, U. (1954). An experimental distinction between perceptual process and verbal response. *Journal of Experimental Psychology*, *47*, 399–402.

Neisser, U. (1967). *Cognitive psychology.* New York: Appleton-Century-Crofts.

Neisser, U. & Harsch, N. (1992). Phantom flashbulb: False recollections about hearing the news about the *Challenger.* In E. Winograd & U. Neisser (Eds), *Affect and*

accuracy in recall: Studies of 'flashbulb' memories (pp. 9–31). New York: Cambridge University Press.
Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories and recognition. *Psychological Review, 105*, 299–324.
Nelson, D. L, Reed, U. S., & Walling, J. R. (1976). Picture superiority effect. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 523–528.
Nelson, D. L., Schreiber, T. A., & McEvoy, C. L. (1992). Processing implicit and explicit representations. *Psychological Review, 99*, 322-48.
Nelson, H. E. (1982). *National Adult Reading Test (NART): Test Manual.* Windsor: NFER-Nelson.
Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133.
Nelson, T. O. (1996a). Consciousness and metacognition. *American Psychologist, 51*, 102–116.
Nelson, T. O. (1996b). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item. Comments on Schraw (1995). *Applied Cognitive Psychology, 10*, 257–260.
Nelson, T. O. & Dunlosky, J. (1991). When peoples' judgments of learning (JOLs) are extremely accurate at predicting subsequent recall – the delayed-JOL effect. *Psychological Science, 2*, 267–270
Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5*, 207–213.
Nelson, T. O., Gerler, D., & Narens, L. (1984). Accuracy of feeling of knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General, 113*, 282–300.
Nelson, T. O. & Narens, L. (1990). Metamemory: A theoretical framework and some new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 26, pp. 125–173). Oxford, UK: Academic Press.
Nelson, T. O. & Leonesio, R. J. (1988). Allocation of self-paced study and the 'Labour in Vain Effect'. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*, 676–686.
Neuschatz, J. S., Lampinen, J. Preston, E. L., Hawkins, E. R., & Toglia, M. P. (2002). The effect of memory schemata on memory and the phenomenological experience of naturalistic situations. *Applied Cognitive Psychology, 16,* 687–708.
Newell, A. & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.
Newman, L. S. & Baumeister, R. F. (1998). Abducted by aliens: Spurious memories of interplanetary masochism. In S. J. Lynn, K. M. McConkey, & N. P. Spanos (Eds), *Truth and memory* (pp. 284–303). New York, NY: Guilford.
Ng, W. & Lindsay, R. C. L. (1994). Cross-race facial recognition: Failure of the contact hypothesis. *Journal of Cross-Cultural Psychology, 25*, 217–232.
Norman, E., Price, M. C., Duff, S. C., & Mentzoni, R. A. (2007). Gradations of awareness in a modified sequence learning task. *Consciousness and Cognition, 16*, 809–837.
Novemsky, N., Dhar, R., Schwarz, N., & Simonson, I. (2007). Preference fluency in choice. *Journal of Marketing Research, 44*, 347–356.

Oppenheimer, D. M. & Frank, M. C. (2008). A rose in any other font would not smell as sweet: Effects of perceptual fluency on categorization. *Cognition*, *106*, 1178–1194.

Paivio, A., Yuille, J. C., & Smythe, P. C. (1966). Stimulus and response abstractness, imagery and meaningfulness, and reported mediators in paired-associated learning. *Canadian Journal of Psychology*, *20*, 362–377.

Pansky, A., Koriat, A., Goldsmith, M., & Pearlman-Avnion, S. (2009). Memory accuracy and distortion in old age: Cognitive, metacognitive, and neurocognitive determinants. *European Journal of Cognitive Psychology*, *21*, 303–329.

Papesh, M. H. & Goldinger, S. D. (2009). Pupil-Blah-Metry: Word frequency reflected in physiology. *Paper presented at the Western Psychology Association conference in Portland, OR (April, 2009).*

Parkin, A. J. (1998). The development of procedural and declarative memory. In N. Cowan (Ed.), *The development of memory in childhood* (pp. 113–137). Hove, UK: Psychology Press.

Parkin, A. J. & Streete, S. (1988). Implicit and explicit memory in young children and adults. *British Journal of Psychology*, *79*, 361–369.

