SECOND EDITION

Cognitive Illusions

Intriguing phenomena in thinking, judgment and memory

Edited by **Rüdiger F. Pohl**

A **Psychology Press** Book

eRESOURCES

ROUTLEDGE

COGNITIVE ILLUSIONS

Cognitive Illusions explores a wide range of fascinating psychological effects in the way we think, judge and remember in our everyday lives. Featuring contributions from leading researchers, the book defines what cognitive illusions are and discusses their theoretical status: Are such illusions proof for a faulty human information-processing system, or do they only represent by-products of otherwise adaptive cognitive mechanisms? Throughout the book, background to phenomena such as illusions of control, overconfidence and hindsight bias are discussed, before considering the respective empirical research, potential explanations of the phenomenon and relevant applied perspectives. Each chapter also features the detailed description of an experiment that can be used as a classroom demonstration.

Featuring six new chapters, this edition has been thoroughly updated throughout to reflect recent research and changes of focus within the field.

This book will be of interest to students and researchers of cognitive illusions, specifically, those focusing on thinking, reasoning, decision making and memory.

Rüdiger F. Pohl is Professor of Psychology at the University of Mannheim, Germany. His research interests include cognitive illusions, heuristics and decision making, and autobiographical memory.

This page intentionally left blank

COGNITIVE ILLUSIONS

Intriguing phenomena in thinking, judgment and memory

Second Edition

Edited by Rüdiger F. Pohl

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2017
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2017 selection and editorial matter, Rüdiger F. Pohl; individual chapters, the contributors

The right of the editor to be identified as the author of the editorial material, and of the authors for their individual chapters, has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record has been requested for this book

ISBN: 978-1-138-90341-8 (hbk)
ISBN: 978-1-138-90342-5 (pbk)
ISBN: 978-1-315-69693-5 (ebk)

Typeset in Bembo
by Apex CoVantage, LLC

Visit the eResources website: https://www.routledge.com/products/9781138903425

CONTENTS

Contributors *viii*

Introduction **1**

Cognitive illusions 3
Rüdiger F. Pohl

PART I
Thinking **23**

1 Conjunction fallacy 25
John E. Fisk

2 Base-rate neglect 44
Gordon Pennycook and Valerie A. Thompson

3 Probability matching 62
Ben R. Newell and Christin Schulze

4 Framing 79
Anton Kühberger

5 Confirmation bias – myside bias 99
Hugo Mercier

6 Illusory correlation 115
Klaus Fiedler

7 Illusions of control 134
Suzanne C. Thompson

8 Wason selection task 150
Jonathan St. B. T. Evans

9 Belief bias in deductive reasoning 165
Jonathan St. B. T. Evans

PART II
Judgment **183**

10 Availability 185
Rolf Reber

11 Judgments by representativeness 204
Karl H. Teigen

12 Anchoring effect 223
Štěpán Bahník, Birte Englich, and Fritz Strack

13 Validity effect 242
Catherine Hackett Renner

14 Mere exposure effect 256
Robert F. Bornstein and Catherine Craver-Lemley

15 Halo effects 276
Joseph P. Forgas and Simon M. Laham

16 Overconfidence 291
Ulrich Hoffrage

17 Pollyanna Principle 315
Margaret W. Matlin

PART III
Memory **337**

18 Revelation effect 339
André Aßfalg

19 Survival processing effect 357
Meike Kroneisen and Edgar Erdfelder

20 Labelling and overshadowing effects 373
Rüdiger F. Pohl

21 Associative memory illusions 390
Henry L. Roediger, III and David A. Gallo

22 Misinformation effect 406
Jacqueline E. Pickrell, Dawn-Leah McDonald, Daniel M. Bernstein, and Elizabeth F. Loftus

23 Hindsight bias 424
Rüdiger F. Pohl and Edgar Erdfelder

24 Retrieval-induced forgetting 446
Oliver Kliegl and Karl-Heinz T. Bäuml

PART IV
Extensions **465**

25 Suggestions and cognitive illusions 467
Günter Molz and Rüdiger F. Pohl

Author Index *485*
Subject Index *493*

CONTRIBUTORS

André Aßfalg
Albert-Ludwigs-University, Freiburg, Germany

Karl-Heinz T. Bäuml
Regensburg University, Regensburg, Germany

Štěpán Bahník
University of Würzburg, Würzburg, Germany

Daniel M. Bernstein
Kwantlen Polytechnic University, Surrey, British Columbia, Canada

Robert F. Bornstein
Adelphi University, Garden City, NY, USA

Catherine Craver-Lemley
Elizabethtown College, Elizabethtown, PA, USA

Birte Englich
University of Cologne, Cologne, Germany

Edgar Erdfelder
University of Mannheim, Mannheim, Germany

Jonathan St. B. T. Evans
University of Plymouth, Plymouth, Great Britain

Klaus Fiedler
University of Heidelberg, Heidelberg, Germany

John E. Fisk
University of Central Lancashire, Preston, Great Britain

Joseph P. Forgas
University of New South Wales, Sydney, Australia

David A. Gallo
University of Chicago, Chicago, IL, USA

Catherine Hackett Renner
UnityPoint Health, Des Moines, IA, USA

Ulrich Hoffrage
University of Lausanne, Lausanne, Switzerland

Oliver Kliegl
Regensburg University, Regensburg, Germany

Meike Kroneisen
University of Mannheim, Mannheim, Germany

Anton Kühberger
University of Salzburg, Salzburg, Austria

Simon M. Laham
University of Melbourne, Melbourne, Australia

Elizabeth F. Loftus
University of California, Irvine, CA, USA

Margaret W. Matlin
SUNY Geneseo, Geneseo, NY, USA

Dawn-Leah McDonald
Kwantlen Polytechnic University, Surrey, British Columbia, Canada

Hugo Mercier
University of Neuchâchtel, Neuchâtel, Switzerland

Günter Molz
Wuppertal University, Wuppertal, Germany

Ben R. Newell
University of New South Wales, Sydney, Australia

Gordon Pennycook
University of Waterloo, Waterloo, ON, Canada

Jacqueline E. Pickrell
University of Washington, Seattle, WA, USA

Rüdiger F. Pohl
University of Mannheim, Mannheim, Germany

Rolf Reber
University of Oslo, Oslo, Norway

Henry L. Roediger, III
Washington University in St. Louis, St. Louis, MO, USA

Christin Schulze
Max Planck Institute for Human Development, Berlin, Germany

Fritz Strack
University of Würzburg, Würzburg, Germany

Karl H. Teigen
University of Oslo, Oslo, Norway

Suzanne C. Thompson
Pomona College, Claremont, CA, USA

Valerie A. Thompson
University of Saskatchewan, Saskatoon, SK, Canada

Introduction

This page intentionally left blank

COGNITIVE ILLUSIONS

Rüdiger F. Pohl

Errare humanum est.

(Cicero, 116–43 BC)

That we as humans do make errors in thinking, judgment, and memory is undisputed. In fact, there is a plethora of phenomena showing that we deviate in our thinking, judgment, and memory from some objective and arguably "correct" standard. However, trying to provide a complete list of these biases and fallacies would be a tedious task. I therefore restricted the selection of cognitive illusions in this book to those that are more well known and that have inspired a wealth of empirical studies and theoretical approaches in order to better understand the underlying mechanisms of the human information-processing system or to dispute their status as a "bias" by questioning the normative standards of comparison and the applied experimental methods. This introductory chapter gives an overview about cognitive illusions in general, discusses their theoretical status, and describes the intentions of this book.

The domain of cognitive illusions

I start this section by asking "What are the defining features of a cognitive illusion?" Then I present the categorization and selection of illusions in this book and briefly review other collections of such illusions. This should give the reader a first impression of the domain in question.

Defining features

The term "cognitive illusion" has evolved in analogy to the better known domain of "optical illusions" (see Roediger, 1996). The first and main feature of a phenomenon

to count as an illusion thus is that it leads to a perception, judgment, or memory that reliably *deviates* from "reality." In cases of optical and memory illusions, it may be immediately evident what constitutes reality (because subjective perception and recall can be compared to external or original stimuli, respectively), but in thinking and judgment, the matter is less clear (Gigerenzer, 1996). The problem concerns how to define an objectively "correct" judgment or decision (see below).

As a second criterion, the observed phenomenon needs to deviate from the normative standard in a *systematic* fashion (i.e., in a predictable direction) rather than just randomly. Therefore, most designs include a control group, assuming that any deviations in the control group's data result from random error alone, while the experimental group shows in addition a systematic effect. Special care has to be taken if demonstrations of an illusion depend on repeated measures because it involves the danger of being prone to regression effects, which could possibly lead to false interpretations of the data (as examples, see Chapters 6 and 23). As a related feature, the mechanisms eventually leading to cognitive illusions typically include a number of probabilistic processes so that an illusion will not necessarily be observed on each and every single trial, but may only become evident as a systematic bias if the data are summed across a larger number of trials or participants.

A third aspect of cognitive illusions is that they appear *involuntarily*, that is, without specific instructions or deliberate will. They just happen. This is analogous to what has been found in research on suggestions (see Chapter 25): The suggested reaction manifests itself in the given situation without any conscious decision to do so. This does not mean that motivational factors or conscious meta-cognitions may not be influential, too, but they are not the ultimate cause of the illusion itself. They only moderate its size (see Pohl, Bender, & Lachmann, 2002, for further discussion). Another aspect is that persons who have fallen prey to a cognitive illusion usually don't realize what has happened: "Illusions mock our belief that what we perceive, remember, and know is in perfect accord with the state of the external world." (Roediger, 1996, p. 76) That is, illusioned persons are still convinced to have judged, decided, or recalled something to the best of their knowledge.

As a consequence, and this constitutes the fourth cornerstone of the proposed definition, an illusion is hard if not *impossible to avoid*. While this is probably true for all optical illusions, the criterion is much weaker for cognitive ones. For some illusions, a proper instruction, careful selection of the material, or other procedural variations may reduce or even eliminate the illusion (as an example, see Gigerenzer, Hertwig, Hoffrage, & Sedlmeier, 2008), while for other illusions, most (if not all) attempts to overcome the effect have failed (as an example, see Pohl & Hell, 1996, and Chapter 23).

And finally, as the fifth point to consider and to distinguish cognitive illusions from other forms of typical errors, misunderstandings, or faulty memories, illusions often appear as rather *distinct* from the normal course of information processing. An illusion somehow "sticks out" as something special that "piques our curiosity" (as Roediger, 1996, put it) and thus attracts researchers to explain this unexpected but robust finding. In other words, ordinary forms of forgetting (leading to omission

errors), drawing schema-based inferences (leading to commission errors), or deviations resulting from simple misunderstandings would not be considered "illusions." Roediger and McDermott (2000) accordingly described such phenomena as "distortions." This is not to say that a cognitive illusion cannot be explained with ordinary and general mechanisms of information processing. In fact, one of the theoretical goals of research on cognitive illusions is to avoid the assumption of any special mechanism that is responsible only for this one phenomenon, but instead to explain the observed effects with what one already knows about cognitive processes in general.

Categories of cognitive illusions

I ordered the list of cognitive illusions that I came up with when I planned this book into three categories, namely, thinking, judgment, and memory. While the *memory* category may appear rather clear (something is remembered but deviates in a systematic way from the original), the distinction between *thinking* and *judgment* may be less sharp. This becomes immediately evident if one looks at earlier collections of cognitive illusions (see below): These generally focussed either on memory alone or both on thinking and judgment, without clearly differentiating the latter two. And indeed, categorizing illusions as being related to either thinking or judgment appears rather difficult. But to make things even worse, *all* illusions involve memory processes like encoding, storage, and retrieval. For some of the judgment illusions, for example, varying the retention interval between experimental phases led to clear effects. On the other hand, memory illusions involve different kinds of thinking and judgment processes. For example, the three classical heuristics proposed by Tversky and Kahneman (1974) are discussed as responsible not only for biases in judgment but also for several memory illusions.

I defined illusions of *thinking* (Chapters 1 to 9) as those that involve application of a certain rule (like Bayes' theorem, hypothesis testing, or syllogistic reasoning). These rules are derived from normative models (like probability theory, falsification principle, or logic) and their results usually serve as standards against which human performance is evaluated. Typical tasks are to estimate a probability, to verify a logical conclusion, or to discover a hidden rule.

In illusions of *judgment* (Chapters 10 to 17), participants are asked to subjectively rate a specific aspect of a given stimulus (e.g., its pleasantness, frequency, or veracity). However, specific features of the situation may bias someone's judgment in a certain direction. These features involve, for example, feelings of familiarity or confidence, the subjective experience of searching one's memory, or the selective activation of one's memory contents. Most importantly, these are all cases of judgments under uncertainty, that is, the participant is bound to rely on subjective impressions only. The classical heuristics described by Tversky and Kahneman (1974) – availability, representativeness, and anchoring (Chapters 10 to 12, respectively) – represent examples of such mechanisms that may in turn lead to judgmental illusions (see the discussion below). In this sense, these heuristics have a different status than

the phenomena described in the other chapters. They are not cognitive illusions themselves but might be involved as cognitive processes in several different illusions.

Finally, illusions of *memory* (Chapters 18 to 24) are those in which earlier encoded material has to be remembered later on. The critical test typically involves recall or recognition. I restricted this category to cases where presentation of the original material proceeds under experimental control. In other cases, the original material might be encoded pre-experimentally outside the laboratory (see, e.g., Jones & Martin, 2004; Park & Reder, 2004).

Earlier collections

Of course, this collection of cognitive illusions is not the first one to appear. One such comprehensive collection with examples from thinking, judgment, and memory was published in Germany 20 years ago by Hell, Fiedler, and Gigerenzer (1993). In addition to some of the illusions included here, Hell et al. also covered problems of attributing causality and developmental misconceptions of physics. Other collections focussed on only a few illusions that were related either to both thinking and judgment or to memory alone and are described below.

Thinking and judgment

Probably one of the most cited and now "classical" papers on biases in judgment is the 1974 *Science* paper of Tversky and Kahneman entitled "Judgment under uncertainty: Heuristics and biases." The authors presented three general heuristics, namely, *availability*, *representativeness*, and *adjustment and anchoring*, together with several examples of each. A collection of some high-impact papers in this domain was later edited by Kahneman, Slovic, and Tversky (1982) under the same title as the 1974 paper. More recent developments of this approach are documented in "Heuristics and biases: The psychology of intuitive judgment," edited by Gilovich, Griffin, and Kahneman (2002). The three major sections of that volume are devoted to (1) theoretical and empirical extensions (including representativeness and availability; anchoring, contamination, and compatibility; forecasting, confidence, and calibration; optimism; and norms and counterfactuals), (2) new theoretical directions (including two systems of reasoning, support theory, and alternative perspectives on heuristics), and (3) real-world applications (including everyday judgment and behavior, and expert judgment).

Another collection, entitled "On cognitive illusions and their implications," was published by Edwards and von Winterfeld (1986). They explored four kinds of intellectual tasks: (1) probability assessments and revision, (2) decision making, (3) intuitive physics, and (4) logic and mental arithmetic. Similarly, Caverni, Fabre, and Gonzalez (1990a) looked at "Cognitive biases" related to (1) different cognitive activities (like reasoning and problem solving, categorization, assessment, and judgments of probability and confidence), (2) characteristics of the situation (as given by the context and the external structure of the information), and (3) possible cognitive aids to correct for biases. Still other treatises of biases in human reasoning and

thinking were provided by Evans (1989), Gilovich (1991), and Piatelli-Palmarini (1994). Later, Krueger and Funder (2004) listed 42 and Evans (2007) even 51 different biases. Stanovich (2008) suggested a taxonomy of such biases and classified them according to his dual-process approach (see further below) into four categories, namely, biases caused by (1) cognitive miserliness, (2) override failure (i.e., System 2 not overriding System 1), (3) mindware gaps (like missing knowledge), and (4) contaminated mindware (like egocentric thinking).

Memory

The earliest book including "memory illusions" that I know of was published by Sully in 1881. He described several illusive phenomena of introspection, perception, memory, and belief under a rather vague definition: "Illusion, as distinguished from correct knowledge, is to put it broadly, deviation of representation from fact." (p. 332)

A few years later, Hodgson and Davey (1886/1887) provided a noteworthy collection of memory errors (also called "illusions" there). They rigorously investigated spiritualism in the form of "psychography" (i.e., writing on a hidden slate without any operation of the medium's muscles). That this "clever conjuring trick" was not detected by the devoted followers in such séances was explained mainly through errors of perception and memory. More specifically, the authors described memory errors of omission, substitution, transposition, and interpolation.

Another noteworthy but more well-known contribution is Bartlett's (1932) book on "Remembering: A study in experimental and social psychology," describing and experimentally investigating several sources of schematic influences on false recall. Yet, the effects of schematic knowledge on encoding and reconstructing memories (see Alba & Hasher, 1983; Brewer & Nakamura, 1984, for summaries) never came to be considered as "cognitive illusions," perhaps because they represent everyday experiences and thus lack the flavor of a special phenomenon. But the cognitive mechanisms proposed by Bartlett and his successors nevertheless appear in modern explanations of some "true" cognitive illusions.

Much later, Schacter, Coyle, Fischbach, Mesulam, and Sullivan (1995) edited a book on "Memory distortion: How minds, brains, and societies reconstruct the past," and Roediger (1996) introduced a special issue of the *Journal of Memory and Language* focussing on "Memory illusions." The collection of papers in that issue was not intended to systematically and completely cover all known types of illusions, but it nevertheless gave a good overview of the wide range of memory distortions and illusions. Roediger discussed the included papers under the following topics (which are also covered in several chapters of the present volume): Fluency illusions, relatedness effects, verbal overshadowing, effects of interference and misleading information, and illusions of reality and source monitoring. In addition, the issue treated a number of phenomena not explicitly covered here: Illusions of perception and memory, illusory conjunctions and memory, and hypnosis and guessing effects (which are related to effects of suggestion and suggestibility; see Chapter 25).

In the "Oxford handbook of memory," Roediger and McDermott (2000) presented a more systematic coverage of "memory distortions." The two general classes of memory errors, omission and commission, were reviewed according to six factors that seem to be responsible for their occurrence: "False memories arise from inferences from series of related pieces of information, from interference from events surrounding the event of interest, from imagination of possible events that did not occur, from retrieval processes, and from social factors. Finally, there are individual differences in susceptibility to these memory illusions." (p. 160) In closing their overview, Roediger and McDermott expressed their hope that from studying memory illusions "we can elucidate both the nature of these curious and interesting phenomena, but also shed light on processes occurring in 'normal' remembering of events." (p. 160)

The debate on cognitive illusions

In this section, I summarize the debate on the so-called "heuristics and biases program" of Tversky and Kahneman (1974). Then I discuss subsequent developments and the theoretical status of cognitive illusions in general.

The debate on heuristics and biases

The underlying idea of Tversky and Kahneman (1974) was that humans employ a small number of simple and quick rules of thumb in many different situations of judgment under uncertainty. With respect to their efficiency, Tversky and Kahneman stated that "in general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors." (p. 1124) Biases caused by these heuristics were thought to represent genuine cognitive strategies of the human information-processing system independent from any motivational influence: "These biases are not attributable to motivational effects such as wishful thinking or the distortion of judgments by payoffs and penalties." (p. 1130) Several of the observed biases and fallacies were thus thought to be explainable with only a few general heuristics. This "heuristics and biases" approach led to an enormous number of studies investigating the intricacies of human judgment (see, e.g., Kahneman et al., 1982, or later, Gilovich et al., 2002), which in turn affected scholarship in economics, law, medicine, management, and political science.

However, this approach led to some controversy. Especially Gigerenzer (1991, 1996, 2004; Gigerenzer et al., 2008) criticized the "heuristics and biases" program for having "narrow norms and vague heuristics" (see also the replies of Gilovich & Griffin, 2002; Kahneman & Tversky, 1996). The criticism centred mainly around three topics, namely, one-sided view, artificiality, and lack of explanatory power.

One-sided view

According to Gigerenzer, the collection of biases has focussed too much on the few faulty cases of judgment and decision making thereby ignoring the majority of cases where heuristics typically lead to correct or at least useful decisions. This one-sided

view may have led some researchers to conclude that "mental illusions should be considered the rule rather then the exception" (Thaler, 1991, p. 4) and that "mistakes of reason rule our minds" (Piatelli-Palmarini, 1994). But this view of human rationality appears overly pessimistic (see the discussion further below).

Artificiality

Gigerenzer (1991) also asserted that not everything that looked like a cognitive illusion really is one. More specifically, he argued that one could make cognitive illusions disappear simply by (a) avoiding too narrow norms to evaluate human performance, by (b) using an adequate statistical format, and by (c) using representative (instead of selected) samples of items.

With respect to the applied norm, Gigerenzer (1991) asserted that probability theory is about frequencies and thus not applicable to judgments of single events (as is the typical task in studies on cognitive illusions). Rather, norms should concern degrees of subjective belief (confidence) in single events (see further below). Using unnatural and thus inadequate statistical formats (e.g., probabilities instead of natural frequencies) and selective (instead of representative) sampling could be other sources for artificial "biases" (see Hertwig & Ortmann, 2005). For example, if an experimenter selects a biased sample of materials (that is not representative of the respective domain) and then shows that thinking, judgment, or decision making is "biased," this just demonstrates that the experimenter was successful in fooling his participants, but not that the underlying cognitive processes are biased in any way. Using adequate statistical formats and representative sampling, Gigerenzer (1991; see also Gigerenzer & Hoffrage, 1995) accordingly showed that some illusions could be substantially reduced or even eliminated, namely, the conjunction fallacy (see Chapter 1), base-rate neglect (see Chapter 2), and overconfidence (see Chapter 16). In a later paper, Gigerenzer (2004) added more phenomena to this list, asserting that all of them could be explained with an "unbiased" mind, that is, without reference to a faulty information-processing system. Instead, environmental structures (of the world or the experimental study) would be responsible for the observed effects. In addition to unrepresentative sampling and inadequate statistical formats already mentioned above, he added as further artificial causes regression toward the mean (see also Fiedler & Krueger, 2011), unequal sample sizes, actually skewed distributions, pragmatic inferences, probability matching, and missing knowledge. A further source for artificial or inflated "biases" could be that many studies used openly available material ("decisions from descriptions") rather than material retrieved from memory ("decisions from experience") which could arguably be seen as more natural (Hertwig, 2012). Hilbert (2012) added that several biases in decision making could be the product of noisy deviations in memory-based information processing and that this "noisiness" could be, moreover, the unifying mechanism underlying many such distortions.

Gigerenzer et al. (2008), however, also admitted that the accusation of artificiality applied only to some illusions, while others are "true illusions" without

doubt. In addition, the proposed methodological remedies generally only *reduced* the respective illusion, so that these illusions still need to be explained apart from any additional (possibly artificial) factor that may have inflated some of the findings of the heuristics and biases program.

Lack of explanatory power

With respect to their theoretical status, Gigerenzer (1991) claimed that the postulated heuristics are not explanations at all, but rather redescriptions of the observed phenomena: "In the 20 years of 'heuristics and biases' research [. . .], a lack of theoretical progress is possibly the most striking result. The absence of a general theory or even of specific models of underlying cognitive processes has been repeatedly criticized [. . .], but to no avail." (p. 101) Similarly, on behalf of the occasion that Daniel Kahneman was awarded the Nobel prize for economics in 2002, one congratulator (MacCoun, 2002) noted that "their [Kahneman and Tversky's] core arguments make only sparing use of hypothetical constructs" and that their papers "rely less on elaborate experimentation and psychometrics than on simple but powerful demonstrations of replicable phenomena." (p. 8) This lack of rigorous investigation and detailed modelling is what Gigerenzer so heavily criticized.

Subsequent developments

Following the by now classical dispute on "heuristics and biases," this section describes some subsequent trends. These concern the fast and frugal heuristics suggested by Gigerenzer and his group, dual-process models, individual differences, and applied perspectives. (A further area, neuropsychological foundations, also appears promising, but is beyond the scope of this introduction. Several of the chapters in this book include according research findings.)

Fast and frugal heuristics

As an alternative to the classic heuristics, Gigerenzer, Todd, and The ABC Research Group (1999) proposed a set of "fast and frugal" heuristics that were fundamentally different from the classic ones. They were more clearly defined, were supposed to be ecologically valid, and could be implemented and tested in computer simulations (see Gigerenzer, Hertwig, & Pachur, 2011; Marewski & Schooler, 2011; Todd, Gigerenzer, & The ABC Research Group, 2012). Gigerenzer et al. (1999) introduced these heuristics in a book entitled "Simple heuristics that make us smart." As the title implies, the authors' focus was less on errors and fallacies, but rather on how such heuristics may lead to good decisions under natural time and knowledge constraints, thus endorsing a more optimistic view of using heuristics. They moreover claimed that these heuristics are adaptive and functional. Decision makers are believed to possess a mental "toolbox" that contains a set of such heuristics from which the decision maker selects the most appropriate one, given the current

environment. One example (which is also the one that triggered the most research in this domain) is the recognition heuristic (see below). Others include the Take-The-Best heuristic and the fluency heuristic.

The recognition heuristic simply states that whenever a decision maker recognizes one of two options, but not the other, she should infer that the recognized option has the larger criterion value (Goldstein & Gigerenzer, 2002). This strategy works well in domains in which recognition is related to the criterion, for example, in cities' population sizes. So when asked which of two cities is the larger one, but recognizing only one of them, choosing the recognized one is a good idea. The recognition heuristic relies on only one piece of information, recognition, assuming that all further cue knowledge, if available, is ignored. This bold claim has led to a wealth of research and intense debates as, for example, documented in three special issues of the journal *Judgment and Decision Making* (Marewski, Pohl, & Vitouch, 2010, 2011a, 2011b). In sum, it seems fair to conclude that the recognition heuristic is indeed a powerful strategy that is often employed, but much less often than the authors originally proposed. The reason for this is that further information is often included in the decision, and not ignored, as the heuristic assumes.

To explain cognitive illusions, one needs to assume that such heuristics are applied more or less automatically to all cases that fulfill the respective requirements. Then one could predict the number and type of errors. But, as I said above, the authors of these heuristics focussed on the impressive accuracy of these fast and frugal strategies, not on their shortcomings.

Dual-process models

The dual- or two-process models in general posit "an associationist, parallel processing system ('System 1') that renders quick, holistic judgments [. . .] and a more deliberate, serial, and rule-based system ('System 2')." (Gilovich & Griffin, 2002, p. 16) Both systems are supposed to operate in parallel, with System 1 always running and System 2 occasionally supplementing or overriding System 1 (Stanovich, 1999). More precisely, System 1 is thought to provide "natural assessments" that are based on general-purpose heuristics (like affect, availability, causality, fluency, similarity, and surprise), while System 2 may supply "strategies or rules that are deliberately chosen to ease computational burden." (Gilovich & Griffin, 2002, p. 16f) Evans (1989) had originally labelled these two systems as "heuristic" and "analytic" and later changed them into "intuitive" and "reflective." In this framework, the classical heuristics could be considered to be more connected to System 1 and the fast and frugal heuristics to System 2, so that a peaceful coexistence of both approaches or even a theoretical integration could be possible (as Gilovich & Griffin, 2002, proposed). However, the situation is more complex. First, many and quite diverse dual-process models have been proposed (and repeatedly changed) in the past 20 years, thus perhaps adding more confusion than clarity. Second, the assumed relation between the two systems and the use of heuristics is not so clear-cut (see, e.g., Evans, 2012; Kahneman & Frederick, 2002).

Evans (2011) gave an overview of dual-process models and listed, among others, two common fallacies of misunderstanding these approaches, namely that all dual-process models are the same, and that cognitive biases are solely based on System 1, whereas System 2 always produces correct answers. Moreover, Stanovich (2008) extended such approaches into a tripartite theory by splitting System 2 into an algorithmic and a reflective mind. The latter ("System 3") is supposed to become active whenever conflicts between Systems 1 (autonomous) and 2 (algorithmic) arise that need to be resolved. This makes the assignment of heuristics to single systems even more difficult.

Keren and Schul (2009) as well as Kruglanski and Gigerenzer (2011) criticized such approaches and maintained that the distinction is not warranted and that a unified approach would suffice. Moreover, they argued that both systems are rule-based and that the same rules could govern intuitive as well as deliberate judgments. Similarly, Evans (2012) remarked that both the classical as well as the fast and frugal heuristics are related to both systems so that heuristic processing may occur in both and, as a consequence, cognitive biases may stem from both. Despite these problems, dual-process models are rather popular and widespread (cf. Chapters 8 and 25).

Individual differences

In a way, individual differences have been the very first focus of studying cognitive illusions (e.g., in research on human suggestibility, intelligence, or repression), but then the scientific community seemed to have lost its interest. The focus shifted to the general architecture of the human information-processing system. Individual differences were thus often seen as annoying, but unavoidable, sampling errors. But then, presumably due to the renewed interest in applied perspectives of cognitive functioning, individual differences gained more attention again. For example, Stanovich (1999) asked "Who is rational?", Dempster and Corkill (1999) looked at "Individual differences in susceptibility to memory illusions," and Read and Winograd (1998) edited a special issue of *Applied Cognitive Psychology* on "Individual differences and memory distortions" (see also Stanovich, Sa, & West, 2004). As examples for individual differences in specific illusions, Stanovich, West, and Toplak (2013) investigated the "myside bias" (Chapter 5) and Musch and Wagner (2007) summarized the research on the hindsight bias (see Chapter 23). Both looked (among other features) at individual intelligence, but whereas hindsight bias was related to intelligence, myside bias was not. Other research focussed on cognitive style or preferred mode of thinking.

Applied perspectives

With respect to their ecological validity, fallacies and biases have meanwhile been demonstrated in a multitude of applied contexts (as is documented in many papers in, e.g., *Applied Cognitive Psychology*). These include, for example, medical decision making, eyewitness testimony, or memories for childhood sexual abuse (cf. Roediger & McDermott,

2000). While in the beginning the focus was on possible errors and their costs, other researchers began to ask how illusions could possibly be reduced or even avoided (Caverni, Fabre, & Gonzalez, 1990b, p. 10): "Focusing on biases should contribute to the development of cognitive aids or correction procedures that bring together the necessary conditions for reducing if not canceling their effects." For example, Dodson, Koutstaal, and Schacter (2000) aimed to reduce false memories or misattributions (cf. Chapter 21) by helping their participants to focus on context-specific encoding and source-specific retrieval. Many other studies used simple warnings or informed their participants about the typically found distortion. However, as research progressed, it became clear that only some of the cognitive illusions could be remedied (see, e.g., Gigerenzer et al., 2008), while others proved their robustness (see, e.g., Pohl & Hell, 1996).

Two recent applied examples come from clinical psychology. For some time, the obsessive-compulsive disorder (OCD) has been linked to cognitive illusions (see, e.g., Moritz & Pohl, 2009). Dèttore and O'Connor (2013) even suggested that a large number of such illusions are not mere symptoms of the disorder but that they may act in "a self-sustaining vicious circle that maintains and aggravates OCD." (p. 118) With respect to pathological health anxiety (formerly known as hypochondriasis), Witthöft et al. (2015) voiced similar concerns. Following the "combined-cognitive-biases" hypothesis, they argued that "cognitive biases are not irrelevant by-products of mental disorders but, rather, serve as causal factors in the pathogenesis of the core psychopathology." (p. 1) More specifically, "cognitive biases at different stages simultaneously and interactively contribute to negative affective states and the maintenance of intrusive negative thoughts, images, and behaviors that are characteristic of a given mental disorder." (p. 2) As such, the question of how to reduce or avoid such illusions gains even more importance.

The status of cognitive illusions

In this section, I first take up the normative issue (continuing the discussion from above), and then turn to the question whether cognitive illusions represent built-in errors of the human system or could even possess adaptive value.

Normative issues

Researchers have continuously disputed which models might be used as norms and which not:

> Note that models, as elements of the theoretical framework of a study, are subject to discussion. They are not necessarily intangible. What is called a bias today may very well lose that status tomorrow if, say, the current framework appears too simplistic, naïve, or based on some superficial apprehension of the situation and the involved processing. In such a situation, the notion of bias loses its relevance.
>
> *(Caverni et al., 1990b, pp. 8–9)*

Although such discussions and according changes of normative standards are most well-known from the domain of thinking and judgment, some debate also concerned the appropriate standards for memory-based illusions. I start with these and then return to thinking and judgment.

In the memory domain, the normative issue might appear easy: One simply compares recall against the original material. Such a norm, however, assumes that literal recall would be the optimum to achieve, but Bartlett (1932) already postulated that "in a world of constantly changing environment, literal recall is extraordinarily unimportant." (p. 204) And we know from uncountable studies since then that memory is shaped to extract meaning, not to store detailled surface features. So maybe to store and recall the gist (of the original material) would be the better norm to measure one's memory performance (which, in turn, would already dissolve a number of illusions as such). However, such a change in standards has not yet received much attention in research on memory illusions. Most research seem to implicitly accept literal recall as the norm (cf. Boyer, 2009).

In contrast, in judgment and thinking, the discussion of normative issues has led to a change of standard. Instead of the bivariate (true-false) logic, most researchers now embrace belief revision in terms of Bayesian principles, the so-called "new paradigm" (Elqayam & Evans, 2013; Evans, 2012). However, they do not necessarily agree on the details. Elqayam and Evans (2013) described a large span of approaches reaching from "soft" to "strict" Bayesianism: "The contrast runs through three basic tenets of Bayesian theory: the link between probability and subjective degrees of belief; the mechanisms of belief revision; and the role of personal goals (utility)." (p. 456) The, in my view, most important feature entails a change from (alleged) objectivity to subjectivity. It is no longer the table of truth that determines the standard, but rather how participants change their subjective beliefs in the light of the given information. With this view, thinking becomes more and more a branch of decision making. But the discussion is still ongoing, leading Hahn and Harris (2014, p. 84) to assert "that the theoretical understanding of norms of rationality is incomplete and still developing. This makes it important to articulate in a given context why something is considered a norm and what relationship its justification bears to the problem at hand."

Functional views

Closely linked to normative issues is the question of what functions (if any) these illusions may serve. Three positions can be distinguished: Cognitive illusions can be understood as (a) dysfunctional errors of the system, (b) faulty by-products of otherwise functional processes, or (c) as adaptive (and thus functional) responses.

Whereas some researchers adopted a rather pessimistic view, seeing cognitive illusions as indicators of built-in errors of the human information-processing system (e.g., Piatelli-Palmarini, 1994; Thaler, 1991), most others endorsed the second view, seeing them as mere by-products (or as the "backside of the coin"). For example, "many illusions, in perception and cognition, represent the backside of

otherwise adaptive and prudent algorithms" as Fiedler (Chapter 6, p. 128) put it. Similarly, Tversky and Kahneman (1974) had acknowledged that heuristics typically lead to useful results and Gigerenzer et al. (1999) demonstrated that the fast and frugal heuristics they proposed fared about equally well compared to more complete (but time-consuming) strategies. But both may, of course, fail at times. Similarly, memory-based illusions (like hindsight bias; Chapter 23) can be understood as a by-product of the highly functional process of knowledge updating (see Hoffrage, Hertwig, & Gigerenzer, 2000; Nestler, Egloff, Küfner, & Back, 2012). Schacter, Guerin, and St. Jacques (2011) considered several such illusions, namely, the associative memory illusion (Chapter 21), the effect of imagination inflation (Chapter 22), and the misinformation effect (also Chapter 22), as reflecting "adaptive processes that contribute to efficient functioning of memory." (p. 467) The authors named the following adaptive functions that may occasionally lead to illusions: simulation of future events, more flexibility in memory search in hierarchical (gist-based) structures, and knowledge updating. In sum, many of the illusions covered in this book could represent such by-products, which moreover entail perhaps only little "costs" to the decision maker so that they are evolutionary considered bearable (cf. Hahn & Harris, 2014).

Some researchers, however, went one step further and asserted that the illusions themselves, not just their underlying mechanisms, could possess adaptive value (cf. Chapters 16 and 25). Generally, the question whether a decision, judgment, or memory is "correct" (in a normative way) usually is secondary to the question whether that decision, judgment, or memory is helpful in the current situation. Boyer (2009, p. 513) accordingly asserted that "it makes evolutionary sense to keep in mind that organisms do not develop cognitive abilities (e.g., retrieval of past experience) for abstract epistemic benefits (knowing what used to be the case). They retrieve information inasmuch as it helps fitness-enhancing decision-making in the present." Also Schacter et al. (2011) argued that under some conditions false memories can have beneficial effects. As an example, they discussed "that people frequently remember their pasts in an overly positive or negative manner in order to inflate their current self-evaluation" (p. 471) (see Wilson & Ross, 2003, 2004). Similarly, McKay and Dennett (2009) discussed the functions of some "misbeliefs" and claimed that many of them should be considered faulty but tolerable by-products, but that some may represent genuine adaptive functions. They especially focussed on "positive illusions" (see Chapter 17), which in their view are highly adaptive. Sutton (2009) further noted that the positivity bias in autobiographical memory is related to enhanced emotion regulation, that forgetting is also an adaptive response, and that "misremembering things in particular positive ways might have direct personal, motivational, and social benefits." (p. 536) Of course, such illusions like the positivity bias could not have evolved unlimited, because otherwise we would all have lost contact to reality (which would have been detrimental to survival). Rather, the amount of bias that is beneficial is also a matter of evolutionary selection, which may have yielded an "optimal margin" of such illusions (Baumeister, 1989).

The intentions of this book

Having done research and having taught courses on "cognitive illusions" for many years, I found that it was about time to have a representative collection of cognitive illusions in one book instead of dozens of scattered, typically much-too-specific papers. So I took my list of illusions and started to look for collaborators (for the first edition). With this enterprise, I had the following intentions in mind.

Hopefully, this dense coverage of cognitive illusions will help to further integrate the field and to foster new *research* and the development of more precise *models* to explain these illusions. This is especially needed, because so far several approaches are still not precise enough to allow rigorous testing. On the other hand, some highly detailed models (even including simulation models) have been developed, thus documenting the progress of the field (see, e.g., Pohl, Eisenhauer, & Hardt, 2003). A related goal is based on the fact that until recently most of the illusions have been studied in complete isolation from one another without any cross-referencing. As a consequence, a multitude of empirical methods as well as theoretical approaches developed in parallel, thus obscuring the view on any common mechanisms that possibly lie behind larger classes of illusions. Therefore, this book also intends to help in discovering such accordances in order to untangle the basic cognitive processes responsible for the observed illusions. One such attempt that draws an analogy to the research on suggestion and suggestibility is presented in more detail at the end of this book (Chapter 25).

Besides its envisioned impact on empirical research and theoretical developments, the book is also intended to serve as a *handbook* both to professionals as well as to informed lay people. It brings together a representative sample of cognitive illusions from various domains, thus allowing a quick overview about this exciting field of cognitive psychology. Each chapter presents the respective state of the art in a comprehensive overview, covering research from the earliest experiments to the most recent ones. The inclusion of applied perspectives should, moreover, make the chapters informative to experts from applied domains (economics, medicine, law, counselling, forecasting, etc.). The discussion of which conditions may increase or reduce (or even eliminate) the respective illusion should also be helpful for practitioners to be aware of these distorting influences and of how to possibly reduce their impact by taking appropriate measures. And finally, the references in each chapter include the most important sources, especially classical demonstrations, high-impact papers, and meta-analytic studies (if available).

As a teacher of cognitive psychology, I find it always beneficial for the students' learning progress when psychological topics are not only talked about, but also personally experienced (which is also more fun). Cognitive illusions offer themselves as such a fruitful enterprise, where the single phenomena are generally easy to replicate. Most of the cognitive illusions are robust and will even work under less-than-optimal conditions, thus rendering them applicable as classroom demonstrations.

So, the last but not least intention of editing this book was to use it as a *textbook* for classes of cognitive psychology. To this end, the chapters are written in such a way as to be informative to the expert but also comprehensible to the beginner. In addition, each chapter (except Chapter 25) follows the same general outline, including three main sections, namely, (1) a survey of empirical findings, (2) a discussion of theoretical explanations, and (3) a classroom demonstration. Each chapter also contains a concise bullet-point summary and suggestions for further reading.

The classroom demonstration is in my view an important part of each chapter, because it provides a detailed description of a prototypical experiment that elicits the illusion and that can easily be conducted in the classroom. This can be a classical example already published elsewhere or a new (maybe more simple) variant. The text includes all procedural and statistical details that are necessary to conduct and analyze the described experiment. The materials, instructions, and possibly more detailed information for running the experiment are mostly given in the respective appendix. Thus it should be fairly easy to incorporate these experiments into a class on "cognitive illusions" (as I and others regularly do).

Changes to the first edition

When the publisher approached me asking for a second edition (about 10 years after the first) and I agreed, I first checked the literature on which of the phenomena described in the first edition still yielded currently active research, and which new fields had emerged in the past decade. As a consequence, I decided to drop a few of the old chapters, namely, those on the Moses illusion (Park & Reder, 2004), the orientation illusion (Jones & Martin, 2004), and on illusions of change or stability (Wilson & Ross, 2004), but to keep the remaining ones (asking the respective authors—if available and willing—to update their chapter; otherwise, I asked new authors). In addition, I included several new chapters (and authors), namely, on probability matching (Chapter 3), the framing effect (Chapter 4), the halo effect (Chapter 15), the revelation effect (Chapter 18), the survival processing effect (Chapter 19), and retrieval-induced forgetting (Chapter 24). I hope that this updated collection of intriguing phenomena in thinking, judgment, and memory gives a comprehensive overview of according research activities and theoretical advancements in the field.

Acknowledgements

The first edition of this work was in part supported by a grant from the Deutsche Forschungsgemeinschaft (Po 315/6–4). For their helpful comments on an earlier version of this chapter, I thank Martin Heydemann, Michael Bender, Gregor Lachmann, and Ulrich Hoffrage. With respect to the whole book, my sincere thanks go to the 34 international authors who participated in this project. It was a truly inspiring enterprise to work with them. Finally, for an always friendly and cooperative

guidance in publishing this second edition, I would like to express my special thanks to Ceri Griffiths and the team from Routledge.

References

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, *93*, 203–231.

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

Baumeister, R. F. (1989). The optimal margin of illusion. *Journal of Social and Clinical Psychology, 8*(2), 176–189.

Boyer, P. (2009). Extending the range of adaptive misbelief: Memory 'distortions' as functional features. *Behavioral and Brain Sciences, 32*, 513–514.

Brewer, W. F., & Nakamura, G. V. (1984). The nature and functions of schemas. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 1, pp. 119–160). Hillsdale, NJ: Erlbaum.

Caverni, J.-P., Fabre, J. M., & Gonzalez, M. (Eds.). (1990a). *Cognitive biases*. Amsterdam: North-Holland.

Caverni, J.-P., Fabre, J. M., & Gonzalez, M. (1990b). Cognitive biases: Their contribution for understanding human cognitive processes. In J.-P. Caverni, J. M. Fabre, & M. Gonzalez (Eds.), *Cognitive biases* (pp. 7–12). Amsterdam: North-Holland.

Dempster, F. N., & Corkill, A. J. (1999). Individual differences in susceptibility to interference and general cognitive ability. *Acta Psychologica, 101*(2–3), 395–416.

Dèttore, D., & O'Connor, K. (2013). OCD and cognitive illusions. *Cognitive Therapy and Research, 37*(1), 109–121.

Dodson, C. S., Koutstaal, W., & Schacter, D. L. (2000). Escape from illusion: Reducing false memories. *Trends in Cognitive Science, 4*(10), 391–397.

Edwards, W., & von Winterfeldt, D. (1986). On cognitive illusions and their implications. In H. R. Arkes & K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (pp. 642–679). Cambridge: Cambridge University Press.

Elqayam, S., & Evans, J. St. B. T. (2013). Rationality in the new paradigm: Strict versus soft Bayesian approaches. *Thinking & Reasoning, 19*(3), 453–470.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ: Erlbaum.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove, UK: Psychology Press.

Evans, J. St. B. T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review, 31*(2–3), 86–102.

Evans, J. St. B. T. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning, 18*(1), 5–31.

Fiedler, K., & Krueger, J. I. (2011). More than an artifact: Regression as a theoretical construct. In J. I. Krueger (Ed.), *Social judgment and decision making* (pp. 171–189). New York: Psychology Press.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 2, pp. 83–115). Chichester: Wiley.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A rebuttal to Kahneman and Tversky (1996). *Psychological Review*, *103*, 592–596.

Gigerenzer, G. (2004). The irrationality paradox. *Behavioral and Brain Sciences, 27*(3), 336–338.

Gigerenzer, G., Hertwig, R., Hoffrage, U., & Sedlmeier, P. (2008). Cognitive illusions reconsidered. In C. R. Plott & V. L. Smith (Eds.), *Handbook of experimental economics results* (pp. 1018–1034). Amsterdam: Elsevier.

Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The foundations of adaptive behavior.* Oxford: Oxford University Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684–704.

Gigerenzer, G., Todd, P. M., & The ABC Research Group. (Eds.). (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life.* New York: The Free Press.

Gilovich, T., & Griffin, D. (2002). Introduction – Heuristics and biases: Then and now. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 1–18). New York: Cambridge University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment.* New York: Cambridge University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*(1), 75–90.

Hahn, U., & Harris, A. J. L. (2014). What does it mean to be biased: Motivated reasoning and rationality. In B. Ross (Ed.), *Psychology of learning and motivation* (Vol. 61, pp. 41–102). Amsterdam, NL: Elsevier.

Hell, W., Fiedler, K., & Gigerenzer, G. (Eds.). (1993). *Kognitive Täuschungen* [Cognitive illusions]. Heidelberg: Spektrum der Wissenschaft.

Hertwig, R. (2012). The psychology and rationality of decisions from experience. *Synthese, 187*, 269–292.

Hertwig, R., & Ortmann, A. (2005). The cognitive illusion controversy: A methodological debate in disguise that matters to economists. In R. Zwick & A. Rapoport (Eds.), *Experimental business research* (Vol. 3, pp. 113–130). Dordrecht, NL: Springer.

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin, 138*(2), 211–237.

Hodgson, R., & Davey, S. J. (1886/1887). The possibilities of mal-observation and lapse of memory from a practical point of view. *Proceedings of the Society for Psychical Research*, *4*, 381–495.

Hoffrage, U., Hertwig, R., & Gigerenzer, G. (2000). Hindsight bias: A by-product of knowledge-updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 566–581.