Parks, C. M. & Yonelinas, A. P. (2007). Moving beyond pure signal-detection models: Comment on Wixted. *Psychological Review*, *114*, 188–202.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker.* Cambridge: Cambridge University Press.

Payne, D. G., Elie, C. J., Blackwell, J. M., & Neuschatz, J. S. (1996). Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Memory and Language*, *35*, 261–285.

Perez, L. A., Peynircioglu, Z. F., & Blaxton, T. A. (1998). Developmental differences in implicit and explicit memory performance. *Journal of Experimental Child Psychology*, *70*, 167–185.

Perfect, T. J., Mayes, A. R., Downes, J. J., & Van Eijk, R. (1996). Does context discriminate recollection from familiarity in recognition memory? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *49A*, 797–813.

Perrotin, A., Belleville, S., & Isingrini, M. (2007). Metamemory monitoring in mild cognitive impairment: Evidence of a less accurate episodic feeling-of-knowing. *Neuropsychologia*, *45*, 2811–2826.

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment – Clinical characterization and outcome. *Archives of Neurology*, *56*, 303–308.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds), *Handbook of self-regulation* (pp. 451–502). New York, NY: Academic Press.

Polya, G. (1945). *How to solve it.* Princeton, NJ: University Press.

Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and counting: Insights from pupillometry. *Quarterly Journal of Experimental Psychology*, *60*, 211–229.

Porter, G. & Troscianko, T. (2003). Pupillary response to grating stimuli. *Perception*, *32*, 156.

Price, J., Hertzog, C., & Dunlosky, J. (2010). Self-regulated learning in younger and older adults: Does aging affect metacognitive control? *Aging, Neuropsychology and Cognition, 17,* 329–359.

Posner, M. I. & Mcleod, P. (1982). Information processing models – in search of elementary operations. *Annual Review of Psychology*, *33*, 477–514.

Postma, A. (1999). The influence of decision criterion upon remembering and knowing in recognition memory. *Acta Psychologica*, *103*, 65–76.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60,* 437–447.

Raghubir, P., Tyebjee, T.T., & Lin, Y.C. (2008). The sense and nonsense of consumer product testing: How to identify whether consumers are blindly loyal? *Foundations and Trends® in Marketing, 3,* 127–176.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory and Cognition*, *21*, 89–102.

Rajaram, S. (1996). Perceptual effects on remembering: Recollective processes in picture recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 365–377.

Rajaram, S. & Geraci, L. (2000). Conceptual Fluency selectively influences knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1070–1074.

Rajaram, S. & Roediger, H. L. (1993). Direct comparison of four implicit memory tests. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*, 765–776.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M., (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 163–178.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*, 518–535.

Reber, A. S. (1967). Implicit Learning of Artificial Grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*, 855–863.

Reber, A. S. & Lewis, S. (1977). Implicit learning: an analysis of the form and structure of a body of tacit knowledge. *Cognition*, *114*, 14–24.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, *8*, 364–382.

Reder, L. M. (1982). Plausibility judgements vs. fact retrieval: Alternative strategies for sentence verification. *Psychological Review*, *89*, 250–280.

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology*, *12*, 90–138.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 294–320.

Reder, L. M. & Schunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 45–78). Mahwah, NJ: Erlbaum.

Reed, N., McLeod, P., & Dienes, Z. (2010). Implicit knowledge and motor skill: What people who know how to catch don't know. *Consciousness & Cognition 19,* 63–76.

Rehder, B. & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, *130*, 323–360.

Rhodes, M. G. & Castell, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, *137*, 615–625.

Rhodes, M. G. & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review, 16*, 550–554.

Rhodes, M. G. & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 305–320.

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

Richardson-Klavehn, A., Gardiner, J. M., & Java, R. I. (1996). Memory: Task dissociations, process dissociations and dissociations of consciousness. In G. Underwood (Ed.), *Implicit cognition* (pp. 85–158). Oxford, UK: Oxford University Press.

Richardson-Klavehn, A., Lee, M. G., Joubran, R., & Bjork, R. A. (1994). Intention and awareness in perceptual identification priming. *Memory &Cognition*, *22*, 293–312.