Jones, G. V., & Martin, M. (2004). Orientation bias in memory. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (pp. 293–308). Hove, UK: Psychology Press.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York: Cambridge University Press.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, *103*, 582–591.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science, 4*(6), 533–550.

Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences, 27*, 313–327.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118*(1), 97–109.

MacCoun, R. (2002, December). Why a psychologist won the Nobel prize for economics. *APS Observer*, *15*, 1 and 8.

Marewski, J. N., Pohl, R. F., & Vitouch, O. (2010). Recognition processes in inferential decision making [Special issue; Part 1]. *Judgment and Decision Making, 5*(4).

Marewski, J. N., Pohl, R. F., & Vitouch, O. (2011a). Recognition processes in inferential decision making [Special issue; Part 2]. *Judgment and Decision Making, 6*(1).

Marewski, J. N., Pohl, R. F., & Vitouch, O. (2011b). Recognition processes in inferential decision making [Special issue; Part 3]. *Judgment and Decision Making, 6*(5).

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review, 118*(3), 393–437.

McKay, R. T., & Dennett, D. C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences, 32*, 493–561.

Moritz, S., & Pohl, R. F. (2009). Biased processing of threat-related information rather than knowledge deficits contributes to overestimation of threat in obsessive-compulsive disorder. *Behavior Modification, 33*(6), 763–777.

Musch, J., & Wagner, T. (2007). Did everybody know it all along? A review of individual differences in hindsight bias. *Social Cognition, 25*(1), 64–82.

Nestler, S., Egloff, B., Küfner, A. C. P., & Back, M. D. (2012). An integrative lens model approach to bias and accuracy in human inferences: Hindsight effects and knowledge updating in personality judgments. *Journal of Personality and Social Psychology, 103*(4), 689–717.

Park, H., & Reder, L. M. (2004). Moses illusion. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (pp. 275–291). Hove, UK: Psychology Press.

Piatelli-Palmarini, M. (1994). *Inevitable illusions: How mistakes of reason rule our minds*. New York: Wiley.

Pohl, R. F., Bender, M., & Lachmann, G. (2002). Hindsight bias around the world. *Experimental Psychology*, *49*, 270–282.

Pohl, R. F., Eisenhauer, M., & Hardt, O. (2003). SARA: A cognitive process model to explain the anchoring effect and hindsight bias. *Memory*, *11*, 337–356.

Pohl, R. F., & Hell, W. (1996). No reduction of hindsight bias with complete information and repeated testing. *Organizational Behavior and Human Decision Processes*, *67*, 49–58.

Read, J. D., & Winograd, E. (Eds.). (1998). Individual differences and memory distortions [Special issue]. *Applied Cognitive Psychology*, *12*.

Roediger, H. L., III (1996). Memory illusions. *Journal of Memory and Language*, *35*, 76–100.

Roediger, H. L., III, & McDermott, K. B. (2000). Distortions of memory. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 149–162). Oxford: Oxford University Press.

Schacter, D. L., Coyle, J. T., Fischbach, G. D., Mesulam, M. M., & Sullivan, L. E. (Eds.). (1995). *Memory distortions: How minds, brains, and societies reconstruct the past*. Cambridge, MA: Harvard University Press.

Schacter, D. L., Guerin, S. A., & St. Jacques, P. L. (2011). Memory distortion: An adaptive perspective. *Trends in Cognitive Sciences, 15*(10), 467–474.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

Stanovich, K. E. (2008). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 55–88). Oxford: Oxford University Press.

Stanovich, K. E., Sa, W. C., & West, R. F. (2004). Individual differences in thinking, reasoning, and decision making. In J. P. Leighton & R. J. Sternberg (Eds.), *The nature of reasoning* (pp. 375–409). New York: Cambridge University Press.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science, 22*(4), 259–264.

Sully, J. (1881). *Illusion: A psychological study*. London: Kegan Paul.

Sutton, J. (2009). Adaptive misbeliefs and false memories. *Behavioral and Brain Sciences, 32*, 535–536.

Thaler, R. (1991). *Quasi-rational economics*. New York: Sage.

Todd, P. M., Gigerenzer, G., & The ABC Research Group. (2012). *Ecological rationality: Intelligence in the world*. New York: Oxford University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Wilson, A. E., & Ross, M. (2003). The identity function of autobiographical memory: Time is on our side. *Memory, 11*(2), 137–149.

Wilson, A. E., & Ross, M. (2004). Illusions of change or stability. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (pp. 379–396). Hove: Psychology Press.

Witthöft, M., Kerstner, T., Ofer, J., Mier, D., Rist, F., Diener, C., & Bailer, J. (2015). Cognitive biases in pathological health anxiety: The contribution of attention, memory, and evaluation processes. *Clinical Psychological Science*, online. DOI: 10.1177/2167702615593474.

This page intentionally left blank

PART I

Thinking

This page intentionally left blank

1

CONJUNCTION FALLACY

John E. Fisk

Violations of the rules of probability theory and associated systematic reasoning biases have been widely demonstrated. The likelihood of certain types of events are systematically overestimated while for other types of events the opposite pattern emerges. One area that has received much attention in the research literature concerns conjunction-rule violations. Formally, the conjunction rule may be expressed as follows:

$$P(A\&B) = P(A) \times P(B|A) \tag{1}$$

In simple terms, the probability of event A and event B both occurring together is equal to the probability of event A multiplied by the (conditional) probability of event B *given that A has occurred*. For example, the probability that I will *study* (event A) AND *pass* my exams (event B) is equal to the probability that I will study multiplied by the probability that I will pass GIVEN that I have studied:

$$P(\text{Study and Pass}) = P(\text{Study}) \times P(\text{Pass} \mid \text{Study}) \tag{2}$$

When the two events A and B are independent then Equation 1 simplifies to

$$P(A\&B) = P(A) \times P(B) \tag{3}$$

since for independent events:

$$P(B) = P(B|A) = P(B|\text{not } A) \tag{4}$$

The extent to which individuals make judgements consistent with the conjunction rule has been one of the most investigated areas of probabilistic reasoning with research dating back over 50 years (e.g., Cohen, Dearnaley, & Hansel, 1958). In the 1980s, starting

with Tversky and Kahneman's (1983) seminal study, the focus of research shifted to a particular type of violation of the conjunction rule known as the conjunction fallacy (Agnoli & Krantz, 1989; Fiedler, 1988; Tversky & Kahneman, 1983; Wells, 1985; Yates & Carlson, 1986). The fallacy occurs when the conjunctive probability is assigned a value exceeding that assigned to one or both of the component events, that is,

$$P(A\&B) > P(A) \text{ and/or} \tag{5}$$
$$P(A\&B) > P(B), \tag{6}$$

Such judgements, which violate the conjunction rule,[1] are commonplace, with between 50% and 90% of individuals responding in this fashion (Fisk & Pidgeon, 1996; Gavanski & Roskos-Ewoldsen, 1991; Tversky & Kahneman, 1983; Yates & Carlson, 1986).

Perhaps the best known example of the conjunction fallacy is the Linda scenario from Tversky and Kahneman's (1983) classic study. In this scenario, a fictitious person, Linda, is described as follows: 'Linda is 31 years old, single, outspoken and very bright. At university she studied philosophy. As a student she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations.' Having read the description, individuals were asked to rank the following three statements in order of their probability:

i Linda is active in the feminist movement.
ii Linda is a bank teller.
iii Linda is a bank teller and is active in the feminist movement.

Clearly, the third statement is a conjunction of the first two and so according to the conjunction rule cannot be more likely than first and second statements. However, in Tversky and Kahneman's (1983) study, 85% of participants committed the fallacy, ranking the third statement as more probable than the second. Since then and continuing up to the present day, numerous studies have produced evidence of the conjunction fallacy (e.g., Agnoli & Krantz, 1989; Costello & Watts, 2014; Fisk & Pidgeon, 1996; Gavanski & Roskos-Ewoldsen, 1991; Hertwig & Chase, 1998; Nilsson, Winman, Juslin, & Hansson, 2009; Rogers, Fisk, & Wiltshire, 2011; Tentori, Crupi, & Russo, 2013; Yates & Carlson, 1986). A similar design is the basis for the detailed experiment given in Text box 1.1, which can be used as a classroom demonstration.

TEXT BOX 1.1 CLASSROOM DEMONSTRATION

Background

The classroom demo accompanying this chapter utilises a concept which has yet to be applied to conjunctive probability judgement: support theory (Tversky & Koehler, 1994). According to the theory, when an implicit disjunction is

decomposed and its component parts are made explicit, support for it and its implied probability increases. For example, the probability of inclement weather on a randomly selected day should be equal to the sum of the probabilities of all the different weather events (e.g., hail, sleet, blizzard) that we would define as inclement on such a randomly selected day. However, it has been demonstrated that such sums are generally significantly larger than the probability assigned to the superordinate category (inclement weather). Similarly, in the case of the Linda scenario, we might argue that the extension of the less likely event, bank teller, is itself an implicit disjunction that can be decomposed to make explicit all of the different mutually exclusive event combinations in which Linda may be a bank teller, e.g., a bank teller and votes Conservative, a bank teller and votes Labour, a bank teller and votes Green, etc. Decomposing the bank teller event in this way might be expected to increase support for it. As a consequence, the probability assigned to it might increase relative to the conjunctive events, which might in turn reduce the incidence of the fallacy. Thus, this experiment presents two versions of the Linda scenario, the standard one in which the bank teller event is left implicit and the other version in which the bank teller event is decomposed into a number of explicit mutually exclusive conjunctions as shown below. The hypothesis to be tested is that the bank teller probability will be significantly larger in the second version and the incidence of the fallacy significantly less.

Method

Design

The independent variable is the problem format (standard version or support theory version). The dependent variables are the incidence of the fallacy and the magnitude of the bank teller probability.

Participants

At least 30 participants should be randomly assigned to each condition.

Materials

- *Standard version*

  ‘Linda is 31 years old, single, outspoken and very bright. At university she studied philosophy. As a student she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations.’ How likely are each of the following statements (for each, enter a percentage probability between 0 and 100):

  i Linda is a bank teller and voted for the Labour Party in the 2015 UK General Election.

ii Linda is an interior designer and voted for the Green Party in the 2015 UK General Election.
iii Linda is a home maker and voted for the Liberal Democrat Party in the 2015 UK General Election.
iv Linda is a financial analyst and voted for the Conservative Party in the 2015 UK General Election.
v Linda is a bank teller.

- *Support theory version*

  'Linda is 31 years old, single, outspoken and very bright. At university she studied philosophy. As a student she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations.' How likely are each of the following statements (for each, enter a percentage probability between 0 and 100):

  i Linda is a bank teller and voted for the Labour Party in the 2015 UK General Election.
  ii Linda is a bank teller and voted for the Green Party in the 2015 UK General Election.
  iii Linda is a bank teller and voted for the Liberal Democrat Party in the 2015 UK General Election.
  iv Linda is a bank teller and voted for the Conservative Party in the 2015 UK General Election.
  v Linda is a bank teller.

Results

The probability assigned to the bank teller event in each version can be analysed using a between-participant *t*-test. The fallacy is committed if the probability assigned to the first statement is greater than the probability assigned to the last. The incidence of the fallacy in both versions can be compared using the chi-squared test.

Explanations

Over the years a diverse range of explanations have been provided for the conjunction fallacy phenomenon. Tversky and Kahneman (1983) attempted to explain the fallacy in terms of the representativeness heuristic (see Chapter 11). Considering the two component statements above, the description of Linda is *representative* of or *similar* to that of a feminist while, by way of contrast, Linda's description is unrepresentative of the characteristics of being a bank teller. However, since Linda's description is representative of one of the statements within the conjunction, overall,

it is judged to be more representative of the conjunctive event than it is of the bank teller event on its own. Since Tversky and Kahneman (1983) proposed that probability judgements are based on differences in representativeness, the conjunctive event is therefore judged to be more likely than the bank teller event. This same scenario along with many others with varying contents has been used in a large number of studies (see above) and the conjunction fallacy has been replicated on numerous occasions. Apart from the potential role of representativeness, many other accounts of the conjunction fallacy have appeared over the past 40 years, including:

1 Linguistic interpretation
2 Signed summation
3 Frequentist interpretations and response format
4 Applying a different probabilistic rule
5 Surprise theory
6 Cognitive experiential self-theory (CEST)
7 Averaging
8 Quantum probability
9 Inductive confirmation
10 Random noise or sampling errors.

Each of these theoretical perspectives will be discussed in turn.

1. Linguistic interpretation

It has been suggested that participants interpret the single event as implicitly implying the negation of the other member of the conjunction. Thus, for example, when asked to evaluate the probability that 'Linda is a bank teller', participants actually estimate the probability that 'Linda is a bank teller and *not* a feminist'. Ranking the latter as less likely than the statement 'Linda is a bank teller and a feminist' would not be a violation of the conjunction rule. However, while different linguistic interpretations might account for a small proportion of conjunction rule violations, the consensus is that they do not offer a comprehensive explanation of the phenomenon (Bonini, Tentori, & Osherson, 2004; Morrier & Borgida, 1984; Yates & Carlson, 1986; but cf. Chiesi, Gronchi, & Primi, 2008).

2. Signed summation

'Signed summation' is a heuristic procedure proposed by Yates and Carlson (1986). It is assumed that individuals represent degrees of likelihood in terms of a 'qualitative likelihood index' (QLI). Unlikely events are represented by negative values on this index, the more unlikely the more negative the likelihood index becomes. Conversely, likely events are represented by positive values on the QLI, with more likely events associated with larger values on the index. Unlike standard representations of probability, the QLI is unbounded and can assume negative as well as positive values.

According to the theory of signed summation, the likelihood of the conjunction of two events is the signed sum of the QLI values corresponding to the two individual events. Thus, the conjunction of a likely event with an unlikely event is the sum of a negative and a positive number. Arithmetically, this signed sum will be greater than the QLI for the unlikely event, thus giving rise to the conjunction fallacy. Signed summation also predicts that the conjunction fallacy should not occur for conjunctions of two unlikely events and that for conjunctions of two likely events a double fallacy should occur with both component events being rated more probable than their conjunction. Yates and Carlson's (1986) results and those of Wells (1985) are broadly consistent with these expectations.

3. Frequentist interpretations and response format

It has been proposed that when responses to conjunction problems are elicited in terms of frequencies, the conjunction fallacy does not occur (Fiedler, 1988). In the frequency format, Linda's description is as it appears above: 'Linda is 31 years old . . .'. Participants are then asked to imagine 100 individuals, all of whom closely resemble Linda's description. They are then asked to respond to the following:

i How many of the 100 would be bank tellers?
ii How many of the 100 would be active in the feminist movement?
iii How many of the 100 would be active in the feminist movement and bank tellers?

When the problem is posed in this kind of way, the incidence of the conjunction fallacy is greatly reduced in that most people indicate that there are a greater number of bank tellers than there are feminist bank tellers. However, Evans and co-workers have shown that frequency versions of probabilistic reasoning problems do not inevitably result in better judgements; indeed, sometimes the opposite outcome occurs. Furthermore, Evans and others have shown that when the extensional nature of the conjunctive relationship is made transparent, problems posed in probabilistic terms also can result in a reduced incidence of the fallacy (Evans, Handley, Perham, Over, & Thompson, 2000).

Aside from the frequency-probability response format distinction, more generally, Hertwig and Chase (1998) have argued that the nature of the response required from individuals will determine what sort of strategy is utilised by the participant. They maintained that problems requiring participants to provide numerical estimates are more likely to elicit rule-based strategies while problems requiring participants to simply rank alternatives produce cue based strategies. Consistent with these propositions, in a series of studies, Hertwig and Chase found that rule use was more prevalent among those individuals who were asked to provide numerical probability estimates. The particular rules used included a 'ceiling rule' (setting the conjunctive probability exactly equal to the smaller component event probability) and a quasi-normative multiplicative rule. Among a group of statistically naïve participants, the

majority of those avoiding the conjunction fallacy responded in accordance with the ceiling rule while the responses among those with statistical training were rather more in tune with the multiplicative than the ceiling rule.

Those participants who were asked by Hertwig and Chase to supply ranks appeared to rely on a cue based strategy. For example, in the Linda scenario, each of the elements in Linda's description may potentially act as a cue. Thus, participating in anti-nuclear demonstrations provides more evidential support for the proposition 'feminist bank teller' than it does for the proposition 'bank teller', resulting in the former being ranked more highly in probabilistic terms than the latter. The notion of evidential support appears similar to the concept of inductive confirmation proposed by Tentori et al. (2013) and discussed below. However, in the case of Hertwig and Chase, the cue values are determined by the reverse probabilities, that is, P(evidence | statement), for example, P(participating in anti-nuclear demonstrations | bank teller) versus P(participating in anti-nuclear demonstrations | feminist bank teller). In this sense, Hertwig and Chase's account appears to echo Wolford, Taylor, and Beck's (1990) assertion that conjunction fallacies arise because individuals estimate the wrong probabilities (see below).

4. Applying a different probabilistic rule

Some researchers have claimed that the manner in which many of the conjunction problems are framed makes it unclear whether the conjunction rule is the appropriate rule to apply. Wolford et al. (1990) argued that in problems like the Linda one above, the participant may believe that one of the statements is actually true and that instead of evaluating the probability of the statement, participants are actually evaluating the probability of Linda's description given the statement, that is, instead of evaluating P(feminist and bank teller | Linda) and P(bank teller | Linda) participants are actually evaluating P(Linda | feminist and bank teller) and P(Linda | bank teller). While producing probabilities such as:

$$P(\text{feminist and bank teller} \mid \text{Linda}) > P(\text{bank teller} \mid \text{Linda}) \tag{7}$$

is clearly a violation of the conjunction rule, making judgements such that

$$P(\text{Linda} \mid \text{feminist and bank teller}) > P(\text{Linda} \mid \text{bank teller}) \tag{8}$$

does not necessarily violate any normative rule.

While this suggestion is an appealing one, one of my own studies (Fisk, 1996), in which participants were asked for *both* kinds of probability judgement, showed that participants still violated the conjunction rule in contexts where it should clearly apply. Furthermore, there are some instances where Bayes' theorem does produce a situation where, depending on the prior probabilities and the information provided:

$$P(\text{Linda} \mid \text{bank teller}) > P(\text{Linda} \mid \text{feminist and bank teller}). \tag{9}$$

However, even in these situations participants still produced estimates such that

$$P(\text{Linda} \mid \text{feminist and bank teller}) > P(\text{Linda} \mid \text{bank teller}) \quad (10)$$

essentially violating Bayes' theorem. Thus, while it may be that some participants apply different probabilistic rules, this does not offer a complete account of the conjunction fallacy, which continues to occur in contexts where the conjunction rule clearly applies.

5. Surprise theory

In a series of studies from our own laboratory, during the 1990s and early 2000s, we attempted to explain conjunction rule violations in terms of Shackle's (1969) theory of 'potential surprise'. According to Shackle, probabilities are systematically related to the potential that events have to surprise us. We may be extremely surprised by the occurrence of certain events, for example, the number 20 seed winning the Wimbledon men's singles final, but less surprised or not surprised at all by other events, e.g., Serena Williams winning the women's title. It is these relative degrees of surprise that Shackle maintained determine the probabilities that we produce when considering the likelihood of events. Crucially Shackle argued that when considering joint events (conjunctions and disjunctions) the component surprise values are not combined but instead the surprise value of the joint event is essentially determined by the surprise value of one or other of the components, the other one having no obvious influence. For conjunctive events, Shackle argued that it is the more surprising (less likely) event that plays this role; for disjunctive events, it is the less surprising (more likely) event.

Shackle's theory can account for many of the results that I and my co-workers have obtained in a number of studies; specifically, we have found that the probability assigned of the likely event more often than not appears to play no statistically significant role in determining the conjunctive probability. It is the unlikely event probability that plays the central role. In the case of disjunctions, my co-workers and I found that the likely component event probability is crucial with the unlikely event accounting for less (or indeed no) unique variance (Fisk, 2002; Fisk & Pidgeon, 1996).

I have also shown that Shackle's account of conditional surprise values and, by extension, conditional probabilities can account for the fact that the conjunction fallacy is more likely in situations where the two component events are positively conditionally related even though neither of them may be particularly consistent with the background evidence (Fisk & Pidgeon, 1998). Tversky and Kahneman (1983) maintained that it is the strength of the causal link which inflates the value assigned to the conjunction, potentially giving rise to the conjunction fallacy. However, while the presence of a positive conditional relationship makes the fallacy more likely, according to the surprise theory, the actual *strength* of the relationship exerts no direct influence on the value assigned to the conjunction nor on the

extent of the fallacy. All that matters is that a positive relationship exists. Consistent with Shackle's theory, both Fabre, Caverni, and Jungermann (1995) and Fisk and Pidgeon (1998) found that the fallacy was more likely to occur when a positive conditional relationship between the components events existed *but* that this tendency was unaffected by whether the positive relationship was strong or weak.

6. Cognitive experiential self-theory (CEST)

In the mid-1990s, Epstein and co-workers attempted to explain conjunction rule violations in terms of their cognitive-experiential self-theory (CEST) concept. With this perspective in mind, Donovan and Epstein (1997) introduced two distinct dimensions (natural-unnatural and concrete-abstract), which they claimed make it possible to identify the conditions under which the fallacy will arise. Concrete representations involve objects or events, which can be directly imagined or experienced (e.g., the Linda scenario set out above). Abstract problems cannot be directly experienced or imagined as, for example, when the task is to work out the value of the conjunction P(A&B) given that P(A) = 0.2 and P(B|A) = 0.6. A natural problem is one that elicits appropriate reasoning strategies, *appropriate* in the sense that they are capable of producing the correct solution if applied properly. Unnatural problems are, conversely, those that elicit inappropriate reasoning strategies. For example, the Linda problem is unnatural because it elicits reasoning based on representativeness, while solution of the problem requires statistical reasoning. Thus, in terms of the two dimensions, the Linda scenario is concrete and unnatural. On the other hand, problems such as the one set out below (Donovan & Epstein, 1997, p. 19) are concrete and natural because they usually elicit statistical reasoning.

> John goes to the auto races. He bets on two races. In one race he bets on a car driven by an experienced driver with a long history of winning races. In a different race, he bets on a car driven by an inexperienced driver who has won very few races. Check which of the following you believe is less likely:
>
> The car driven by the inexperienced driver wins.
> The car driven by the inexperienced driver and the car being driven by the experienced driver both win.

Consistent with this, Donovan and Epstein went on to show that the number of conjunction errors in concrete natural problems is significantly fewer than the number of errors in concrete unnatural problems.

7. Averaging

A further account of conjunctive probability judgement is based on the notion that individuals resort to some form of averaging process in order to derive conjunctive judgements. The idea that individuals utilise some sort of averaging strategy

featured in the earliest research into conjunctive judgement and remains topical today. For example, Wyer (1970) proposed that conjunctive judgements might be (erroneously) derived by simply taking the average of the two component probabilities. Wyer found that participants' conjunctive probability judgements were substantially higher than the normative multiplicative values and in fact closer in absolute magnitude to the simple average of the component event probabilities. However, the estimates were actually more closely *correlated* with the product of the component event probabilities and with the normative values themselves [i.e., with $P(A) \times P(B)$ and $P(A) \times P(B|A)$]. Following on from Wyer, Goldsmith (1978) found evidence for a number of arithmetic combination rules underpinning conjunctive judgements. Some participants apparently based their judgement on the smaller component event probability (see surprise theory and Hertwig and Chase's 'ceiling rule'), others used either the simple average, or the product of the component event probabilities. However, there was less evidence that the larger event played a significant role (again consistent with surprise theory) and less evidence that the conditional probability was used in forming the conjunctive judgement.

Later on, Birnbaum, Anderson, and Hynan (1990) proposed a geometric averaging model of conjunctive probability judgement. Birnbaum et al. maintained that, in any given context, actual probability judgements will be related to the real subjective probabilities plus some error component. Elicited conjunctive judgements will conform to the following equation:

$$q(A\&B) = s(A)^{\beta}s(B|A)^{\alpha} + e_{A\&B} \tag{11}$$

where $s(A)$ and $s(B|A)$ are the subjective component event and conditional probabilities and $e_{A\&B}$ is the random error component. The parameters α and β are estimated and when they are both equal to one, the above equation is equivalent to the normative model. In an experimental study, Birbaum et al. found that the above equation achieved the best fit to the data when values of the two parameters α and β were substantially less than one. The conjunction fallacy in turn was most likely to occur in situations where participants produced high values for the conditional probability and low values for the component event probability or vice versa.

More recently, Nilsson et al. (2009) have proposed a weighted averaging model in an attempt to account for both conjunctive and disjunctive reasoning errors. Nilsson et al. argued that conjunction and disjunction fallacies occur because individuals combine probabilities in an additive manner so that conjunctive and disjunctive probabilities are essentially a weighted average of the component probabilities. However, in Nilsson and co-workers' configural weighted averaging (CWA) model, the weights on the component events vary according to whether the judgement is conjunctive or disjunctive. Nilsson et al. maintained that the less likely event plays the dominant role in determining the probability assigned to the conjunctive event while this role is performed by the more likely component in determining the probability assigned to the disjunctive event. The CWA model predicts that joint probabilities are derived from the component event probabilities in a manner

consistent with the two equations below where the subscripts $_L$ and $_H$ refer to the component with the lower or higher probability value:

$$P(A_L \,\&\, B_H) = (1 - b) \times P(A_L) + b \times P(B_H) \quad (12)$$
$$P(A_L \text{ or } B_H) = b \times P(A_L) + (1 - b) \times P(B_H) \quad (13)$$

Nilsson et al. suggested that in general $0 \leq b \leq 0.5$, although in their experimental studies they assumed a value for $b = 0.2$.

In a series of experiments, Nilsson and co-workers' participants assessed the likelihood of statements concerning the relative populations of countries, capital cities, land areas, etc. (for example, the probability that 'France has more inhabitants than Peru'). They also estimated the likelihood of conjunctive statements containing these components, e.g., 'France has more inhabitants than Peru and Sweden has more inhabitants than Estonia' and equivalent disjunctive statements where 'and' was replaced by 'or'. Nilsson et al. found that participants' conjunctive and disjunctive estimates were better predicted by the CWA model compared to the normative model. It is worthy of note that the application of Equations 13 and 14 will always result in fallacies being committed, so in its simplest form the model could not account for the fact that some judgements were not fallacious. Nilsson et al. addressed this problem by proposing that CWA judgements were distorted by random noise that sometimes was sufficiently large to push the conjunctive (disjunctive) probability below (above) the less (more) likely component event probability, thereby avoiding the fallacy. In subsequent analyses, they produced evidence supporting this proposition.

8. Quantum probability

Among the other more recent theories of probability judgement, Busemeyer, Pothos, Franco, and Trueblood (2011) have proposed what is perhaps the most complex and conceptually demanding accounts of fallacious reasoning. In order to illustrate the framework proposed by Busemeyer et al., we have made use of the Bill scenario from Tversky and Kahneman (1983):

> Bill is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In college, he was strong in mathematics but weak in social studies and literature. Keeping this description in mind how likely is it that:
>
> Bill is an accountant.
> Bill plays jazz for a hobby.
> Bill is an accountant and plays jazz for a hobby.

Typically the conjunctive statement, 'Bill is an accountant and plays jazz for a hobby', is assigned a higher probability than the component 'Bill plays jazz for a hobby'.

Busemeyer et al. argued that the conjunction fallacy can be explained within a quantum probability framework in which events are represented as basis vectors within a multidimensional space. The number of dimensions (and basis vectors)

depends on the number of events that the individual attempts to evaluate concurrently. For example, whether or not Bill is an accountant, works in a tax office and is computer literate, would define an eight dimensional space with each dimension defined in terms of a basis vector representing the presence or absence of each particular event. For example, one dimension or basis vector could represent Bill being an accountant NOT working in a tax office and NOT being computer literate while another could represent Bill NOT being an accountant but working in a tax office and being computer literate, and so on. The background information or evidence concerning Bill is represented as a state vector within this vector space. The state vector representing the background information is projected onto each of the basis vectors within this multidimensional vector space. Essentially, this involves determining an amplitude on each of the basis vectors, with this amplitude representing the degree of concurrence between the qualities defining the basis vector (e.g., that Bill is an accountant, works in a tax office and is computer literate) and the details set out in the background evidence (Bill's description). If the evidence is broadly consistent with the state defined by the basis vector then the amplitude will be high. It might be argued that being an accountant, working in a tax office and being computer literate fits well with the background evidence about Bill and that therefore the state vector would be assigned a high amplitude on this particular basis vector. Importantly, the probabilities of events and combinations of events are said to be determined by these amplitudes and where all of the events in question can be represented within the same multidimensional space, judgements should be broadly consistent with probability theory. However, some events, for example, whether or not Bill plays jazz for a hobby, may not be readily incorporated in a particular vector space. Hobbies may not naturally be represented within the accountancy-tax office-computer literate space, in that consideration of occupations and working conditions may not be readily merged with consideration of hobbies, that is, the two relevant vector spaces are incompatible. In this situation, in order to evaluate the conjunctive probability 'Bill is an accountant and plays jazz for a hobby', the two incompatible vector spaces have to be considered separately and sequentially. Busemeyer et al. propose that the vector space relating to Bill being an accountant is considered first, after which the vector space relating to playing jazz is rotated so that a judgement can be made concerning the two (incompatible) events comprising the conjunction. Crucially, this rotation modifies the amplitudes on the hobbies vector space, potentially increasing the perceived likelihood that Bill plays jazz and thereby inflating the conjunctive probability. This may occur because focussing on the occupations/working conditions vector space results in the other aspects of Bill's description, leaving the focus of attention so that, following rotation, playing jazz for a hobby seems more plausible given the new perspective defined by having first focussed on the accountancy vector space.

9. Inductive confirmation

Moving away from the perspective of quantum probability, a substantially different account of the conjunction fallacy has been proposed by Tentori et al. (2013).

Central to their account is the concept of inductive confirmation. It has already been noted that the fallacy is most commonly found where the conjunction pairs a likely with an unlikely (LU) event. Other accounts of the fallacy propose that, whether via representativeness, signed summation, weighted averaging, or some other process, the probability of the likely event distorts the estimated conjunctive probability, often giving rise to the conjunction fallacy. Tentori et al. maintained that it is not the probability assigned to the likely event that potentially gives rise to the fallacy but rather it is the extent to which the likely event derives its probability by being inductively confirmed. That is to say, the likely event will only give rise to the fallacy when it derives its probability through being congruent and consistent with background information that is provided (i.e., where it is inductively confirmed by the evidence provided in the vignette; e.g., that Linda is a feminist). Where the probability, although likely, has nothing to do with the background information, for example, where it is based on the base rate or frequency of the event in the general population (e.g., Linda owns an umbrella), Tentori et al. maintained that the fallacy will not occur. Consistent with this proposition, Tentori et al. found that the incidence of the fallacy was indeed higher where the likely event was also inductively confirmed relative to where this was not the case.

10. Random noise or sampling errors

In sharp contrast to most of the existing research literature, Costello and Watts (2014) adopted a radically different perspective in which the fallacy is a by-product of an error-prone sampling process in an otherwise generally normative and rational process. In Costello and Watts' conceptual framework, it is assumed that discrete episodes, potentially consisting of a number of specific events, are stored separately in memory. For each episode, information about the different events making up that episode is retained in a binary manner indicating whether, for that particular episode, the event in question occurred. For example, an episode might be the morning commute to work and specific events might include whether or not the train was on time and whether or not it was raining. A whole series of episodes can be processed in order to determine the likelihood of specific events, for example, the probability that the train was on time or that it was raining. This is achieved by some automatic process in which the number of episodes (morning commutes) in which it was raining is tallied; then, by dividing by the total number of episodes, the likelihood of rain can be determined. This process is, however, subject to error in that a minority of episodes in which it was raining are processed as being dry while some episodes when it was dry are processed as having rain. Importantly, both single events and joint events (conjunctions and disjunctions) are stored independently in episodic memory, so when evaluating the likelihood of a conjunction, for example, the relevant conjunctive event is processed directly, removing the need for individuals to combine the two component event probabilities by applying the appropriate normative rule.

Costello and Watts demonstrated that, although giving rise to biases in individual probability judgements and sometimes to fallacies, such a process produces relatively

normative outcomes when the probabilities are combined. In essence, the errors or biases cancel out. For example, the following expression derived from the addition rule in probability has an expected value of zero:

$$P(A) + P(B) - P(A \,\&\, B) - P(A \text{ or } B) = 0 \tag{14}$$

Costello and Watts have shown that when individuals' estimates of the events (PE) are combined, even though each individual estimate is subject to error, the resulting distribution is broadly normal and centred on zero, that is,

$$PE(A) + PE(B) - PE(A \,\&\, B) - PE(A \text{ or } B) = 0 \tag{15}$$

Costello and Watts (2014) also demonstrated that other combinations of single and joint events, on average, similarly yield values consistent with probability theory. Crucially, they maintained that for the individual, deviations from the key normative values are non-systematic and, as a product of this error prone sampling process, essentially random. In their original account, Costello and Watts asserted that the magnitude of the sampling error, which they represented by some parameter 'd', is constant across single event and joint probabilities. However, in one of my own as yet unpublished studies, I have demonstrated that d is significantly larger for joint events and that for some individuals this results in *systematic* violations of the addition rule in that for them the above expression reliably deviates from zero. I have also demonstrated that judgements deviate systematically from the addition rule depending on the magnitude of the component events, for example, where P(A) and P(B) are both unlikely events, the above expression tends to have a negative or at least significantly smaller overall value (Fisk, Marshall, Stock, & Rogers, under review). In a more recent version of his theory, Costello (personal communication) has modified the original assumptions to allow the value of d to vary, which he claims can then account for my own results.

Conclusions

Virtually all accounts of the conjunction fallacy have placed emphasis on the role of the representative, more likely, inductively confirmed event in giving rise to the fallacy. Regarding the representativeness heuristic (Tversky & Kahneman, 1983; see Chapter 11), the extent to which the background evidence is representative of the likely event gives rise to the conjunction fallacy because the presence of the likely event in the conjunction renders the evidence more representative of the conjunctive event compared to the other (generally unrepresentative) component. Thus, those viewing the evidence as particularly representative of the likely event will be more likely to commit the conjunction fallacy, producing a larger estimate of the conjunctive probability relative to those who do not view the evidence as representative. Similarly, according to the signed summation account (Yates & Carlson, 1986), as the QLI associated with the likely event increases, so does the conjunctive

QLI. Because the QLI determines the probability, it follows that as the probability assigned to the likely event increases so would the conjunctive probability. Thus both these accounts are consistent with the proposition that increases in the more likely component event probability will lead to increases in the conjunctive probability, in some cases leading to the conjunction fallacy. The same proposition follows from the CEST (Donovan & Epstein, 1997) theory because where problems are concrete and unnatural it is claimed that judgements will be based on heuristics such as representativeness thereby giving rise to the conjunction fallacy. Clearly, averaging models (e.g., Birnbaum et al., 1990; Nilsson et al., 2009) also allow a role for both component probabilities in influencing the value assigned to the conjunctive probability. Likewise, in the case of quantum probability (Busemeyer et al., 2011), the amplitude assigned to the likely component event will directly influence both the estimated likely component event probability and the conjunctive probability. Tentori et al. (2013) have adopted a rather more refined position. As I have noted above, they have claimed that the likely event does not inevitably give rise to the fallacy in LU combinations. Rather, they have observed that the likely event only appears to be associated with the fallacy when that event is consistent with the background evidence. Nonetheless, under these circumstances, the likely event would be expected to influence the conjunctive probability and this tendency would increase as the degree of inductive confirmation increases.

The crucial point is that virtually all of the existing theories that we have summarised above, which place emphasis on the more likely or inductively confirmed event, are inconsistent with my own findings. Indeed, in several studies from my laboratory, conducted over a number of years, my co-workers and I examined the role played by the less likely and the more likely component event probabilities in determining the probability assigned to the joint (conjunctive or disjunctive) event. Our results consistently revealed that it was the less likely component probability which played the major role in determining the probability assigned to the conjunction while the more likely event played the same role in determining the probability assigned to the disjunction. Indeed, it was often the case that the other component event probability failed to account for any statistically significant variance in the probability assigned to the joint event. These results were confirmed using a number of different methodological and statistical approaches, ranging from factorial ANOVA (Fisk, 2002) to multiple regression (Fisk & Pidgeon, 1996, 1997) and with different problem contents and formats (Fisk et al., under review; Fisk & Pidgeon, 1998; Kariyazono, 1991; Rogers et al., 2011). Our findings, which have been largely ignored in the subsequent literature (with the notable exception of Nilsson and co-workers), are thus quite clearly incompatible with most existing accounts of conjunction errors/fallacies.

My own view of the fallacy has continued to evolve and has now perhaps reached a resting state. I now see the fallacy as a process of fixation and adjustment. For conjunctive (disjunctive) judgements, the fixation point is the less (more) likely component event probability. Such a proposition is consistent with surprise theory and the evidence for it comes from the numerous studies in which I have

demonstrated that only the fixation point is statically significant in determining the joint event probability or at least accounts for the overwhelming proportion of explained variance (Fisk, 2002; Fisk & Pidgeon, 1996, 1997, 1998; Rogers et al., 2011). Indeed, crucially, many participants actually set the joint probability exactly equal to the fixation point (Fisk et al., under review; Hertwig & Chase, 1998; Sloman, Over, Slovak, & Stibel, 2003). There may be deviations from this pattern where, for example, events are conditionally related (Fisk & Pidgeon, 1998) or where the two component probabilities are in close proximity. I believe that the process of adjustment is essentially qualitative in nature and not amenable to quantification. Many decades ago, Yates and Carlson (1986) identified a number of strategies that their participants used in determining the conjunctive probability. The responses of some participants were consistent with representativeness or signed summation (both sometimes resulting in the fallacy), while other participants, although not apparently utilising the normative rule, had a basic understanding of the extension principle (i.e., that the joint event was contained within the extension of the less likely component event and therefore cannot have a greater probability). To these decision processes can also be added the role of inductive confirmation (Tentori et al., 2013). I have argued that these basic understandings determine whether the adjustment from the fixation point is a positive or negative one but crucially that the magnitude of the adjustment itself is initially arbitrary in nature. Although initially arbitrary, the adjustment may be guided by cues. For example, it is clear from our own research (Fisk et al., under review) and that of others (e.g., Costello & Watts, 2014) that many participants simply select the neighbouring point on the response scale with respect to the particular fixation point in question. Importantly our own research has shown that the magnitude of the adjustment does not appear to be related to the other (non-fixated) component probability estimate.

Interestingly and paradoxically, very recently, Costello (personal communication) has pointed out that our regression results (in which only the fixation point is statistically significant in predicting the joint event) are also consistent with normatively derived probabilities. In order to demonstrate this, using a quasi-random process, Costello generated normatively consistent component and conjunctive probabilities. He then observed that if the resulting conjunctive probability is regressed on the components, the smaller component paradoxically does indeed account for virtually all of the variance in the conjunctive probability. On reflection, this is perhaps not difficult to understand and indeed I made an equivalent observation in the previous edition of this volume (Fisk, 2004). Specifically, it is clear that the same absolute change in the component probability will result in very different changes in the normative conjunctive probability, depending on whether the change is in the less likely or more likely component. Specifically, given the multiplicative nature of the normative relationship, changes in the magnitude of the less likely component event probability will have a far greater influence on the magnitude of the conjunctive probability than similar changes in the more likely component event probability. In the previous edition of this book, I argued that the focus on the

less likely event, to the exclusion of the other, was an ecologically sound strategy because it was the changes in the less likely event which had the most influence on the conjunctive probability.

In conclusion, there has been a paradox at the heart of the conjunction fallacy literature for several decades. While my own research and that of others has failed to show that the likely, representative, or inductively confirmed event has any direct influence on the probability assigned to the conjunction, most accounts of the fallacy focus entirely on this aspect. It is up to the reader to decide whether my own account is plausible or whether decision makers are essentially rational after all (Costello & Watts, 2014).

Summary

- The conjunction fallacy arises when individuals assign probabilities to conjunctive events that exceed the probabilities assigned to the component events making up the conjunction.
- The incidence of the fallacy can be as high as 90% but varies as a function of problem content, the manner in which the problem is framed, and the type of response that is required.
- Early explanations of the fallacy, including representativeness, linguistic interpretation, 'signed summation', applying a different probabilistic rule, and averaging, have failed to provide a satisfactory account of the phenomenon.
- More recent accounts focussing on inductive confirmation, quantum probability, configural weighted averaging, and the attribution of the fallacy to random errors also appear to have their limitations.
- My own view is that the fallacy is the product of a two-stage judgement process in which the smaller component probability serves as a fixation point after which a quasi random adjustment is made based on a particular qualitative judgement process, for example, similarity or the extension principle.

Further reading

Tversky and Kahneman's (1983) classic exposition of the conjunction fallacy is essential reading for anyone interested in the phenomenon. Tentori et al. (2013) present their own perspective on the existing research literature while promoting their own inductive confirmation account of the fallacy. Lastly, the rather different combinatorial approach adopted by Nilsson et al. (2009) makes very worthwhile reading.

Note

1 The conjunction of two events is defined as: $P(A\&B) = P(A) \times P(B|A)$. Since probabilities cannot exceed 1, then $P(B|A) \leq 1$; therefore $P(A) \times P(B|A) \leq P(A)$ and so $P(A\&B) \leq P(A)$. The same reasoning holds for P(B).

References

Agnoli, F., & Krantz, D. H. (1989). Suppressing natural heuristics by formal instruction: The case of the conjunction fallacy. *Cognitive Psychology*, *21*, 515–550.

Birnbaum, M. H., Anderson, C., & Hynan, L. G. (1990). Theories of bias in probability judgment. In J.-P. Caverni, J. M. Fabre, & M. Gonzalez (Eds.), *Cognitive biases* (pp. 477–498). Amsterdam: North-Holland.

Bonini, N., Tentori, K., & Osherson, D. (2004). A different conjunction fallacy. *Mind & Language*, *19*(2), 199–210.

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review*, *118*(2), 193–218.

Chiesi, F., Gronchi, G., & Primi, C. (2008). Age-trend-related differences in tasks involving conjunctive probabilistic reasoning. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *62*(3), 188–191.

Cohen, J., Dearnaley, E. J., & Hansel, C. E. M. (1958). Skill and chance: Variations in estimates of skill with an increasing element of chance. *British Journal of Psychology*, *49*, 319–323.

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463–480.

Donovan, S., & Epstein, S. (1997). The difficulty in the Linda conjunction problem can be attributed to its simultaneous concrete and unnatural representation, and not to conversational implicature. *Journal of Experimental Social Psychology*, *33*, 1–20.

Evans, J. St. B. T., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, 77, 197–213.

Fabre, J., Caverni, J. P., & Jungermann, H. (1995). Causality does influence conjunctive probability judgements if context and design allow for it. *Organizational Behavior and Human Decision Processes*, *63*, 1–5.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, *50*, 123–129.

Fisk, J. E. (1996). The conjunction effect: Fallacy or Bayesian inference? *Organizational Behavior and Human Decision Processes*, *67*, 76–90.

Fisk, J. E. (2002). Judgements under uncertainty: Representativeness or potential surprise? *British Journal of Psychology*, *93*, 431–449.

Fisk, J. E. (2004). Conjunction fallacy. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (pp. 23–42). Hove, UK: Psychology Press.

Fisk, J. E., Marshall, D. A., Stock, R., & Rogers, P. (under review). Probability judgements do not appear to be 'surprisingly rational' after all.

Fisk, J. E., & Pidgeon, N. (1996). Component probabilities and the conjunction fallacy: Resolving signed summation and the low component model in a contingent approach. *Acta Psychologica*, *94*, 1–20.

Fisk, J. E., & Pidgeon, N. (1997). The conjunction fallacy: The case for the existence of competing heuristic strategies. *British Journal of Psychology*, *88*, 1–27.

Fisk, J. E., & Pidgeon, N. (1998). Conditional probabilities, potential surprise, and the conjunction fallacy. *Quarterly Journal of Experimental Psychology*, *51A*, 655–681.

Gavanski, I., & Roskos-Ewoldsen, D. R. (1991). Representativeness and conjoint probability. *Journal of Personality and Social Psychology*, *61*, 181–194.

Goldsmith, R. W. (1978). Assessing probabilities of compound events in a judicial context. *Scandinavian Journal of Psychology*, *19*, 103–110.

Hertwig, R., & Chase, V. M. (1998). Many reasons or just one: How response mode affects reasoning in the conjunction problem. *Thinking & Reasoning*, *4*, 319–352.

Kariyazono, A. (1991). The study of conjunction fallacy in the probability judgment task of compound event. *Japanese Journal of Psychonomic Science*, *10*, 57–64. (English abstract).

Morrier, D. M., & Borgida, E. (1984). The conjunction fallacy: A task specific phenomenon? *Personality and Social Psychology Bulletin*, *10*, 243–252.

Nilsson, H., Winman, A., Juslin, P., & Hansson, G. (2009). Linda is not a bearded lady: Configural weighting and adding as the cause of extension errors. *Journal of Experimental Psychology: General*, *138*(4), 517–534.

Rogers, P., Fisk, J. E., & Wiltshire, D. (2011). Paranormal belief and the conjunction fallacy: Controlling for temporal relatedness and potential surprise differentials in component events. *Applied Cognitive Psychology, 25*, 692–702.

Shackle, G. L. S. (1969). *Decision, order and time in human affairs*. Cambridge: Cambridge University Press.

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes, 91*, 296–309.

Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General, 142*, 235–255.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547–567.

Wells, G. L. (1985). The conjunction error and the representativeness heuristic. *Social Cognition*, *3*, 266–279.

Wolford, G., Taylor, H. A., & Beck, J. R. (1990). The conjunction fallacy? *Memory & Cognition*, *18*, 47–53.

Wyer, R. S. (1970). Quantitative prediction of belief and opinion change: A further test of a subjective probability model. *Journal of Personality and Social Psychology*, *16*, 559–570.

Yates, J. F., & Carlson, B. W. (1986). Conjunction errors: Evidence for multiple judgement procedures including "Signed Summation". *Organizational Behaviour and Human Decision Processes*, *37*, 230–253.