Roberts, K. P. & Blades, M. (2000). *Children's source monitoring*. Mahwah, NJ: Erlbaum.

Roebers, C. M., von der Linden, N., Schneider, W., & Howie, P. (2007). Children's metamemorial judgments in an event recall task. *Journal of Experimental Child Psychology*, *97*, 117–137.

Roediger, H. L., III. & Blaxton, T. A. (1987). Retrieval modes produce dissociations in memory for surface information. In D. S. Gorfein & R. R. Hoffman (Eds), *Memory and Learning*. Hillsdale, NJ: Erlbaum.

Roediger, H. L., III. & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 803–814.

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

Rosenthal, D. M. (2002). Consciousness and higher order thought. *Encyclopedia of cognitive science* (pp. 717–726). New York, NY: Nature.

Rotello, C. M. & Heit, E. (*2000*). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition*, *28*, 907–922.

Rotello, C. M. & Macmillan, N. A. (2008). Response bias in recognition memory. In A. S. Benjamin, B. H. Ross, A. S. Benjamin, & B. H. Ross (Eds), *Skill and strategy in memory use* (pp. 61–94). San Diego, CA: Elsevier Academic Press.

Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum-difference theory of remembering and knowing: A two-dimensional signal-detection model. *Psychological Review*, *111*, 588–616.

Rotello, C. M., Macmillan, N. A., Reeder, J. A., & Wong, M. (2005). The remember response: Subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin and Review*, *12*, 865–873.

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type 1 error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics*, *70*, 389–401.

Rubin, D. C. & Friendly, M. (1986). Predicting which words get recalled – measures of free-recall, availability, goodness, emotionality, and pronounciability for 925 nouns. *Memory & Cognition*, *14*, 79–94.

Sanna, L. J., Schwarz, N., & Small, E. M. (2002). Accessibility experiences and the hindsight bias: I knew it all along versus it could never have happened. *Memory & Cognition*, *30*, 1288–1296.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 1–17.

Schacter, D. L. (2001). *The seven sins of memory: How the mind forgets and remembers*. Boston, MA: Houghton, Mifflin and Company.

Schacter, D. L. & Tulving, E. (1994). What are the memory systems of 1994? In D. L. Schacter & E. Tulving (Eds), *Memory systems 1994* (pp.1–38). Cambridge: MIT Press.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Science*, *4*, 299–310.

Schraw, G. (1995). Measures of feeling-of-knowing accuracy: a new look at an old problem. *Applied Cognitive Psychology*, *9*, 321–332.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*, 33–45.

Schneider, W., & Pressley, M. (1997). *Memory development between two and twenty*. Mahwah, NJ: Erlbaum.

Schwartz, B. L. & Metcalfe, J. M.(1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1074–1083.

Schwarz, N. (2004). Meta-cognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology*, *14*, 332–348.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, *61*, 195–202.

Scoboria, A., Mazzoni, G., Kirsch, I., & Relya, M. (2004). Plausibility and belief in autobiographical memory. *Applied Cognitive Psychology*, *18*, 791–807.

Scott, R. B., & Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology-Learning Memory and Cognition*, *34*, 1264–1288.

Scott, R. B., & Dienes, Z. (2009). The metacognitive role of familiarity in artificial grammar learning: transitions from unconscious to conscious knowledge. In A. Efklides & P. Misailidi (Eds), *Trends and prospects in metacognition research*. New York, NY: Springer.

Scott, R. B., & Dienes, Z. (2010a). Fluency does not express implicit knowledge of artificial grammars. *Cognition, 114 (3),* 372–388.

Scott, R. B., & Dienes, Z. (2010b). Knowledge applied to new domains: the unconscious succeeds where the unconscious fails. *Consciousness and Cognition, 19,* 391–398.

Scott, R. B., & Dienes, Z (2010c). The influence of prior familiarity on implicit knowledge and the conscious status of structural knowledge. *Consciousness and Cognition, 19,* 413–418.

Seamon, J. G., Luo, C. R., Schlegel, S. E., Greene, S. E., & Goldenberg, A. B. (2000). False memory for categorized pictures and words: The category associates procedure for studying memory errors in children and adults. *Journal of Memory and Language*, *42*, 120–146.

Seger, C. A. (1994). Implicit learning. *Psychological Bulletin*, *115*, 163–196.

Serra, M. J., & Metcalfe, J. (2009). Effective implementation of metacognition. In D. Hacker, J. Dunlosky, & A. Graesser (Eds), *Handbook of metacognition in education* (pp. 278–298). NY, NY: Rutledge.

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 592–608.

Shah, A. K., & Oppenheimer, D. (2008). Heuristics made easy: An effort reduction framework. *Psychonomic Bulletin*, *134*, 207–222.

Shanks, D. R. (2005). Implicit Learning. In K. Lamberts & R. Goldstone (Eds), *Handbook of cognition* (pp. 202–220). London, UK: Sage.

Siegel, L. S. & Ryan, E. B. (1989). The development of working memory in normally achieving and subtypes of learning disabled children. *Child Development, 60*, 973–980.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99–118.

Singer, M. (2009). Strength-based criterion shifts in recognition memory. *Memory & Cognition, 37*, 976–984.

Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory & Cognition, 34*, 125–137.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592–604.

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin, & D. Kahneman (Eds), *Intuitive judgment: Heuristics and biases* (pp. 397–420). New York, NY: Cambridge University Press.

Sluzenski, J., Newcombe, N., & Ottinger, W. (2004). Changes in reality monitoring and episodic memory in early childhood. *Developmental Science, 7*, 225–245.

Smallwood, J. & Schooler, J. W. (2006). The restless mind. *Psychological Bulletin, 132*, 946–958.

Smith, S. M., Ward, T. B., Tindell, D. R., Sifonis, C. M., & Wilkenfeld, M. J. (2000). Category structure and created memories. *Memory & Cognition, 28*, 386–395.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117, 34–50.*

Sommer, W., Heinz, A., Leuthold, H., Matt, J., & Schweinberger, S. R. (1995). Metamemory, distinctiveness, and event-related potentials in recognition memory for faces. *Memory & Cognition, 23*, 1-11.

Son, L. K., & Kornell, N. (2008). Research on the allocation of study time: Key studies from 1890 to the present (and beyond). In J. Dunlosky & R. A. Bjork (Eds), *Handbook of metamemory and memory*, (pp. 333–351). New York, NY: Psychology Press.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*, 204–221.

Song, H., & Schwarz, N. (2009). If it's difficult to pronounce, it must be risky: Fluency, familiarity, and risk perception. *Psychological Science, 20*, 135–138.

Souchay, C. (2007). Metamemory in Alzheimer's disease. *Cortex, 43*, 987–1003.

Souchay, C., Isingrini, M., & Gil, R. (2002). Alzheimer's disease and feeling-of-knowing in episodic memory. *Neuropsychologia, 40*, 2386–2396

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, & Law, 7*, 36–97.

Starns, J. J., Lane, S. M., Alonzo, J. D., & Roussel, C. C. (2007). Metamnemonic control over the discriminability of memory evidence: A signal-detection analysis of warning effects in the associative list paradigm. *Journal of Memory and Language, 56*, 592–607.

Strack, F., & Forster, J. (1995). Reporting recollective experience: Direct access to memory systems? *Psychological Science, 6*, 352–358.

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, and Cognition, 24*, 1379–1396.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24*, 1397–1410.

Thapar, A., & Westerman, D. L. (2009). Aging and fluency based illusions of recognition memory. *Psychology and Aging*, *24*, 595–603.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-paced study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1024–1037.

Titus, T. G. (1973). Continuous feedback in recognition memory. *Perceptual and Motor Skills*, *37*, 771–776.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*, 1-12.

Tulving, E. & Schacter, D. L. (1990). Priming and human memory systems. *Science*, *247*, 301–306.

Tunney, R. J., & Bezzina, G. (2007). Effects of retention intervals on receiver operating characteristics in artificial grammar learning. *Acta Psychologica*, *125*, 37–50.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Uleman, J. S., Saribay, S. A., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology*, *59*, 329–360.

Unkelbach, C. (2006). The learned interpretation of cognitive fluency. *Psychological Science*, *17*, 339–345.

Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 219–230.

Vaihinger, H. (1911/1924). *The Philosophy of 'As If': A system of the theoretical, practical and religious fictions of mankind* (translation by C. K. Ogden). London, UK: Routledge and Kegen Paul.