2

BASE-RATE NEGLECT

Gordon Pennycook and Valerie A. Thompson

The "base-rate" refers to the *a-priori* probability of an event or outcome. For example, there are 19 professional hockey players who play for the Toronto Maple Leafs at any given moment during the hockey season. On game day, 38 out of 2.5 million people in Toronto are National Hockey League (NHL) players (i.e., the Leafs and their opponent). Thus, the base-rate probability that a randomly encountered person in Toronto on game day is a NHL player is 38/2,500,000 or .00152%. Base-rate neglect refers to the phenomenon whereby people ignore or undervalue that probability, typically in lieu of less informative, but more intuitively appealing information about an individual case (Kahneman & Tversky, 1973). Thus, even if a Toronto resident were to come across a tall, burly, hockey-stick-wielding man wearing a Maple Leafs jersey, the probability that he actually plays for the team (and is not simply a fan wearing the jersey on his way to a recreational hockey game) is very small. An everyday example of how base-rates such as this can be neglected can be illustrated with a thought experiment.

Imagine owning a car that constantly breaks down and, after a few years of this, you have finally found someone to purchase it. This additional bit of money allows you to purchase a new car – one that will hopefully be more reliable – though you do not have a very large budget. You have narrowed the list of potential cars to two options (which are approximately the same cost): a Subaru and a Fiat. The most recent issue of *Consumer Reports* indicates that Subaru owners typically have fewer mechanical problems than do the Fiat owners and that the Subaru was more highly rated by experts. However, you also happen to have an uncle who once owned a Subaru. He informs you that his Subaru had multiple, very costly problems. His suggestion is to go with the Fiat, which he feels is a more reliable car.

Which car do you purchase? There is a strong temptation in situations such as this to ignore or underweight the base-rate probability of mechanical issues (i.e., based on the large sample of owners' experiences and expert opinion described

in *Consumer Reports*) in lieu of the more appealing single case (i.e., based on your trusted uncle's experience). Indeed, when given hypothetical scenarios of this sort, participants often choose the 'Fiat' response – that is, the car that is probabilistically more likely to have mechanical issues but that has an intuitive appeal (Fong, Krantz, & Nisbitt, 1986). Clearly, neglecting the base-rates can be expensive, if one opts for the repair-needy Fiat over the more reliable Subaru. The neglect or underweighting of base-rate probabilities has been demonstrated in a wide range of situations in both experimental and applied settings (Barbey & Sloman, 2007). In this chapter we will outline some of the ways that the base-rate fallacy has been investigated, discuss a debate about the extent of base-rate use, and, focusing on one particular form of base-rate neglect, we will outline recent work on the cognitive mechanisms that underlie the tendency to underweight or ignore base-rate information.

Base-rate neglect in many forms

The term *base-rate neglect* applies to any case where a prior probability is not sufficiently weighted in reasoning. As a consequence, base-rate neglect takes many forms, a selection of which is illustrated in Text box 2.1. The purpose of the first

TEXT BOX 2.1 FREQUENTLY INVESTIGATED VARIETIES OF BASE-RATE NEGLECT

1 Person 'A' was selected at random from a group consisting of all University of Waterloo students majoring in either GENERAL ARTS or STATISTICS. Person 'A' is orderly, organized, precise, practical and realistic. Is Person A's major more likely to be: GENERAL ARTS or STATISTICS? (Pennycook et al., 2012)

2 In a study 1000 people were tested. Among the participants there were 995 nurses and 5 doctors. Paul is a randomly chosen participant of this study. Paul is 34 years old. He lives in a beautiful home in a posh suburb. He is well spoken and very interested in politics. He invests a lot of time in his career. Is Paul more likely to be: a doctor or a nurse? (Pennycook & Thompson, 2012)

3 The probability of breast cancer is 1% for a woman at age forty who participates in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? ___%. (Gigerenzer & Hoffrage, 1995)

two problems (1–2) is to create a conflict between base-rate and individuating (stereotype) information and see what proportion of individuals select the base-rate response. The first example (1) is referred to as an "implicit base-rate" problem, because the relevant base-rate is not mentioned. Instead, there is an explicit description of a set of stereotypes (orderly, precise, etc.) that suggests to most people that "Person A" is more likely to be a Statistics major. Although implicit, the base-rate is nonetheless relevant to deciding which option is more likely. At most universities, there are far more students in General Arts than there are in Statistics (the ratio at the University of Waterloo in Canada is ~24:1); this discrepancy is so large that an individual who is orderly, precise, etc. is far more likely to major in General Arts than Statistics, despite the stereotypical association with a Statistics major. Nonetheless, when students at the University of Waterloo were given a set of these problems, only 21% selected the response consistent with the base-rate (in this case, General Arts). Moreover, response time analyses indicated that participants did not appear to recognize the relevance of the base-rate probability (i.e., they spent the same amount of time reasoning as when the problem contained no conflict between base-rate and stereotype); indicating that the 21% of the time when base-rate responses were given likely resulted from individuals having atypical stereotypes and not an understanding of the base-rate (Pennycook, Fugelsang, & Koehler, 2012). Thus, at least about 80% of the students in the study completely neglected the base-rates, but probably more did as well.

Consider now the second example (2), a base-rate problem modelled on Kahneman and Tversky's (1973) "lawyer-engineer problem" (described subsequently). This problem is very similar to the first example (1), with the key exception that the base-rate probability (995 nurses, 5 doctors) is explicitly stated. The conflict between base-rate probability (indicating that Paul is a nurse) and the stereotypical information (indicating that Paul is a doctor) is now, in theory, plainly obvious. Nonetheless, Pennycook et al. (2012) found that participants selected the base-rate response only 24% of the time on a set of problems like this. Thus, participants typically fail to sufficiently weight base-rate information even when the prior probability is extreme (the prior probability that Paul is a doctor is 0.5%) and explicitly stated in the problem. However, in this case, participants do take longer when giving the stereotypical response to versions of this problem where the base-rate and stereotypes point to alternative responses, suggesting that they do successfully recognize (at some level) that the base-rate conflicts with the stereotype (Pennycook et al., 2012).

The third example (3) – referred to as the "mammography problem" (e.g., Eddy, 1982) – is quite different in form and plainly more complex than the previous two. The problem contains a base-rate (1% of women have breast cancer), but also includes information about the hit-rate (80% chance of a positive mammogram if breast cancer is present) and the false-alarm rate (9.6% chance of a positive mammogram if breast cancer is absent). Given the hypothesis (H) that a random 42-year-old woman has a positive mammogram (the observed datum, D), the probability that she actually has breast cancer [$P(H|D)$] can be determined using Bayes' theorem

(the specific details of which are not important for present purposes; interested readers can see Birnbaum, 2004; Kurzenhäuser & Lücking, 2004): (0.01)(0.80) / [(0.01)(0.80) + (0.99)(0.096)]. Based on this calculation, there is a 7.8% chance that the woman has breast cancer given a positive mammogram. Given the complexity of this operation, it is perhaps no surprise that very few people are able to generate the correct solution (e.g., 16% in Gigerenzer & Hoffrage, 1995). Indeed, even physicians have a great deal of difficulty with problems such as this (Hammerton, 1973). The median response on these types of problem is to report a number close to the hit-rate (i.e., 80%; Barbey & Sloman, 2007), ignoring the fact that the base-rate indicates that the probability of cancer is very rare.

Results and discussion

The experiment described in Text box 2.2 was the first in Kahneman and Tversky's (1973) seminal work on base-rate neglect. Their results can be found in Table 2.1. They found a very strong positive correlation between similarity (Condition 1) and likelihood (Condition 2), $r = .97$. The rankings in the two groups were nearly identical! In contrast, not only was there no *positive* correlation between mean judged

TEXT BOX 2.2 A CLASSROOM DEMONSTRATION OF BASE-RATE NEGLECT BASED ON KAHNEMAN AND TVERSKY (1973)

Method

Participants

This is a between-participant experiment and requires three roughly equal groups. Fortunately, the phenomenon under investigation is very robust and groups of 10 or more individuals should suffice. If necessary, the experiment can be changed to a within-participant design and participants can complete each condition in order (Condition 1, Condition 2, Condition 3).

Materials and procedure

Participants in each of the conditions will be given a slightly different task. It would be best for each participant to only see the task assigned to their condition, although the experiment should work regardless. The conditions are as follows:

Condition 1

Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail

finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and by flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feeling and little sympathy for other people and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense.

How similar is Tom W. to the typical graduate student in [your country] in each of the following nine fields of graduate specialization? Please rank the following nine fields of graduate specialization in order of the relative similarity of Tom W. relative to the prototypical student in [your country]. Rank from 1 to 9, using each rank only once.

Business Administration
Computer Science
Engineering
Humanities and Education
Law
Library Science
Medicine
Physical and Life Sciences
Social Science and Social Work

Condition 2

Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and by flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feeling and little sympathy for other people and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense.

The preceding personality sketch of Tom W. was written during Tom's senior year in high school by a psychologist, on the basis of projective tests. Tom W. is currently a graduate student in [your country]. Please rank the following nine fields of graduate specialization in order of the likelihood that Tom W. is now a graduate student in each of these fields. Rank from 1 to 9, using each rank only once.

Business Administration
Computer Science
Engineering
Humanities and Education
Law
Library Science
Medicine
Physical and Life Sciences
Social Science and Social Work

Condition 3

Consider all first year graduate students in [your country] today. Please write down your best guesses about the percentage of these students who are now enrolled in each of the following nine fields of specialization:

Business Administration
Computer Science
Engineering
Humanities and Education
Law
Library Science
Medicine
Physical and Life Sciences
Social Science and Social Work

Note: If universities in your country do not offer exactly these fields of specialization, please replace them with the fields that come closest.

Analysis

Participants in Condition 3 provided subjective base-rates. Note that it does not matter if these perceived base-rates are accurate; they just need to accurately represent the typical opinions of the participants in the other conditions. This can be double-checked by having participants in Conditions 1 (the "similarity" group) and 2 (the "likelihood" group) also complete Condition 3. To compute the base-rates, one needs to compute a mean of the estimated base-rates for each graduate specialization (in %). To compute the measure of similarity and of likelihood, compute the mean ranks for each graduate specialization for Conditions 1 and 2 (respectively).

On this task, base-rate neglect occurs when participants' likelihood ratings are informed by the similarity of each person to a stereotype rather than the base-rates. To demonstrate this, one needs to correlate the likelihood judgments with both the similarity rankings and the base-rate estimates. For this, put the responses for each condition in separate columns of the same table (see Table 2.1). If participants used the base-rate of graduate specialization (i.e., the responses for Condition 3) when judging the likelihood of graduate specialization (Condition 2), then there should be a positive correlation between the responses for Conditions 2 and 3. If, on the other hand, participants used stereotypes (i.e., the responses for Condition 1) to determine the likelihood of graduate specialization, there should be a positive correlation between the responses for Conditions 1 and 2. Finally, these two correlation coefficients can be compared to assess which source of information was more influential. If the similarity was more influential than the base-rate probability, the correlation between responses for Conditions 1 and 2 should be larger than the correlation between responses for Conditions 2 and 3.

TABLE 2.1 Estimated base-rates of the nine areas of graduate specialization and summary of similarity and likelihood ratings for Tom W. (Kahneman & Tversky, 1973)

Graduate specialization area	*Mean similarity rank*	*Mean likelihood rank*	*Mean judged base-rate (in %)*
Business Administration	3.9	4.3	15
Computer Science	2.1	2.5	7
Engineering	2.9	2.6	9
Humanities and Education	7.2	7.6	20
Law	5.9	5.2	9
Library Science	4.2	4.7	3
Medicine	5.9	5.8	8
Physical and Life Sciences	4.5	4.3	12
Social Science and Social Work	8.2	8.0	17

base-rate (Condition 3) and likelihood judgments (Condition 2), but the correlation was actually *negative*, $r = -.65$. This is because Tom W. sounds most like someone in computer science or engineering (associated with relatively low base-rates) and least like someone in humanities and social sciences (associated with relatively high base-rates). Clearly, base-rates were not taken into account when participants were asked to judge the *likelihood* that Tom W. was a student in these graduate specializations. Thus, this is an example of base-rate neglect.

Theoretical accounts

The term "base-rate neglect" implies that information about base-rates is completely ignored. In this section, we will summarize research that examines whether this is true or whether the term "neglect" is a bit of a misnomer. We will then move to some more recent research on the cognitive mechanisms that underlie base-rate neglect. As was made evident in Text box 2.1, there are many different manifestations of base-rate neglect. Not surprisingly, therefore, there is no unified theoretical account of base-rate neglect. Instead, the theorizing tends to be centred on the cognitive mechanisms thought to underlie particular forms of base-rate neglect, which may not be applicable to other forms.

Are base-rates ignored?

The earliest data on base-rate neglect seemed to indicate that people essentially ignore base-rates when making judgments. The Tom W. problem (Text box 2.2) from Kahneman and Tversky (1973) is a particularly striking example of this. In the case of the more complex mammography problem (Text box 2.1) (and others like it), Eddy (1982) found that fewer than 5% of respondents were able to correctly solve the problem and Hammerton (1973) found only nominally better

performance in a group of physicians (10% correct). Results such as this led many researchers to conclude that base-rates are ignored. However, subsequent research showed that this conclusion was too pessimistic (e.g., Gigerenzer & Hoffrage, 1995).

For example, there are a number of conditions under which people can be made sensitive to base-rate information (Birnbaum, 2004): Participants are more sensitive to base-rates when given multiple problems with varying base-rate probabilities (Fischhoff & Bar-Hillel, 1984) or if given problems where the base-rates come *after* individuating information (e.g., stereotypes; Krosnick, Li, & Lehman, 1990). Sensitivity to base-rates is also facilitated by manipulations that make a causal link between the base-rate and the judged case explicit (e.g., Bar-Hillel, 1980). Consider the two examples in Text box 2.3. In the first example (1), the Cab problem, the color distribution of the cabs is the base-rate information (85% blue, 15% green) and the accuracy of witness identification is the individuating information (80%

TEXT BOX 2.3 CAUSALITY IN BASE-RATE PROBLEMS (BAR-HILLEL, 1980)

1 Two cab companies operate in a given city, the Blue and the Green (according to the color of cab they run). 85% percent of the cabs in the city are Blue, and the remaining 15% are Green. A cab was involved in a hit-and-run accident at night. A witness later identified the cab as a Green cab. The court tested the witness' ability to distinguish between Blue and Green cabs under night time visibility conditions. It found that the witness was able to identify each color correctly about 80% of the time, but confused it with the other color about 20% of the time.

What do you think are the chances that the errant cab was indeed Green, as the witness claimed?

2 A large water-pumping facility is operated simultaneously by two giant motors. The motors are virtually identical (in terms of model, age, etc.), except that a long history of breakdowns in the facility has shown that one motor, call it A, was responsible for 85% of the breakdowns, whereas the other, B, caused 15% of the breakdowns only. To mend a motor, it must be idled and taken apart, an expensive and drawn out affair. Therefore, several tests are usually done to get some prior notion of which motor to tackle. One of these tests employs a mechanical device which operates, roughly, by pointing at the motor whose magnetic field is weaker. In 4 cases out of 5, a faulty motor creates a weaker field, but in 1 case out of 5, this effect may be accidentally caused. Suppose a breakdown has just occurred. The device is pointed at motor B.

What do you think are the chances that motor B is responsible for this breakdown?

hit-rate and 20% false-alarm rate). According to Bayes' theorem, the probability that the cab was green is 41% because the base-rate and individuating information need to be integrated [(0.8/0.2) * (0.15/0.85)]. However, the typical response for this problem is 80% (i.e., the hit-rate; Bar-Hillel, 1980). However, if there is a causal link between the base-rate and individuating information, people are more inclined to combine them. To understand why, consider the second example (2) in Text box 2.3, the Motor problem. This problem is exactly the same in terms of base-rate (A = 85%, B = 15%) and individuating information (80% hit-rate and 20% false-alarm rate), and the correct answer is therefore also 41%. The key difference between the problems is that the Motor problem makes it clear that the base-rate is an *attribute* of the two motors. Or, in other words, it is readily apparent that the base-rate is causally linked to the function of the motors. This manipulation highlighted the importance of the base-rate and, as a consequence, over 60% of the participants gave a response that indicated sensitivity to the base-rate (Bar-Hillel, 1980).

These results would not be expected if base-rates are completely ignored. Importantly, there is also evidence that base-rates can not only enter into judgment, but that people are capable of using them correctly. Consider a modified version of the mammography problem:

> 10 out of every 1,000 women at age forty who participate in routine screening have breast cancer. 8 out of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography. Pretend that you were presented a new representative sample of women at age forty who got a positive mammography in routine screening. How many of these women do you expect to actually have breast cancer? __ out of 100

Here, the probabilities (1%; 80%; 9.6%) have been presented in terms of natural frequency formats (10 out of 1,000; 8 out of 10; 95 out of 990). This relatively straightforward manipulation was sufficient to increase performance by a factor of ~3 (46% v. 16% accuracy; Gigerenzer & Hoffrage, 1995, see also Kurzenhäuser & Lücking, 2004; Tversky & Kahneman, 1983). On the basis of these results, Gigerenzer and Hoffrage (1995) argued that frequency formats are more easily understood by participants because they are consistent with the sequential way that information is acquired in the context of natural sampling. This view is typically associated with evolutionary psychology and, in particular, the idea that humans have evolved an intuitive way of dealing with base-rates that requires the right sort of conditions to be triggered but that does not require conscious deliberation (see Barbey & Sloman, 2007, for a review).

Now consider Kahneman and Tversky's (1973) "lawyer-engineer" problem (mentioned above):

> A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written.

> You will find on your forms a description, chosen at random from the 100 available descriptions. Please indicate your probability that the person described is an engineer, on a scale from 0 to 100.
>
> The same task has been performed by a panel of experts, who were highly accurate in assigning probabilities to the various descriptions. You will be paid a bonus to the extent that your estimate comes close to those of the expert panel.
>
> Here is the description: Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles. The probability that Jack is one of the 30 engineers in the sample of 100 is ___%.

This problem contains base-rate information that is easily understood, such that most participants successfully give the base-rate response on versions of this problem that do not contain stereotypes (Pennycook & Thompson, 2012). Nonetheless, as discussed above, participants do not typically give the base-rate response when base-rates are in conflict with stereotypical information. Rather, they rely primarily on the representative information and underweight (but not necessarily neglect) the base-rate.

In a recent experiment, Pennycook and Thompson (2012) investigated the degree to which participants used base-rates in a set of problems (18 in total) of the lawyer-engineer type. They included two between-participant conditions (see Text box 2.4): A standard base-rate (BR) condition (1) and a no base-rate (NoBR) condition (3). The goal of this manipulation was to see what sort of influence base-rates had on probability judgments. If base-rates are completely ignored, judgments should not differ between conditions.

TEXT BOX 2.4 CONDITIONS FROM PENNYCOOK AND THOMPSON (2012)

1. In a study 1000 people were tested. Among the participants there were 995 nurses and 5 doctors. Paul is a randomly chosen participant of this study. Paul is 34 years old. He lives in a beautiful home in a posh suburb. He is well spoken and very interested in politics. He invests a lot of time in his career. What is the probability (0–100) that Paul is a nurse? [Base-rate Condition (BR); Incongruent]
2. In a study 1000 people were tested. Among the participants there were 995 who live in a condo and 5 who live in a farmhouse. Kurt is a randomly chosen participant of this study. Kurt works on Wall Street and is single. He works long hours and wears Armani suits to work. He likes

wearing sunglasses. What is the probability (0–100) that Kurt lives in a condo? [Base-rate Condition (BR); Congruent]

3 In a study 1000 people were tested. Among the participants there were nurses and doctors. Paul is a randomly chosen participant of this study. Paul is 34 years old. He lives in a beautiful home in a posh suburb. He is well spoken and very interested in politics. He invests a lot of time in his career. What is the probability (0–100) that Paul is a nurse? [No base-rate Condition (NoBR)]

In addition, Pennycook and Thompson (2012) also included a within-subject manipulation of congruency such that in the BR condition the base-rates were sometimes inconsistent with the stereotypes, akin to the lawyer-engineer problem, and sometimes consistent with the stereotypes. For example, in Text box 2.4 the first problem (1) is incongruent because the base-rate indicates that Paul is very likely to be a nurse but the stereotypes suggest that Paul sounds more like a doctor. In contrast, the second problem (2) is congruent because the base-rate indicates that Kurt is likely to own a condo and the stereotypes are more consistent with a condo owner than a farmhouse owner. [Note: Congruency could not be manipulated in the NoBR condition due to the lack of base-rates.] This manipulation was included because it is possible that base-rates may be used differently depending on their association with the individuating stereotypical information.

Pennycook and Thompson's (2012) key results can be found in Figures 2.1 and 2.2, which show the distribution of probability estimates across the conditions. The comparison between the distributions of probability estimates given base-rates (BR) or not (NoBR) shows that base-rates had a substantial influence on judgments. Moreover, the form that this took differed depending on whether base-rates and stereotypes were consistent (congruent) or inconsistent (incongruent). As is evident from Figure 2.1, the vast majority of probability estimates for congruent problems in the BR condition were 90% or higher (*Mean* = 88.6%). In contrast, probability estimates ranged fairly equally from 50% to 90% when participants were only given stereotypical information in the NoBR condition (*Mean* = 68.5%). This pattern of results indicates that base-rates not only informed participants' judgments, but that the modal response was a *combination* of base-rate and individuating (stereotypical) information. Participants integrated the two sources of information, as is necessary for Bayes' theorem.

An entirely different pattern was evident when base-rates and stereotypes were inconsistent (incongruent). In the BR condition (Figure 2.2), one cluster of responses was similar to that found for the congruent problems (i.e., responses 90% and higher); however, there was a second cluster more consistent with traditional base-rate neglect findings. Namely, many responses were 10% and lower, which is very *inconsistent* with the presented base-rates (and therefore very *consistent* with the presented stereotypes).

FIGURE 2.1 Distribution of probability estimates for congruent problems (from "Reasoning with base rates is routine, relatively effortless, and context dependent" by Gordon Pennycook and Valerie A. Thompson, 2012, *Psychonomic Bulletin & Review, 19*, 531, © Psychonomic Society, Inc. 2012. Adapted with permission of the publisher). For BR condition, high responses are consistent with both stereotypes and base-rates. For No-BR condition, high responses are consistent with stereotypes. Note that problems were not "congruent" in the No-BR condition due to the lack of base-rate information. They are the exact congruent problems from the base-rate condition with base-rates removed. The counterbalancing was such that, in the no base-rate condition, the problems would have been "congruent" or "incongruent" if base-rates had been included.

FIGURE 2.2 Distribution of probability estimates for incongruent problems (from "Reasoning with base rates is routine, relatively effortless, and context dependent" by Gordon Pennycook and Valerie A. Thompson, 2012, *Psychonomic Bulletin & Review, 19*, 532, © Psychonomic Society, Inc. 2012. Adapted with permission of the publisher). For the BR condition, high estimates are consistent with base-rates and low estimates are consistent with stereotypes. For the No-BR condition, high responses are consistent with stereotypes. Note that problems were not "incongruent" in the No-BR condition due to the lack of base-rate information. They are the exact incongruent problems from the base-rate condition with base-rates removed. The counterbalancing was such that, in the no base-rate condition, the problems would have been "congruent" or "incongruent" if base-rates had been included.