Van Overschelde, J. P., (2008). Metacognition: Knowing about knowing. In D. Hacker, J. Dunlosky, & A. Graesser (Eds), *Handbook of metacognition in education* (pp. 47–71). New York, NY: Psychology Press.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 582–600.

Verde, M. F., & Rotello, C. M. (2004). ROC curves show that the revelation effect is not a single phenomenon. *Psychonomic Bulletin & Review*, *11*, 560–566.

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition*, *35*, 254–262.

Verfaellie, M., and Giovanello, K. S. (2006). Conceptual priming in semantic dementia: A window into the cognitive and neural basis of conceptual implicit memory. *Cognitive Neuropsychology*, *23*, 606–620.

Vokey, J. R., & Higham, P. A. (2005). Abstract analogies and positive transfer in artificial grammar learning. *Canadian Journal of Experimental Psychology*, *59*, 54–61.

Wagner, A. D., & Gabrieli, J. D. E. (1998). On the relationship between recognition familiarity and perceptual fluency: Evidence for distinct mnemonic processes. *Acta Psychologica*, *98*, 211–230.

Wais, P. E., Mickes, L., & Wixted, J. T. (2008). Remember/know judgments probe degrees of recollection. *Journal of Cognitive Neuroscience*, *20*, 400–505.

Wallace, W. P. (1982). Distractor-free recognition tests of memory. *American Journal of Psychology*, *95*, 421–440.

Wallsten, T. S., & Budescu, D. V. (1983). Encoding subject probabilities: A psychological and psychometric review. *Management Science*, *29*, 151–173.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science*, *39*, 176–190.

Wan, L. L., Dienes, Z., & Fu, X. L. (2008). Intentional control based on familiarity in artificial grammar learning. *Consciousness and Cognition*, *17*, 1209–1218.

Watkins, M. J., & Gibson, J. M. (1988). On the relation between perceptual priming and recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *14*, 477–483.

Watkins, M. J., & Peynircioglu, Z. F. (1990). The revelation effect: When disguising test items induces recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *16*, 1012–1020.

Weaver, C. A., III, & Kelemen, W. L. (2003). Processing similarity does not improve metamemory: Evidence against transfer-appropriate monitoring. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29*, 1058 –1065.

Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, *60*, 53–85.

Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *Journal of Applied Psychology*, *3*, 490–499.

Weingardt, K. R., Leonesio, R. J., & Loftus, E. F., (1994). Viewing eyewitness research from a metacognitive perspective. In J. Metcalfe & A. P. Shimamura (Eds), *Metacognition: Knowing about knowing* (pp. 157–184). Cambridge: MIT Press.

Westerman, D. L. (2008). Relativity and fluency-based illusions of recognition memory. *Psychonomic Bulletin & Review*, *15*, 1196–1200.

Westerman, D. L., & Greene, R. L. (1996). On the generality of the revelation effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *22*, 1147–1153.

Westerman, D. L., Lloyd, M. E., & Miller, J. K. (2002). The attribution of perceptual fluency in recognition memory: The role of expectation. *Journal of Memory and Language*, *47*, 607–617.

Westerman, D. L., Miller, J. K., & Lloyd, M. E. (2003). Change in perceptual form attenuates use of the fluency heuristic. *Memory and Cognition*, *31*, 619–629.

Whittlesea. B. W. A. (1987). Preservation of specific experiences in the representation of general knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 3–17.

Whittlesea, B. W. A. (1993). Illusions of Familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1235–1253.

Whittlesea, B. W. A. (1997). Production, evaluation, and preservation of experiences: Constructive processing in remembering and performance tasks. In D. L. Medlin (Ed.), *Psychology of learning and motivation* (Vol. 37, pp. 211–264). New York: Academic Press.

Whittlesea, B. W. A. (2002). False memory and the discrepancy-attribution hypothesis: The prototype-familiarity illusion. *Journal of Experimental Psychology: General*, *131*, 96–115.

Whittlesea, B. W. A. (2003). On the construction of behavior and subjective experience: The production and evaluation of performance. In J. S. Bowers & C. J. Marsolek (Eds), *Rethinking implicit memory* (pp. 239–260). New York: Oxford University Press.