This contrasts starkly with the responses for the No-BR conditions and the congruent BR condition. It seems that when base-rates and stereotypes point to different responses, participants no longer integrate them. Rather, they select one or the other source of information and give a relatively extreme response. If participants successfully integrated base-rate and individuating information for incongruent problems most of the probability judgments would be clustered somewhere in the middle of the distribution. As is evident from Figure 2.2, this did not happen. This pattern of results indicates that base-rates are sometimes very influential (i.e., when they are consistent with stereotypes) but are also sometimes neglected in lieu of stereotypes. Moreover, some people give base-rate responses even when base-rates and stereotypes are in conflict. This may be due, in part, to the fact that participants received multiple problems with slightly variable base-rate information (Kahneman & Frederick, 2005). Thus, to understand these findings, we need to move past the debate about whether base-rates are ignored and into a more theoretical discussion of the cognitive mechanisms that underlie the use (or neglect) of base-rates.

Dual-process theory and base-rate neglect

The dominant explanation for why some types of individuating information (e.g., stereotypes) are favored so heavily over base-rates appeals to dual-processing. Dual-process theory relates to the idea that there are two types of processes by which humans make judgments and decisions (Evans & Stanovich, 2013): Type 1 processes that are autonomous, fast, and high capacity, and Type 2 processes that are reflective, slow, and resource demanding. The role of Type 1 processes is to provide default outputs which can be accepted, rejected, or modified as explicit representations in working memory via Type 2 processing (Evans & Stanovich, 2013).

The initial explanation of base-rate neglect anticipated these later developments in dual-process theory. Namely, participants were thought to form a rapid response using a "representativeness heuristic" (Kahneman & Tversky, 1973; see Chapter 11 in this volume). That is, rather than answering the rather difficult question regarding the probability of group membership, participants formed their judgment on the basis of which group the personality description seemed more representative of. This explanation is quite amenable with dual-process theory. Specifically, stereotypes cue an intuitive "Type 1" response (based on representativeness) and base-rates require deliberate "Type 2" reasoning processes to enter into judgment (e.g., Kahneman, 2003). Since humans typically forego costly Type 2 processing in favor of less effortful Type 1 processing (Stanovich & West, 2000), stereotypes are naturally favored over base-rates. Indeed, participants who are more disposed to analytic thought (as indexed by both self-report and performance measures) are more likely to give the base-rate response for problems of the lawyer-engineer type (e.g., Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014).

Although all seem to agree that individuating information like stereotypes is a very intuitive source of information, there is clearly some disagreement about how

difficult base-rates are to use. Kahneman's (2003) dual-process account holds that base-rates require resource-demanding reasoning processes, whereas other accounts hold that base-rates do not require any deliberation at all and are actually quite intuitive (at least, when they are in the right format; e.g., Gigerenzer & Hoffrage, 1995). Fortunately, recent experiments have started to clarify this issue.

Recall the Pennycook and Thompson (2012) experiment where participants were given problems with base-rates (BR) or without base-rates (NoBR; Text box 2.4). The researchers also included an additional within-participant manipulation that was quite revealing. Participants were asked to respond to each problem twice: First they provided whatever response initially popped into their head (an intuitive response given under a time deadline) and then they responded to the same question again with a final answer given over free time. When offered this chance to rethink their intuitive response, participants were just as likely to shift toward the stereotype as they were toward the base-rate. Moreover, many participants gave responses consistent with the base-rates even when they gave the first response that came to mind. These results indicate that responses based on both stereotypes and base-rates can be either intuitive *or* reflective.

This conclusion was supported by a set of experiments by Pennycook, Trippas, Handley, and Thompson (2014). Participants were given a set of base-rate problems of the lawyer-engineer type and were explicitly instructed how to respond to each problem: 1) Statistics instructions highlighted the importance of base-rates in determining the likelihood of group membership, and 2) Belief instructions highlighted the importance of belief-based information (stereotypes) in determining the likelihood of group membership. If base-rates require slow Type 2 processing and belief judgments are made using fast Type 1 processing, then responding according to the base-rates should not interfere with judgments when responding based on beliefs. Instead, across three experiments, participants had just as much difficulty responding according to belief instructions as they did with statistics instructions. Namely, probability estimates were less accurate, confidence was lower, and response time was longer when base-rates and stereotypes conflicted *regardless of the instruction manipulation*. This result was replicated when participants were put under a strict time deadline. This represents rather striking evidence that the use of base-rates can be intuitive.

If base-rates use is (at least sometimes) intuitive, how can we explain the preponderance of stereotypical responses to problems of the lawyer-engineer type? The answer to this question requires a more nuanced understanding of what it means for something to be "intuitive" (Thompson, 2014). It may be that responses based on *both* stereotypes and base-rates can be intuitive, but that the former are typically *more* accessible in that stereotypes cue a response that comes to mind more quickly and fluently than the response cued by the base-rate information (Pennycook, Fugelsang, & Koehler, 2015). This would leave the stereotype as the default response and, as a consequence, even if participants recognized the importance of the base-rates they would still need to inhibit and override the default stereotypical response (De Neys & Franssens, 2009). Since humans are miserly information

processors (Stanovich & West, 2000), this resource-demanding Type 2 response is often foregone; hence, base-rate neglect.

The miserly processing account explains why base-rate responses are more common among individuals who are more disposed to analytic thought (e.g., Pennycook, Cheyne et al., 2014) and more intelligent (e.g., Thompson & Johnson, 2014). It also explains why base-rate responding has been linked with additional psychological factors. For example, those who are less prone to base-rate neglect are more likely to be skeptical of religious and paranormal claims (e.g., Pennycook, Cheyne et al., 2014). In other words, those who are more likely to question their initial intuitions about stereotypes in the context of a base-rate problem are also more likely to question widely held and often quite intuitive supernatural beliefs. Moreover, people who are better able to detect the conflict between base-rates and stereotypes may also be better able to detect the intrinsic conflict between ubiquitous materialistic intuitions (e.g., that beings cannot pass through solid objects) and immaterial beliefs (e.g., that an angel can pass through solid objects; Pennycook, Cheyne et al., 2014). The low-level conflict between base-rates and stereotypes evident for (at least some) base-rate problems may also be a key trigger of Type 2 processing (i.e., something that *causes* people to think; Pennycook et al., 2015) and the way people respond to this type of conflict (which would presumably be prevalent in many different domains, such as the conflict between supernatural and materialist beliefs) is a key aspect of human cognition. Moreover, an understanding of these mechanisms could be used to potentially devise interventions that help reduce or even overcome base-rate neglect. For example, in order to facilitate conflict detection, Pennycook et al. (2015) gave participants multiple problems with extreme base-rates that were presented *after* stereotypes. In that condition, participants actually gave *more* base-rate responses than stereotypical ones.

Conclusions

Base-rate neglect is a very robust phenomenon that comes in many forms (Barbey & Sloman, 2007). Nonetheless, much evidence suggests that base-rates are not always neglected (e.g., Gigerenzer & Hoffrage, 1995). Although the calculation of probability estimates may require deliberative reasoning at least some of the time for at least some people, base-rates influence judgment at a lower, more intuitive level (Pennycook, Trippas, et al., 2014). Further, people appear to be able to detect the conflict between base-rates and individuating information (e.g., stereotypes) that is common in certain forms of base-rate neglect (Pennycook et al., 2012). This conflict detection is an important bottom-up source of analytic reasoning (Pennycook et al., 2015) and, as a consequence, base-rate neglect has been linked to psychological phenomena not typically associated with reasoning and decision making (e.g., religious belief; Pennycook, Cheyne, et al., 2014). Thus not only does base-rate neglect have important consequences in applied areas (e.g., medical decision making; Eddy, 1982), but at a more theoretical level, the study of base-rate neglect has revealed novel insights about the interaction between deliberate and analytic

thinking in ways that has informed our understanding of a wide array of cognitive illusions and reasoning biases.

Summary

- People often neglect or underweight base-rate probabilities when other (typically more intuitive) information is available.
- Base-rate neglect has been demonstrated using a wide range of tasks across many experiments.
- There are ways to improve people's reasoning with base-rates, though they are typically still underweighted.
- Base-rates can be processed without deliberative reasoning (though typically not as intuitively as stereotypical individuating information).
- Base-rate neglect emerges as a consequence of an interaction between intuitive and reflective processes.

Further reading

Barbey and Sloman (2007) represents an extensive review of base-rate neglect research in the context of competing models. Pennycook, Trippas, et al. (2014) outline typical dual-process account of base-rate neglect and provide evidence for a revised version of that model. Pennycook et al. (2015) use base-rate problems to illustrate the key role of conflict detection as a source of analytic engagement.

Acknowledgements

The cited studies of the authors were funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada. This came in the form of a master's and doctoral funding for GP and a Discovery Grant for VT.

References

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioural and Brain Sciences, 30*, 241–256.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*, 211–233.

Birnbaum, M. H. (2004). Base rates in bayesian inference. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (pp. 43–60). New York, NY: Psychology Press.

De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. *Cognition, 113*, 45–61.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–67). New York: Cambridge University Press.

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives in Psychological Science, 8*, 223–241.

Fischhoff, B., & Bar-Hillel, M. (1984). Focusing techniques: A shortcut to improving probability judgments? *Organizational Behavior and Human Performance, 34*, 175–194.

Fong, G. T., Krantz, D. H., & Nisbitt, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18*, 253–292.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704.

Hammerton, M. (1973). A case of radical probability estimation. *Journal of Experimental Psychology, 101*, 252–254.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58,* 697–720.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgement. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–293). Cambridge: Cambridge University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

Krosnick, J. A., Li, F., & Lehman, D. R. (1990). Conversational conventions, order of information acquisition, and the effect of base rates and individuating information on social judgments. *Journal of Personality and Social Psychology, 59*, 1140–1152.

Kurzenhäuser, S., & Lücking, A. (2004). Statistical formats in Bayesian inference. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (pp. 61–77). New York, NY: Psychology Press.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition, 42*, 1–10.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition, 124*, 101–106.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology, 80*, 34–72.

Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context-dependent. *Psychonomic Bulletin & Review, 19*, 528–534.

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base-rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 544–554.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23,* 645–726.

Thompson, V. (2014). What intuitions are. . . and are not. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 35–75). Burlington: Academic Press.

Thompson, V., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning, 20*, 215–244.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 91,* 293–315.

3

PROBABILITY MATCHING

Ben R. Newell and Christin Schulze

Imagine you are in your local casino simultaneously playing two slot machines. You feed money into each machine, press the buttons, watch the reels spin, and occasionally win. You notice that the machine on your left pays off more regularly—about 70% of the time—than the one on your right—only about 30% of the time. Both machines pay out the same amount when you win. In terms of the amount of money you feed into each machine, what should you do?

A few options present themselves: You could abandon the *left* machine, reasoning that the "time was coming" for the right machine to start paying out; you could abandon the *right* machine and concentrate your efforts on the higher-paying left machine; or you could try to play both machines in rough proportion to how often they are paying out (i.e., play the left machine 70% of the time and the right 30%).

To maximize your winnings in any simple repeated binary choice task like this—in which a payoff is available with higher probability given one response than another—one should *always* choose the higher probability option. In the slot machine example, this amounts to only playing the *left,* higher-paying machine.

Surprisingly, participants faced with an analogous experimental situation often fail to identify the optimal choice strategy. Instead, many participants show a striking violation of rational choice theory by tending to "match" probabilities: That is to say, they try to play the machines in proportion to how often they are paying out (70% on the left and 30% on the right). This phenomenon is known as *probability matching* and this chapter examines why it arises and the factors that affect it.

Definition

Probability matching refers to a behavior in which people allocate responses to two options in proportion to their relative pay-off probabilities. Whenever one option has a higher relative pay-off probability than the other, probability matching is

irrational (even if the payoff probabilities only differ by a very small amount). It is easy to see why. Let's suppose, like in the slot machine example, that a payout of fixed size—say \$0.10—is given with probability $P(A) = .70$ for choosing option A and $P(B) = .30$ for choosing option B, in a simple two-option experiment (A and B could be buttons on a computer screen). A probability-matching strategy yields 58% accuracy because a reward will be received on 70% of the 70% of trials on which A is chosen ($.70 \times .70 = .49$) and on 30% of the 30% of trials on which B is chosen ($.30 \times .30 = .09$), adding to $.49 + .09 = .58$. Thus on average, across 100 trials of an experiment a participant would earn \$5.80. In contrast, choosing option A on every trial yields an accuracy of 70% ($.70 \times 1.0 = .70$) and thus a payoff of \$7.00. So a participant should only ever choose option A.

This example makes it clear that the best strategy for *each separate decision* is to select the option with the higher probability of payoff. On any trial, the expected payoff for choosing A is higher than the expected payoff for choosing B. There is never a time at which the expected payoff is higher for the lower-yielding option, so one should never choose it.

There are caveats to this advice, some of which were already hinted at in the opening slot-machine example. First, in situations where the payoff probabilities are not known in advance, there needs to be an initial period of experimentation to discover which option (if either) is dominant. Second, the advice to perseverate on option A assumes that the payoff probabilities are stationary (they do not systematically change over time). If one has suspicions that the probabilities will change then further periods of experimentation might make sense (Mehlhorn et al., 2015; Navarro, Newell, & Schulze, 2016). We consider different situations in which probability matching is observed, namely situations where choices are made under uncertainty about payoff probabilities and those in which there is no uncertainty and nothing—from a normative perspective—to learn about the problem facing participants.

Varieties of probability matching

Probability matching when probabilities are unknown

Classic demonstrations of probability matching are akin to the slot-machine example in that people must learn via experience and experimentation which of two options has a higher relative pay-off probability. For instance, in Neimark and Shuford's (1959) study, one response alternative was correct on 67% of trials and the other on 33%. Participants were given 100 trials to learn these relative probabilities: The optimal response on every trial is, of course, to choose the response that has the higher likelihood of reward. However, at the end of training participants were choosing the 67% correct option on about 67% of trials—that is, they were probability *matching*.

Participants do not always "slavishly" match the probabilities, however; sometimes responding is more characteristic of "over-matching", that is, a tendency to choose the more rewarded option with a probability closer to 1.0. In Edwards'

(1961) study, for example, participants' asymptotic choice probability for a response that had a payoff probability of .70 was .83. Vulkan (2000) reviewed several other demonstrations of matching and over-matching in such standard probability-learning tasks. These failures to maximize are surprising given the seeming simplicity of the two-option problem facing participants. Moreover, unlike many examples of apparently irrational behavior detailed in other chapters of this book, participants make repeated (rather than "one-shot") choices and receive a steady flow of feedback from their behavior, which should provide a strong impetus to help them find the optimal choice strategy.

Not everyone, however, has drawn such a pessimistic conclusion about the (ir) rationality of behavior in probability-learning tasks. Shanks, Tunney, and McCarthy (2002) presented a more optimistic outlook. They explored simple probability-learning tasks in which large performance-related financial incentives were provided (over USD$100 in some cases), together with meaningful and regular feedback, and extensive training (up to 1,800 learning trials in one experiment). The results were fairly clear in demonstrating that large proportions of participants (about 70%) can maximize their payoffs when maximizing is defined as a run of at least 50 consecutive choices of the optimal response. Many participants quite comfortably exceeded this criterion. Each of the three factors mentioned above contributed to participants' ability to maximize: Both feedback and payoffs affected the overall likelihood of exclusively choosing the best alternative, and stable performance was not reached until after many hundreds of trials.

These results suggest that given enough feedback and the right kind of incentives, many participants will learn which of two options is dominant, and will be convinced that pay-off probabilities remain stationary. One would imagine, then, that in situations in which there is no uncertainty about pay-off probabilities and no reason to suspect change, probability matching would be very rare. Surprisingly, even when there is "nothing" to learn, probability matching remains prevalent—as we shall see in the next section.

Probability matching when probabilities are known

Consider the following problem (referred to hereafter as the "die problem"): A die with 4 black faces and 2 white faces will be rolled 60 times. Before each roll you will be asked to predict which color (black or white) will show up once the die is rolled. You will be given 1$ for each correct prediction. Assume that you want to make as much money as possible. What strategy would you use in order to make as much money as possible by making the most correct predictions?

a *Go by intuition, switching when there have been too many of one color or the other.*
b *Predict the more likely color (black) on most of the rolls but occasionally, after a long run of blacks, predict white.*
c *Make predictions according to the frequency of occurrence (4 of 6 for black and 2 of 6 for white). That is, predict twice as many blacks as whites.*

d *Predict the more likely color (black) on all of the 60 rolls.*
e *Predict more black than white, but switching back and forth depending on "runs" of one color or the other.*

The description of this problem tells you everything you need to know in order to make the rational choice. The relative pay-off probabilities are known, the pay-off amount is known, and assuming you believe that the die is fair and will not be altered throughout the 60 rolls, then you know that pay-off probabilities will remain stationary.

Yet, people seem to find this problem quite difficult: Newell and Rakow (2007), for example, found that a majority of their participants chose Strategy C. That is to say, they chose the strategy which exemplifies a "probability matching"—and hence incorrect—response. Strategy D is the maximizing strategy, the correct answer in the sense that it will yield more correct predictions than any of the other strategies. It will yield around 40 correct predictions ($4/6 \times 60$), a level that cannot be exceeded given the randomness of the outcomes. Strategy C—the matching strategy—is expected to yield only around $(4/6 \times 4/6 \times 60) + (2/6 \times 2/6 \times 60) = 33$ correct predictions, because whenever a white outcome is predicted there is only a 1 in 3 probability that it will occur. Such predictions are "wasted" because a black outcome is twice as likely. Why are people tempted to match under such transparent conditions?

Subsequent research has examined two factors which seem to be important. Newell and Rakow (2007) gave participants the scenario above, but also allowed some participants to see the outcome of each roll after making their prediction. This "hybrid" design ensured that participants had all of the necessary information for solving the problem via a description, but were also given the opportunity to "learn" about the problem via trial-by-trial prediction and feedback. Somewhat surprisingly, the experience of making repeated choices had an initial *negative* effect on responding; during the first block of 50 trials, participants who both read a description and experienced making choices showed fewer maximizing responses than did participants who only read the description. Interestingly, sufficient trial-by-trial outcome feedback eroded this initial negative influence of experience; asymptotic trial-by-trial choice (after 300 trials) was slightly superior to description-based decisions (Newell & Rakow, 2007). Notably, these effects of feedback were observed despite the feedback being completely uninformative and therefore normatively irrelevant—the die was fair, the probabilities were stationary, and the outcome probabilities were already known before any feedback was received. The finding that feedback nonetheless influenced the rate of maximizing is puzzling because it suggests that people only gradually "learn" to choose optimally in a task that (normatively) requires no learning.

A second notable aspect is that a proportion of participants may simply fail to realize that maximizing is the optimal response without being explicitly encouraged to consider different ways of responding, and so are seduced into responding in a way that is "representative" of the probabilities underlying the task (Kahneman &

Tversky, 1972; see Chapter 11 in this volume). When a hint is provided to explicitly contrast the matching and maximizing strategies, and to consider how frequently each would pay off, some participants are pushed toward maximizing (Koehler & James, 2010; Newell, Koehler, James, Rakow, & van Ravenzwaaij, 2013).

The fact that most people need such extensive support to maximize probability in repeated choice (in the form of feedback, hints, incentives, and training) suggests that identifying the optimal solution in this context may require considerable cognitive effort (Koehler & James, 2009, 2014). Consistent with that idea, Schulze and Newell (2015) have shown that probability matching is virtually absent when the cognitive power of a small group of people is combined via group discussion. Both the emergence of erroneous probability matching in individual choice and its reduction in a group choice context can be illustrated in the classroom by comparing the choices of individual students with those of small groups of students in a described task, as outlined in Text box 3.1.

TEXT BOX 3.1 PROBABILITY MATCHING IN GROUP AND INDIVIDUAL CHOICE: CLASSROOM DEMONSTRATION 1

Method

For an effective demonstration of probability matching in the classroom, we recommend a basic version of the "die problem" introduced above that has been used in numerous studies of risky choice (e.g., Gal & Baron, 1996; James & Koehler, 2011; Newell et al., 2013; Newell & Rakow, 2007; Schulze & Newell, 2015).

Participants

The general magnitude of previously reported effect sizes is typically large. To accommodate the size of typical classroom settings, we recommend comparing 10 groups (of three to four members) to 10 individuals, that is, an overall sample size between 40 and 50 people.

Material

Participants read a summary of the die problem (see Appendix), which asks them to make a series of predictions about the outcome of many rolls (e.g., 50) of a fair, colored die on a sheet of paper. On this die, more sides are of one color than another (e.g., a ten-sided die with 7 black and 3 white sides), and it is used to determine outcomes probabilistically.

Design

A simple design for demonstrating both the emergence and the elimination of probability matching in repeated choice involves two between-subjects conditions, namely individual choice and group choice. Participants in the individual-choice condition predict outcomes for a series of rolls of a fair die by themselves; participants in the group-choice condition discuss strategies for solving the die problem within a small group of people and make predictions as a group. Group-choice research typically contrasts the performance of group members to that of an equivalent number of independent individuals to determine how group performance compares to the performance of the best, second-best, third-best, and fourth-best individuals (see Laughlin, VanderStoep, & Hollingshead, 1991). To accommodate typical classroom sizes we propose a simplified design which compares groups with the average individual. The dependent measure is participants' proportion of choices to the more probable outcome, that is, the proportion of choices to the dominant color on the die.

Procedure

The two conditions of the experiment can be run in parallel. Participants in the group-choice condition are asked to assemble into small groups of three to four people; participants in the individual-choice condition should be encouraged to work by themselves. All participants are instructed that the choice task involves a fair die that is going to be rolled numerous times by the experimenter and that their job is to predict which side will come up for each roll. For each correct prediction, participants earn one point and it should be stressed that the goal of the task is to earn as many points as possible. Participants are given 10 minutes to think about/discuss effective strategies for solving this "die problem" and to generate a series of outcome predictions on a piece of paper. In order to encourage students' best effort, instructions should emphasize that the die is not biased or loaded in any way but fair in the sense that each side has an equal chance of turning up once the die is rolled (see the Appendix for example instructions). Once students have completed their predictions, the die is rolled repeatedly and participants can assess their prediction accuracy and ensuing point score.

Analysis

Participants' choices to the die's dominant color are counted and converted into proportions for each individual and each group. The average choice proportions of the interacting groups can be compared to those of the individuals

with an independent-samples *t* test. Additionally, each dominant color-choice proportion that is within 5% of the dominant outcome probability can be classified as probability matching and each dominant color-choice proportion that is 95% or higher can be classified as probability maximizing to determine the number of probability maximizers and matchers in the group and individual choice contexts (see, e.g., Newell & Rakow, 2007). These numbers can then be compared between groups and individuals using chi-squared tests.

Changing the probability distribution: Classroom demonstration 2

A very simple classroom demonstration can use the die problem but vary, between groups, the distribution of colors on the die.

For example, a class of 30 students could be divided into two even groups. One group is given a description in which a 10-sided die has 6 black sides and 4 white sides and the other group is told the die has 8 black and 2 white sides. All individuals are then invited to make a series of 50 predictions (as per Classroom demonstration 1). The results can be analyzed in the same way too: average proportions of dominant color choices compared via independent samples *t* test, and counts of "maximizers" and "matchers" according to the +/– 5% classification.

Following precedent (e.g., Newell & Rakow, 2007) an effect of distribution is likely to be found (more dominant color choices in the 8/2 than the 6/4 group) but evidence of matching or over-matching is still likely to be present in both groups. A class discussion could focus on why this pattern arises: It suggests some sensitivity to the implications of the probability distribution but that, on its own, this knowledge is insufficient to push everyone to maximize probability.

Classroom demonstration

Results

Comparing group and individual choice in two variants of the "die problem", Schulze and Newell (2015) found that groups performed as well as their best individual members and better than almost everyone else. This group-choice advantage emerged when participants generated a sequence of choices before receiving outcome feedback (as suggested in the classroom demonstration; see Schulze & Newell, 2015, Experiment 1a) and when participants made trial-by-trial choices while receiving outcome feedback following each decision (see Schulze & Newell, 2015, Experiment 1b). Figure 3.1 depicts the results of their first experiment. To assess groups' and individuals' engagement in probability matching and maximizing, each

FIGURE 3.1 From left to right, proportion of dominant color choices for groups, best, second-best, third-best, and fourth-best performing individuals in the individual-choice condition of Experiment 1a reported in Schulze and Newell (2015). Individuals' level of choice performance was determined by grouping sets of four, randomly drawn, independent individuals and ranking their dominant color choice proportions from highest to lowest. Where two or more independent individuals in a grouped set returned identical choice proportions, rank positions were assigned randomly. Bar graphs plot mean ± standard error choice proportions across groups/individuals. Small squares/circles plot the choice proportion of one group/individual in the respective decision making condition. Probability matching is indicated by the dashed line at .70.

dominant color-choice proportion that fell within 5% of the average outcome probability was classified as probability matching and each dominant color-choice proportion that was 95% or higher was classified as probability maximizing (see also Newell & Rakow, 2007; Schulze, van Ravenzwaaij, & Newell, 2015). Group choice nearly eradicated probability matching in these experiments, and the majority of groups maximized probability. By contrast, individuals who completed a choice task by themselves showed similar propensities towards probability matching and maximizing; both strategies were adopted with comparable frequency. Why did this group advantage arise? A key determinant of groups' capacity to overcome (or succumb) to individual biases such as probability matching is the nature of the group process (Kerr, MacCoun, & Kramer, 1996), which, in turn, hinges on the demonstrability of the correct solution (Laughlin & Ellis, 1986). If the demonstrability is low, individual opinions need to be aggregated (e.g., via a majority voting process), which may disseminate biases held by individual members. If the bias can be demonstrated, it may be overcome by the group based on a single member's insight. In post-task assessments of the group dynamics, most group members indicated that their group either followed a "truth wins" decision process and implemented the most effective strategy proposed by one of the members—even if they would not

have thought of the same solution by themselves—or was unanimous in identifying the correct solution (see Schulze & Newell, 2015, Experiment 1b). This suggests that the erroneous nature of probability matching can be demonstrated and overridden through small group discussions—even if a large portion of the group's members initially held different beliefs.

Theoretical accounts

Our review so far highlights that probability matching is a pervasive phenomenon. It is observed when pay-off probabilities are known in advance and when they have to be learned via experience. In this section we consider two theoretical accounts that can roughly be categorized into two conflicting views of the phenomenon. The classic perspective dismisses probability matching as a simple error that is driven by cognitive and environmental limitations. The alternative view points to countless real-world choice environments in which probability matching would normally occur and posits that people may simply *mis*apply an adaptive response to experimental tasks where it is no longer appropriate. In other words, researchers disagree about whether probability matching is a simple mistake or a sophisticated adaptation to complex environments.

Probability matching as an error

From the perspective of rational-choice theory, probability matching is unequivocally an error, and a surprising one. Indeed, the implications of the phenomenon for a *homo-economicus* view have caught the attention of many decision theorists. For example, the Nobel Prize-winning economist Kenneth Arrow (1958, p. 14) noted that:

> The remarkable thing about this is that the asymptotic behavior of the individual, even after an indefinitely large amount of learning, is not the optimal behavior [. . .] We have here an experimental situation which is essentially of an economic nature in the sense of seeking to achieve a maximum of expected reward, and yet the individual does not in fact, at any point, even in a limit, reach the optimal behavior.

Our review is consistent with this characterization of probability matching as an error, but an error that can be corrected—or at least reduced—provided the learning environment is appropriately structured (Friedman, 1998). As we have seen, probability matching can be reduced by providing hints about the effectiveness of different strategies, by detailed regular feedback, extensive opportunities to learn, and large financial incentives. Putting heads together in a group-choice context can also reduce or even eliminate the tendency to match probabilities.

However, labeling probability matching as an error and outlining the situations in which it can be reduced or eliminated still begs the question *why* the error is

committed in the first place. A rather uninteresting, but nonetheless important, consideration is that the error arises because people simply misapprehend the requirements of the task (see, e.g., Hilton, 1995). This is a possibility in the "die problem" (where payoff probabilities are known in advance), in which participants might think that they are being asked to produce a potential *outcome distribution* for the die rolls, rather than to make independent predictions for *each* trial. If this were the case, probability matching would be an appropriate strategy. While possible, this interpretation seems unlikely because the incentive structure of the task (payment for *each correct prediction*) clearly emphasizes the importance of independent predictions (see Text box 3.1 and the Appendix).

Such emphasis notwithstanding, some research does point to the role of thinking about sequence-wide (or distributional) expectations rather than individual trial outcomes. James and Koehler (2011) found that instructions designed to diminish expectations regarding the aggregate outcome of an upcoming sequence of events (such as rolls of a die) reduced probability matching and increased maximizing behavior. For example, they gave participants a version of the die problem in which, before making predictions, they were asked either "In 10 rolls of the die, how many times would you expect each outcome?" or "On any individual roll of the die, which color is more likely to be rolled?". The former question—claimed to prime a global, sequence-wide expectation—led to more probability-matching behavior than the latter "local focus" question, which encouraged consideration of individual trial contingencies. Gao and Corter (2015) found a similar result demonstrating that instructions emphasizing correct predictions of 4-trial sequences led to more probability matching than instructions that emphasized single-trial predictions.

Proponents of dual-system views of judgment and decision making argue that these kinds of results can be explained via the interaction of two systems, namely an intuitive system that provides the spontaneous (but incorrect) matching response, and a deliberative system which overcomes this initial tendency and then maximizes provided instructional and/or environmental scaffolding is in place (e.g., James & Koehler, 2011; Koehler & James, 2014). As such, probability matching can be subsumed within the more general framework of "attribute substitution" (Kahneman & Frederick, 2002), which has been used to explain several of the biases reviewed in other chapters in this volume (e.g., availability and representativeness; see Chapters 10 and 11, respectively).

Although such a dual-process view is appealing, it can be questioned on both theoretical and empirical grounds. From an empirical perspective, the primacy of the matching response is challenged by data showing that young children appear to exhibit less probability matching than adults (Derks & Paclisanu, 1967). In addition, cognitive load seems to *decrease* the tendency to probability match (e.g., Wolford, Newman, Miller, & Wig, 2004—though see the discussion in the next section, "Probability matching as an adaptive response"). Such findings appear difficult to reconcile with the idea that probability matching arises from an intuitive, effortless system. At a theoretical level, there are broader challenges for the dual-process perspective in terms of how the purported systems interact, dominate, and control

choice (e.g., Keren, 2013; Newell et al., 2015; Newell & Shanks, 2014). Much of this debate is beyond the scope of the current chapter but nevertheless should give pause for thought when considering the adequacy of a two-system explanation of probability matching.

Perhaps one of the most enduring explanations for the persistence of probability matching, and one that does not require recourse to competing systems, is that the utility of predicting the rare event might outweigh the often small monetary benefit of maximizing, and might alleviate the boredom of continually making the same response (e.g., Goodnow, 1955; Shanks et al., 2002). In the words of one participant from the Goodnow study (in which the reinforcement was 70 % left, 30 % right): "It's easy to win on the left; the real skill comes in winning on the right." As was pointed out by Goodnow, "One could win on the right key by betting on it very often but . . . this would be no more a test of skill than deer-hunting with a machine gun" (p. 113). In the "die problem", such attempts at "skilled" predictions of the rare event seem unlikely given the use of a familiar chance device (a die) with precise outcome probabilities (and no memory). However, in probability *learning* tasks, the temptation to predict the rare event could be stronger, especially if sequences are not truly random (i.e., not independent and identically distributed—see Fiorina, 1971) or monetary rewards are low (Vulkan, 2000). Giving people reason to believe or help to understand that a random process cannot "skillfully" be outdone decreases rates of probability matching. Fantino and Esfandiari (2002), for instance, told decision makers that the best they could do in a probability-learning task was to be correct on a subset of choice trials (bound by the frequency of the more likely outcome) and found an increase in maximizing responses under certain conditions. Likewise, allowing people to conclusively infer the randomness of the process that generates outcomes (Morse & Runquist, 1960; Peterson & Ulehla, 1965) or increasing the perceived randomness of a choice sequence (Wolford et al., 2004) facilitates maximization.

A further contributing factor to matching—or, more specifically, overmatching—is the reluctance to accept stationarity of payoffs. If a participant is confident about which option provides the higher pay-off (and thus chooses it most of the time) but also "checks" the other alternative at semi-regular intervals it will lead to responding that is superior to matching but remains suboptimal (i.e., overmatching). Indeed, recent research in related probability-learning tasks showed that it takes several hundred trials before participants respond as if they believe that an environment is truly stable (Navarro et al., 2016) as also seen in probability-matching experiments (Shanks et al., 2002).

Why do people show this exploratory persistence in simple choice tasks? One possibility is the relatively higher costs associated with assuming stationarity when, in fact, the environment is dynamic, compared to the cost of occasional exploration in an unchanging context (Navarro & Newell, 2014). Another likely reason is that most real-life decisions are made in *non-stationary* environments that change under the influence of countless external factors and in often highly systematic ways. Systematic patterns of probabilistic change, however, are inconsistent with the

independence assumption underlying typical repeated choice experiments (as well as repeated play at casino slot machines). Therefore, when participants bring their everyday experience to the lab, they may reasonably expect some form of temporal dependence in the outcome sequence and respond accordingly (Ayton & Fischer, 2004). In other words, probability matching may be an overlearned response from frequently encountered real-world settings in which spreading choices across available alternatives is, in fact, adaptive.

Probability matching as an adaptive response

Consider, for instance, an exploratory strategy that is highly adaptive in most natural environments but futile when outcome probabilities are stationary and independent: the search for patterns in an outcome sequence. If a fixed pattern did exist, a rule-searching strategy would prove superior to static probability maximizing. Because any predictable pattern must match the outcome frequencies, probability matching would emerge as a by-product of successful search behavior. Consistent with this idea, Gaissmaier and Schooler (2008) found that people who probability matched in the absence of patterns were more likely to detect regularities in the outcome sequence when patterns were introduced. Unturbe, Corominas, and Josep (2007) showed that the complexity of rules participants reported to have followed during a repeated choice task was positively related to probability matching. Moreover, in the absence of a fixed pattern but under the assumption that the learner holds the erroneous—but ecologically plausible—belief that outcomes share some form of temporal interdependence, probability matching is adopted by an *optimal* Bayesian learner using a maximizing decision rule (Green, Benson, Kersten, & Schrater, 2010).

Assuming that people search for patterns in outcome sequences also quite comfortably accounts for reduced probability-matching rates when cognitive resources are taxed by a secondary task (Wolford et al., 2004). This is because occupying cognitive resources also needed for pattern search would impede the ability to detect patterns and, by extension, reduce probability matching observable at the choice level. Yet findings from dual-task paradigms do not unequivocally support the idea that probability matching represents an ecologically rational response associated with the search for patterns. An alternative explanation, compatible with the "simple mistake" account described above, suggests that the effort needed to implement a choice strategy is favorably skewed toward probability maximizing (see Koehler & James, 2014). Optimal maximization typically involves only a simple repeated response to the same option—that is, little effort to carry out. By contrast, to implement an (inferior) matching strategy, choice proportions need to be tracked over many choice trials. This effort likely involves similar cognitive mechanisms as occupied by a concurrent (typically a working memory) task.

Addressing this issue, Schulze and Newell (2016) manipulated the allocation of choice options to physical response options in a repeated choice task so that repeatedly choosing the same physical response key resulted in either probability

matching or maximizing. Applied to a dual-task context, this manipulation can illuminate whether cognitive load causes people to repeatedly choose the same physical option rather than the maximizing option. The results indicated that increasing the implementation effort of maximizing under load markedly reduced the adoption of this strategy. That is, when the mapping of choice options to physical response keys changed and simple repetitions of the same motor response no longer resulted in optimal maximizing (but inferior matching), participants had a harder time choosing optimally. These results suggest that the effort involved in implementing a choice strategy is another key factor in shaping repeated risky choice.

Another characteristic of most real-world choice situations is that decision makers rarely choose in social isolation but often compete with others for the exploitation of limited resources. Yet the more competing individuals crowd the richest resource, the less each individual's gain. This means that spreading choices across available options can become beneficial (Gallistel, 1990; Gigerenzer, 2000). Schulze et al. (2015) examined the adequacy of human choice under competition in a repeated choice paradigm that involved a computerized opponent. Their choice task was identical to standard probability-learning paradigms—two choice options that award the same payoff with unequal odds—but the computer opponent and participant had to share potential rewards equally when both selected the same alternative. The behavior of the computer opponent reinforced one of two strategies: If the opponent probabilistically imitated participant choices, probability *matching* was optimal; if the opponent was indifferent, probability *maximizing* was optimal. Under both conditions, participants adopted accurate choice strategies at asymptote (either optimal matching or maximizing), which demonstrates that adaptive probability matching indeed emerges under competition.

In summary, people's tendency to probability match has been explained in terms of a simple mistake facilitated by deficiencies in the structure of task environments and in terms of an adaptive response to the structure of natural choice environments. These findings emphasize that appropriately structured choice contexts are paramount to promoting appropriate choice behavior.

Conclusions

Probability matching is of enduring interest to scholars of judgment and decision making because it provides a very simple—yet informative—window into how, when, and why people learn to make optimal choices. Indeed, insights from such simple two-option choice tasks have led to major developments in fundamental theories and models of learning (see Newell et al., 2015). Although much of the "classic" work on probability matching was done in the 1950s and '60s, the recent popularization of dual-system perspectives in the judgment and decision making literature (e.g., Kahneman, 2011) along with a resurgence of interest in simple feedback-based decisions (i.e., "decisions from experience", see Rakow & Newell, 2010) has thrust research on probability matching back

into the limelight. This renewed interest will no doubt bring further insight into whether probability matching should be best viewed as an error or a misapplied adaptive response.

Summary

- Probability matching refers to a behavior in which people allocate responses to two options in proportion to their relative pay-off probabilities. Whenever one option has a higher relative pay-off probability than the other, this behavior is irrational.
- Probability matching is highly robust and is observed when pay-off probabilities have to be learned from experience *and* when they are known in advance of any choices.
- Extensive training, adequate feedback, and appropriate incentives can reduce probability matching and promote rational probability maximizing.
- Providing hints about the effectiveness of strategies and giving groups the opportunity to discuss the problem reduces probability matching in described contexts (e.g., the die problem).
- Theoretical accounts contrast the idea that probability matching is an error with one suggesting it is a misapplied adaptive response to the structure of most *natural* environments.

Further reading

Koehler and James (2014) provide an excellent recent overview of probability matching, highlighting the dual-process perspective. Vulkan (2000) provides an earlier review of the phenomenon from an economist's perspective. Newell et al. (2015) discuss probability matching and maximizing in the context of a wider thesis on the importance of learning for optimal decision making.

Acknowledgements

The cited studies of the authors have been supported by grants from the Australian Research Council (ARC). The first author acknowledges salary support from an ARC Future Fellowship (FT110100151).

References

Arrow, K. J. (1958). Utilities, attitudes, choices: A review note. *Econometrica, 26*, 1–23.

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition*, *32*, 1369–1378.

Derks, P. L., & Paclisanu, M. I. (1967). Simple strategies in binary prediction by children and adults. *Journal of Experimental Psychology*, *73*, 278–285.

Edwards, W. (1961). Probability learning in 1000 trials. *Journal of Experimental Psychology, 62*, 385–394.

Fantino, E., & Esfandiari, A. (2002). Probability matching: Encouraging optimal responding in humans. *Canadian Journal of Experimental Psychology*, *56*, 58–63.

Fiorina, M. P. (1971). A note on probability matching and rational choice. *Behavioral Science, 16*, 158–166.

Friedman, D. (1998). Monty Hall's three doors: Construction and deconstruction of a choice anomaly. *American Economic Review, 88*, 933–946.

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition, 109,* 416–422.

Gal, I., & Baron, J. (1996). Understanding repeated simple choices. *Thinking & Reasoning*, *2*, 81–98.

Gallistel, C. R. (1990). *The organization of learning.* Cambridge, MA: MIT Press.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world.* New York, NY: Oxford University Press.

Goodnow, J. J. (1955). Determinants of choice-distribution in two-choice situations. *The American Journal of Psychology, 68,* 106–116.

Green, C. S., Benson, C., Kersten, D., & Schrater, P. (2010). Alterations in choice behavior by manipulations of world model. *Proceedings of the National Academy of Sciences, 107*, 16401–16406.

Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin, 118,* 248–271.

James, G., & Koehler, D. J. (2011). Banking on a bad bet: Probability matching in risky choice is linked to expectation generation. *Psychological Science*, *22*, 707–711.

Kahneman, D. (2011). *Thinking, fast and slow.* New York, NY: Allen Lane.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York, NY: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430–454.

Keren, G. (2013). A tale of two systems: A scientific advance or a theoretical stone soup? Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, *8*, 257–262.

Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in judgment: Comparing individuals and groups. *Psychological Review*, *103*, 687–719.

Koehler, D. J., & James, G. (2009). Probability matching in choice under uncertainty: Intuition versus deliberation. *Cognition*, *113*, 123–127.

Koehler, D. J., & James, G. (2010). Probability matching and strategy availability. *Memory & Cognition*, *38*, 667–676.

Koehler, D. J., & James, G. (2014). Probability matching, fast and slow. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 61, pp. 103–131). San Diego, CA: Elsevier.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*, 177–189.

Laughlin, P. R., VanderStoep, S. W., & Hollingshead, A. B. (1991). Collective versus individual induction: Recognition of truth, rejection of error, and collective information processing. *Journal of Personality and Social Psychology*, *61*, 50–67.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fielder, K., & Gonzalez, C. (2015). Unpacking the exploration-exploitation tradeoff: A synthesis of human and animal literatures. *Decision, 2*, 191–215.

Morse, E. B., & Runquist, W. N. (1960). Probability-matching with an unscheduled random sequence. *The American Journal of Psychology*, *73*, 603–607.

Navarro, D. J., & Newell, B. R. (2014). Information versus reward in a changing world. In P. Bellow, M. Guarani, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the Cognitive Science Society* (pp. 1054–1059). Austin, TX: Cognitive Science Society.

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore-exploit dilemma in static and dynamic environments. *Cognitive Psychology, 85*, 43–77.

Neimark, E. D., & Shuford, E. H. (1959). Comparison of predictions and estimates in a probability learning situation. *Journal of Experimental Psychology, 57*, 294–298.

Newell, B. R., Koehler, D. J., James, G., Rakow, T., & van Ravenzwaaij, D. (2013). Probability matching in risky choice: The interplay of feedback and strategy availability. *Memory & Cognition*, *41*, 329–338.

Newell, B. R., Lagnado, D. A., & Shanks, D. R. (2015). *Straight choices: The psychology of decision making* (2nd ed.). Hove, UK: Psychology Press.

Newell, B. R., & Rakow, T. (2007). The role of experience in decisions from description. *Psychonomic Bulletin & Review*, *14*, 1133–1139.

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences, 37,* 1–63.

Peterson, C. R., & Ulehla, Z. J. (1965). Sequential patterns and maximizing. *Journal of Experimental Psychology*, *69*, 1–4.

Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making, 23*, 1–14.

Schulze, C., & Newell, B. R. (2015). More heads choose better than one: Group decision making can eliminate probability matching. *Psychonomic Bulletin & Review*. Advance online publication. doi:10.3758/s13423-015-0949-6

Schulze, C. & Newell, B. R. (2016). Taking the easy way out? Increasing implementation effort reduces probability maximizing under cognitive load. *Memory & Cognition*. Advance online publication. doi:10.3758/s13421-016-0595-x

Schulze, C., van Ravenzwaaij, D., & Newell, B. R. (2015). Of matchers and maximizers: How competition shapes choice under risk and uncertainty. *Cognitive Psychology, 78,* 78–98.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making, 15*, 233–250.

Unturbe, J., & Corominas, J. (2007). Probability matching involves rule-generating ability: A neuropsychological mechanism dealing with probabilities. *Neuropsychology*, *21*, 621–630.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14,* 101–118.

Wolford, G., Newman, S. E., Miller, M. B., & Wig, G. S. (2004). Searching for patterns in random sequences. *Canadian Journal of Experimental Psychology*, *58*, 221–228.

APPENDIX

Instructions for classroom demonstration

The current game involves rolling a fair 10-sided die 50 times. On this particular 10-sided die:

7 sides are black,

3 sides are white.

FIGURE 3.2

Figure 3.2 shows all 10 faces of the die that you will be playing with. Your job is to predict which side will show up once the die is rolled. You will be asked to make all predictions before all rolls of the die on a piece of paper. Once you have made your predictions, the experimenter will roll the die 50 times and you will be able to observe the outcome of each roll. For each correct prediction, you will earn one point and your goal is to earn as many points as possible. Note that the die is not biased or loaded in any way; it is a *fair* die. This means that each of the 10 sides has an equal chance of turning up when the die is rolled. You have 10 minutes to complete this task.

4

FRAMING

Anton Kühberger

Is telling half a lie morally worse than telling half the truth? Does a glass that is described as half full contain the same amount of liquid as a glass described as half empty? Can I like ground beef containing 80% lean, but at the same time dislike ground beef containing 20% fat?

These are examples of using different compositions of words for communicating things differently. The important point is that the difference lies in communication rather than in substance: It is at the level of the wording while the actual state of affairs is identical. That is, the same reality is differently described in language. Research that deals with the causes and consequences of using different linguistic descriptions of the same state of affairs comes under the heading of "framing". It presupposes that it is possible to highlight, in a message, different aspects of a state of affairs, without changing the message's content. Obviously, it is a hard problem to define what can be changed in a message without changing its content (for a thorough treatment, see Sher & McKenzie, 2011). Framing research shows that seemingly innocent wording changes can have important consequences.

That some states of affairs can be described in various different ways is in itself not an interesting insight. What makes the issue interesting is that there is a *systematic framing effect*. This effect refers to a class of well-established empirical findings showing that the wording of an utterance has predictable effects on people's responses even in equivalent situations or outcomes. In the loose definition of *framing*, which political scientists and communication scholars tend to use, a framing effect occurs when, due to the specific way of describing an issue or event, identical situations are construed differently, leading to different opinions. In the stricter view used in psychology and economics, a framing effect occurs when logically equivalent statements of a problem lead decision makers to choose different options or to evaluate options differently.