Whittlesea, B. W. A. (2002). Two routes to remembering (and another to remembering not). *Journal of Experimental Psychology: General*, *131*, 325–348.
Whittlesea, B. W. A. (2004). The perception of integrality: Remembering through the validation of expectation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 891–908.
Whittlesea, B. W. A., & Dorken, M. D. (1997). Implicit learning: Indirect, not unconscious. *Psychonomic Bulletin and Review*, *4*, 63–67.
Whittlesea, B. W. A., Jacoby, L. L., & Girard, K. (1990). Illusions of immediate memory: Evidence of an attributional basis for feelings of familiarity and perceptual quality. *Journal of Memory & Language*, *29*, 716–732.
Whittlesea, B. W. A., & Leboe, J. P. (2000). The heuristic basis of remembering and classification: Fluency, generation, and resemblance. *Journal of Experimental Psychology: General*, *129*, 84–106.
Whittlesea, B. W. A., & Leboe, J. P. (2003). Two fluency heuristics (and how to tell them apart). *The Journal of Memory and Language*, *49*, 62–79.
Whittlesea, B. W. A., Masson, M. E. J., & Hughes, A. D. (2005). False memory following rapidly presented lists: The element of surprise. *Psychological Research*, *69*, 420–430.
Whittlesea, B. W. A., & Price, J. R. (2001). Implicit/explicit memory versus analytic/nonanalytic processing: Rethinking the mere exposure effect. *Memory & Cognition*, *29*, 234–246.
Whittlesea, B. W. A., & Williams, L. D. (1998). Why do strangers feel familiar, but friends don't? The unexpected basis of feelings of familiarity. *Acta Psychologica*, *96*, 141–165.
Whittlesea, B. W. A., & Williams, L. D. (2000). The source of feelings of familiarity: The discrepancy-attribution hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 547–565.
Whittlesea, B. W. A., Williams L. D. (2001a). The discrepancy-attribution hypothesis: I. The heuristic basis of feelings and familiarity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 27, 3-13.
Whittlesea, B. W. A., Williams L. D. (2001b). The discrepancy attribution hypothesis: II. Expectation, uncertainty, surprise, and feelings of familiarity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 27, 14–33.
Whittlesea, B. W. A., & Wright, R. L. (1997). Implicit (and explicit) learning: Acting adaptively without knowing the consequences. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *23*, 181–200.
Wilken, P., Bayne, T., & Cleeremans, A. (Eds), (2008). *The Oxford companion to consciousness*. Oxford, UK: Oxford University Press.
Willems, S., Grandjean, J. Steaniak, N., & Van der Liden, M. (submitted). The evolution of the fluency heuristic in amnesia.
Winkielman, P., Schwarz, N., Fazendeiro, T., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch & K. C. Klauer (Eds), *The psychology of evaluation: Affective processes in cognition and emotion* (pp.189–217). Mahwah, NJ: Erlbaum.
Winne, P. H. & Hadwin, A. F. (1998). Studying as self-regulated learning. In Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds), *Metacognition in educational theory and practice*. (pp. 277–304). Hillsdale, NJ: Erlbaum.
Witherspoon, D., & Allan, L. G. (1985). The effects of prior presentation on temporal judgments in a perceptual identification task. *Memory & Cognition*, *13*, 101–111.
Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 681–690.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*, 152–176.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin and Review*, *11*, 616–641.

Woll, S. B. & Graesser, A. C. (1982). Memory discrimination for information typical and atypical of person schemata. *Social Cognition*, *1*, 287–310.

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, *110*, 611–617.

Yano, M., Satoshi, U., & Mimura, M. (2008). Preserved priming but insensitivity to perceptual fluency on recognition judgments in Alzheimer's Disease. *Psychogeriatrics*, *8*, 178–187.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, *30*, 132–156.

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. (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. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517.

Zimmerman B. J., Greenspan D., & Weinstein C. E. (1994). Self-regulating academic study time: A strategy approach. In D. H. Schunk & B. J. Zimmerman (Eds), *Selfregulation of learning and performance: Issues and educational applications* (pp.181–99). Hillsdale, NJ: Erlbaum.

Zacks, R. T. (1969). Invariance of total learning time under different conditions of practice. *Journal of Experimental Psychology*, *82*, 441–447.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*, 151–175.

Index