The framing effect is a reliable empirical finding, reported in hundreds of papers. Framing research gained its momentum from the interpretation that it shows a violation of a basic normative principle, sometimes called the invariance (or extensionality) principle: "Different representations of the same choice problem should yield the same preference. That is, the preference between options should be independent of their description" (Tversky & Kahneman, 1986, p. 253). This principle captures the essence of consequentionalism: Variations of form should not affect judgment and choice as long as they do not affect actual outcomes.

Varieties of framing effects

Framing effects come in many varieties, which can be distinguished along the following dimensions:

Frames in communication versus frames in thought (Druckman, 2001)
A frame in communication is simply the way a situation is described by a speaker in a communicative act. The frame in communication is manipulated by the wording, leading to a frame in thought (i.e., a specific way of looking at things), if the framing is effective.

Emphasis frames versus equivalency frames (Chong & Druckman, 2007a; Druckman, 2004)
Emphasis frames (also called issue frames) highlight a subset of potentially relevant considerations, and speakers thus can direct hearers' focus on this subset. For example, the issue of encryption of e-mails can be emphasized as a matter of public safety or as a matter of individual privacy. Emphasis framing does not involve logically equivalent ways of making the same statement; for another example, see Text box 4.1. Equivalency framing involves casting the same information in different descriptions, typically positive or negative. The Asian disease problem (see Text box 4.1) is the classic example (Tversky & Kahneman, 1981). Emphasis framing is a hot topic in sociology and political science using mainly content analysis as a method. Equivalency framing is a hot topic in psychology and economy, using mainly the experimental method. A study found that among about 380 papers published on framing between 1997 and 2007, about 60% were on emphasis framing, while only about 20% were on equivalence framing using the experimental method (Borah, 2011).

Attribute framing, goal framing, and risky-choice framing

Levin, Schneider, and Gaeth (1998) proposed a typology that probably is the best known typology of equivalency-framing tasks, where the critical information is cast in either a positive or a negative light. They distinguished among attribute framing, goal framing, and risky-choice framing.

TEXT BOX 4.1 EXAMPLES OF FRAMING TASKS

1 Emphasis framing (Druckman, 2001, Exp. 1): Giving assistance to the poor by the government in a humanitarian as opposed to a government expenditures frame.

a Humanitarian frame. *In the next few weeks, the US Congress will likely accept one of two proposals that will alter the amount of federal assistance to the poor. One proposal is to increase assistance while the other is to decrease assistance. An increase in assistance to the poor would ensure help for many people who need it. A decrease in assistance would prevent people from receiving basic support. Do you think Congress should increase or decrease assistance to the poor?*

☐ *Increase* ☐ *Decrease*

b Government expenditures frame. *In the next few weeks, the US Congress will likely accept one of two proposals that will alter the amount of federal assistance to the poor. One proposal is to increase assistance while the other is to decrease assistance. An increase in assistance to the poor would lead to an increase in government spending. A decrease in assistance would allow the government to cut excessive expenditures. Do you think Congress should increase or decrease assistance to the poor?*

☐ *Increase* ☐ *Decrease*

2 Attribute framing (Levin, 1987): Evaluating ground beef described by percentage lean or fat.

a Lean frame. *In this brief survey we want to know what associations or thoughts come to mind when making consumer purchases. We will present you with pairs of possible associates. In each pair we want you to indicate by filling in one of the squares which item in the pair you are most apt to associate with a purchase of 75% lean ground beef and the extent to which you associate the purchase with that item rather than the other item in the pair.*

☐ ☐ ☐ ☐ ☐ ☐ ☐

good-tasting *bad-tasting*

b Fat frame. *In this brief survey we want to know what associations or thoughts come to mind when making consumer purchases. We will present you with pairs of possible associates. In each pair we want you to indicate by filling in one of the squares which item in the pair you are most apt to associate with a purchase of 25% fat ground beef and the*

extent to which you associate the purchase with that item rather than the other item in the pair.

☐ ☐ ☐ ☐ ☐ ☐ ☐

good-tasting *bad-tasting*

3 Goal framing (Meyerowitz & Chaiken, 1987): Emphasizing the positive consequences of doing or the negative consequences of not doing breast self-examination (BSE).

a Positive goal. *By doing BSE now, you can learn what your normal, healthy breasts feel like so that you will be better prepared to notice any small, abnormal changes that might occur as you get older. Research shows that women who do BSE have an increased chance of finding a tumor in the early, more treatable stage of the disease. You can gain several potential health benefits by spending only 5 minutes each month doing BSE. Take advantage of this opportunity.*

b Negative goal. *By not doing BSE now, you will not learn what your normal, healthy breasts feel like so that you will be ill prepared to notice any small, abnormal changes that might occur as you get older. Research shows that women who do not do BSE have a decreased chance of finding a tumor in the early, more treatable stage of the disease. You can lose several potential health benefits by failing to spend only 5 minutes each month doing BSE. Don't fail to take advantage of this opportunity.*

4 Risky-choice framing (Tversky & Kahneman, 1981): Describing the outcome of a health-related risky choice in gain or loss terms.

a Positive frame. *Imagine that the US is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows*:

If Program A is adopted, 200 people will be saved.
If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.
Which of the two programs would you favor?

b Negative frame. *Imagine that the US is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows*:

If Program C is adopted, 400 people will die.
If Program D is adopted, there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.
Which of the two programs would you favor?

Attribute framing

In attribute framing, probably the simplest way of doing equivalency framing, a single attribute is framed. As the dependent variable, evaluative ratings are required of favorability, or acceptance. For instance, a frequently used attribute-framing task describes the expected outcome of a program as % success, or % failure, and asks for acceptance of the program. Another example is given in Text box 4.1. Levin and Gaeth (1988) suggested an explanation of attribute-framing effects in terms of memory: The labeling of an attribute leads to the encoding of the information in a way that evokes the respective associations (e.g., % success evokes positive associations, % failure evokes negative associations). This leads to a valence-consistent shift in the evaluation of the entire issue.

Goal framing

In goal framing, messages are framed to either stress the positive (negative) consequences of performing (failing to perform) an act. Attention thus is focused on attaining the positive goal or preventing the negative outcome. A well-known example of a goal-framing task has been used by Meyerowitz and Chaiken (1987; see Text box 4.1). Notice the difference from attribute framing: Attribute framing would portray a situation as positive or negative (e.g., by informing about either the success rate or the failure rate), whereas in goal framing an act is described: The positive goal frame describes the good consequences of performing the act, whereas the negative goal frame describes the same good consequences in terms of their potency of avoiding losses associated with failing to perform the act.

Explanations of goal-framing effects resolve around two issues: emphasizing the construal of a behaviour as risky or uncertain, as opposed to safe and certain (Rothman & Salovey, 1997); and emphasizing people's dispositional sensitivity to favourable or unfavourable outcomes (Mann, Sherman, & Updegraff, 2004). Inspired by prospect theory (see below), loss-framed messages were thought to be more persuasive when the behavior involves a risk of an unpleasant outcome, while gain-framed messages should be more persuasive for adopting a behavior that seems relatively safe and free of an unpleasant outcome. Thus, in the context of health, messages inviting illness-prevention behaviors (safe and free of unpleasant outcomes) will be more persuasive when framed as gains, while messages inviting illness-detection behaviors (risky with unpleasant outcomes) will be more persuasive when framed as losses.

Rothman, Wlaschin, Bartels, Latimer, and Salovey (2008) have proposed a somewhat different account for goal-framing effects: They proposed that the self-regulatory orientations introduced by Higgins (1998) – prevention focus (being concerned with safety, security, the fulfilment of obligations, and the absence of unfavourable outcomes), and promotion focus (being concerned with accomplishments and ideals, and the presence of favourable outcomes) – mediate the persuasiveness of framed messages. Specifically, Rothman et al. (2008) argued that

health-promotion behaviors (e.g., using sunscreen) induce a promotion-focused mindset, thus rendering the presence or absence of favourable outcomes especially salient and persuasive. In contrast, illness-detection behaviors (e.g., mammography) induce a prevention-focused mindset, rendering the presence or absence of unfavourable outcomes especially salient and persuasive.

Risky-choice framing

The prototypical example of risky-choice framing comes from Tversky and Kahneman's (1981) Asian disease problem (see Text box 4.1), where it was found that choices between a risky and a sure option of equal expected value depended on whether the options were described in positive terms (i.e., lives saved) or in negative terms (i.e., lives lost). Tversky and Kahneman found that a majority of participants under positive framing preferred the sure option (i.e., they preferred *saving 200 people for sure* over *saving 600 people with probability 1/3 and saving none with probability 2/3*), whereas a majority of participants under negative framing preferred the risky option (i.e., they preferred *losing 600 people with probability 2/3 and losing none with probability 1/3* over *losing 400 people*). Tversky and Kahneman (1981) explained this choice pattern in terms of their prospect theory. Prospect theory (Kahneman & Tversky, 1979) holds that during evaluation processes, outcomes are coded relative to a reference point, and probabilities are translated into decision weights. Prospect theory's essential characteristic is the functional form of subjective valuation, which depends on the reference point: Above the reference point, in the domain of gains, the value function is concave, and below the reference point, in the domain of losses, the value function is convex. In addition, the slope of the value function is steeper for losses than for gains, giving rise to the phenomenon of loss aversion (i.e., the reaction to losses is stronger than the reaction to gains).

The prospect-theory explanation of risk attitude in the Asian disease problem (and for risky-choice framing in general) goes like this: In the positive frame, a reference point is adopted such that the disease is allowed to take its toll of 600 lives. Relative to this reference point, the outcomes of the two options are gains, and – by virtue of the value function v – the sure option is valued by $v(+200)$ and the risky option by $\frac{1}{3}v(+600) + \frac{2}{3}v(0)$. Because the value of 0 is 0 by agreement, the choice is between $v(+200)$ (sure gain) and $\frac{1}{3}v(+600)$ (risky gain). Due to the concavity for gains of the value function $v(+200) > \frac{1}{3}v(+600)$, people prefer the sure gain. In the negative frame, the reference point is zero people dying (rather than 600 people dying). Relative to this reference point, the outcomes are perceived as losses, with the sure loss valued by $v(-400)$, and the risky loss by $\frac{2}{3}v(-600) + \frac{1}{3}v(0)$. Again the choice is between $v(-400)$ and $\frac{2}{3}v(-600)$. Since the value function for losses is convex, $v(-400) < \frac{2}{3}v(-600)$. The risky loss thus is less aversive and is therefore preferred over the sure loss. We end up with the risky-choice framing effect: risk aversion in the domain of gains, and risk seeking in the domain of losses.

Valuation of outcomes as described by the value function of prospect theory is the dominant explanation of risky-choice framing effects. However, there are a variety of related explanations that also take the weighting of probabilities into account. For instance, cumulative prospect theory (Tversky & Kahneman, 1992) predicts a fourfold pattern: risk aversion for gains and risk seeking for losses of high probability, but risk seeking for gains and risk aversion for losses of low probability, because risk attitude is determined jointly by outcome valuation and probability weighting. Finally, in venture theory (Hogarth & Einhorn, 1990) framing effects are predicted solely from the peculiarities of probability weighting.

The size of the framing effect

Researchers are not only interested in whether or not they can reject the null hypothesis of no framing effect in a significance test but also in the size of the effect. Effect sizes are often reported in meta-analyses. There are general meta-analyses of framing effects (Kühberger, 1998; Levin et al., 1998; Pinon & Gambara, 2005), and meta-analyses specifically on message-framing effects (Gallagher & Updegraff, 2012; O'Keefe & Jensen, 2008, 2009), and on risky-choice framing effects (Kühberger, Schulte-Mecklenbeck, & Perner, 1999).

Before we discuss the framing-effect size, it is important to point to a problem with the measurement of equivalency-framing effects. Naïvely, if two options are equivalent, there should not be a definite preference for either option, and therefore a framing effect could be operationally defined as a discrepancy from the 50:50 even-split. However, equivalency in expected value does not entail equivalency in subjectively expected utility as there can be many reasons to prefer an option over another one, different from expected value. An "unframed" condition, or a condition framed in both ways (termed *competitive* in research by Druckman; e.g., Chong & Druckman, 2007b) might provide a plausible benchmark for comparison, however. For example, Druckman (2001), using the Asian disease task, presented the sure option as "200 people will be saved and 400 people will die," and compared the positive and negative framing condition to this competitive condition. His findings suggest that exposure to competitive frames leads individuals to convert to their previous beliefs/opinions, thereby annulling the effects frames typically have. However, research on competitive framing is rare (about 3% according to Borah, 2011), and may not be feasible in many cases, thus an operational definition of framing as a difference between framing conditions makes more sense.

Two approaches have been taken to evaluate the existence/strength of framing effects between conditions: the unidirectional approach testing a choice shift, and the bidirectional approach testing a choice reversal (Druckman, 2001; Kühberger, 1998; Wang, 1996). Consider the following case: Imagine you run a risky-choice framing experiment with 30 participants in either framing condition. With gains, 10 participants (33%) choose the sure option, whereas with loss, 3 (10%) do so. The

statistical test yields $\chi^2(1) = 4.81$, $p = .029$, showing significantly more risk seeking with losses than with gains. This is a unidirectional framing effect: It shows increased risk seeking in the negative framing condition compared to the positive condition. However, there is no bidirectional effect in these data as there is no risk aversion for gains. A bidirectional framing effect would require just this: that there is both risk aversion for gains (i.e., more than 50% choices of the sure option), as well as risk-seeking for losses (i.e., more than 50% choices of the risky option). Specifically, in our example the bidirectional test (discrepancy from the 50:50 expectation in either condition) results in no framing effect for gains (10 of 30 is not significantly different from 50% ($\chi^2(1) = 3.33$, $p = .068$; there is even a tendency in the wrong direction), but a highly significant effect for losses ($\chi^2(1) = 19.2$, $p < .0001$). Thus the very same finding can be interpreted either as (i) a significant framing effect as predicted by prospect theory (the unidirectional interpretation), (ii) no framing effect for gains but a framing effect for losses, providing partial support of prospect theory, or (iii) no framing effect, contrary to the predictions of prospect theory (the bidirectional interpretation). In the literature, the test for evaluating direction and size of framing effects varies. Sometimes unidirectional and sometimes bidirectional framing effects are reported.

Some meta-analyses report effect sizes for framing effects. The first comprehensive meta-analysis (Kühberger, 1998) summarized 136 studies with 230 single effects sizes based on a sample of about 30,000 participants. Mean (unidirectional) Cohen's d was $d = .33$, weighted by the reciprocal of variance, which takes sample size into account $d = .31$. Pinon and Gambara (2005) essentially repeated Kühberger's (1998) meta-analysis with studies that were published between 1997 and 2003 (i.e., in the years following the analysis by Kühberger). They used the typology of Levin et al. (1998) and reported a mean unweighted (weighted) $d = .41$ (.44) for risky-choice framing, $d = .39$ (.26) for attribute framing, and $d = .54$ (.44) for goal framing in a sample of 151 effect sizes. These effect sizes seem to be higher, but actually they are similar. Kühberger (1998) did not use the Levin et al. typology, but made a distinction between what he called *risk manipulation by reference point* versus *risk manipulation by outcome salience*. The latter, risk manipulation by outcome salience, is the type of tasks that Levin et al. call attribute framing. In Kühberger's analysis, outcome-salience studies have a negative effect size of $d = -.11$, indicating that the results go against the prospect-theory prediction. Indeed, prospect theory is not applicable for attribute-framing tasks. For reference-point framing tasks (i.e., risky-choice and goal framing combined) Kühberger found $d = .50$, quite similar to what Pinon and Gambara reported for the two framing conditions (i.e., $d = .41$ and $d = .54$, respectively).

For the subgroup of risky-choice framing tasks, a meta-analysis on more than 40 such experiments reported the bidirectional framing affect (discrepancy from 50:50 proportion; Kühberger et al., 1999). Risk-aversion effect size with gains was 13% (63% choices of the sure gain), while risk-seeking effect size with losses was 9% (59% choices of the risky loss). For the subgroup of message-framing tasks, O'Keefe

and Jensen (2008) reported the effect sizes of framing on message engagement (i.e., thinking about the message, memory of the message), and on message persuasiveness (O'Keefe & Jensen, 2009). Based on 42 cases, they reported that gain-framed messages produced significantly greater message engagement than did loss-framed messages (d = .12), while, based on 53 cases, loss-framed appeals were significantly more persuasive than gain-framed appeals (d = .08).

Gallagher and Updegraff (2012) did a meta-analysis on health-message framing effects distinguishing effects of positive/negative framing on attitudes, intentions, and behavior. Of 94 published studies, 189 effect-sizes were analyzed. The results were similar to O'Keefe and Jensen (2008, 2009), showing very small effects – indeed non-significant effects – of framing on the persuasiveness of health messages when the persuasiveness was assessed as either attitudes toward behavior or as intention toward behavior. The only significant finding was that, for illness-prevention behaviors, gain-framed messages were more effective in influencing actual behavior (d = .17); no such effect was found for illness-detection behaviors.

Taken together, there is good evidence for a framing effect of medium size in attribute and risky-choice framing tasks of about d = .45 (r = .22). To illustrate its practical importance, a binomial effect-size display (Rosenthal & Rubin, 1982) may be helpful. Assuming an experiment with 200 participants (100 participants in the positive and 100 in the negative framing condition) and equal attractiveness of unframed options (i.e., 50:50; we do not imply that this is actually true), positive framing would increase the percentage of participants choosing risk-aversive to 61% (.50 + r/2), and negative framing would decrease the percentage of participants choosing risk-aversive to 39% (.50 – r/2). In contrast to attribute framing and risky-choice framing, the evidence for goal-framing effects is weak, even after a considerable amount of empirical work.

TEXT BOX 4.2 RUNNING A FRAMING EXPERIMENT IN CLASS

Method

Divide participants into two groups roughly equal in size; these are the two different framing conditions, manipulated between subjects. Distribute booklets containing a framed Asian disease task and instruct individual participants to decide between the two programs described in the booklet. Here is the task framed positively:

Imagine that Europe is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease

have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:

If Program A is adopted, 200 people will be saved.
If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.
Which program would you favour?

☐ Program A ☐ Program B

Here is the task framed negatively:

Imagine that Europe is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:

If Program C is adopted, 400 people will die.
If Program D is adopted, there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.
Which program would you favour?

☐ Program C ☐ Program D

Design

Framing experiments vary widely in number of participants, and therefore in power. A minimum of 50 participants in either framing condition is recommended to avoid a disappointing outcome in terms of a failure to replicate the effect.

Analysis

Participants are required to choose one of the two programs, usually termed "sure" (Programs A and C) and "risky" (Programs B and D). The results thus can be reported in a 2 (framing: positive vs. negative) × 2 (choice: sure vs. risky option) frequency table. Alternatively, they can be depicted by a diagram showing the relative preference of sure (more rarely, risky) choices.

Two basic analyses are possible: unidirectional and bidirectional. The unidirectional test compares the two framing conditions directly in a 2 × 2 table. Do a χ^2 ($df = 1$)-test. The bidirectional test is done for single framing conditions and tests whether the choice proportions differ from 50% in either condition. Again, run χ^2 ($df = 1$)-tests.

Results

Chances are good to find a unidirectional effect, showing a difference between framing conditions (estimated power with 50 participants in each condition is 0.60). Chances to find both bidirectional effects are less good, but a bit higher in the positive than in the negative framing condition.

Instructive variations of the classroom demonstration

Order of programs

The sure programs are listed before the risky programs. Solid experimental methodology requires randomization of the order of programs. This is rarely reported in experiments and it must be feared that randomization is rarely done. There is evidence that the order of program matters (at least) for risky-choice framing (e.g., Kühberger & Gradl, 2013, Exp. 2). Discussion of the effects of presentation order may increase students' understanding of the role of the reference point, which presumably is stronger in the sure/risky order than in the reversed order (see Kühberger & Gradl, 2013, for a more thorough discussion).

Response mode

Usually, forced choice is used in framing experiments. Thus, participants have to choose one option and cannot remain undecided. Making a "don't know" option available is sensible, given that the difference in attractiveness of options may be limited (note that the options' expected values are identical). Empirically, allowing for indifference reduces the effect. Discussion of the effect of allowing indifference helps students realize that the task design is apt to artificially increase the effect (for another explanation of the moderating role of indifference, see Huang, Su, & Chang, 2015).

Design: Between vs. within subjects

Manipulating framing within participants is possible and does not result in a disappearance of the effect, contrary to what students tend to assume. This counterintuitive finding can be used to show that there is indeed substance to the framing effect, since (at least some people) see a real difference between the two framing conditions (even in within-subjects presentation) that is stronger than the desire to appear consistent in one's own choices.

Description of options

The sure option only states that 200 people will be saved; it says nothing about the remaining 400 people. Adding the implicit complement (i.e., that 400 people will not be saved) spoils the effect. The same applies to the negative framing condition: Adding "400 people will not die" to "200 people will die" in the sure option will make the effect disappear. These findings can be used for discussing the role of implicit complements and of the ambiguity of language (see also Tombu & Mandel, 2015).

Size of numbers

The magical number 600 in conjunction with the probability levels of 1/3 and of 2/3 is used frequently in many risky-choice framing experiments. There is no deeper theoretical reason for these levels. However, they result in strong effects. Be careful: The use of small numbers (e.g., 6 people in danger, 2 of them saved) with the Asian disease task leads to consistent risk-seeking in both framing conditions (Wang, 1996), as does the use of small payoffs in framed gambles (Kühberger, Schulte-Mecklenbeck, & Perner, 2002).

Domain

The Asian disease problem pertains to the health domain. The models explaining the framing effect are domain-dependent to a degree. For instance, ideas about leakage and the pragmatic use of language and quantifiers (e.g., that people understand *at least* 200 will be saved, rather than *exactly* 200 will be saved; Mandel, 2014) are plausible in health, or disaster, situations, but are implausible in gambling situations. Nobody would seriously argue that, in a gamble, an option described as "with Option A you win €200", would be interpreted as you win *at least* € 200. Rather, in the gambling domain, numbers are treated as exact. Nevertheless, framing effects with monetary outcomes are robust and not weaker than with human lives (Kühberger et al., 1999).

Moderators and mediators

A variety of moderators and mediators for framing effects have been proposed (e.g., affect, expertise, deliberation, effort, gender, involvement, need for cognition, numeracy). It is beyond the scope of this chapter to give an overview on the pertinent findings, but the overall result is clear: There is no consistency in the findings.

Explaining the framing effect

The enormous amount of empirical work inspired various theoretical ideas to explain the framing effect. Some ideas identify the source of the effect in formal

characteristics, others in cognitive or motivational processes, still others in considerations following from the pragmatics of language and communication.

Formal characteristics

(Cumulative) prospect theory (see above) is the most influential formal model of framing. Although it explains a variety of framing results, researchers began to doubt the assumption that people are passive information processors, feeding in information as it is presented to them. Rather, people actively select information, interpret incoming information on the basis of plausible background knowledge, and process the data on different levels. An example for the role of selective attention is our own research (Kühberger, 1995; Kühberger & Gradl, 2013; Kühberger & Tanner, 2010; Schulte-Mecklenbeck & Kühberger, 2014). We argued that, in the classic risky-choice framing task, the sure option is incompletely specified: The description states that 200 people are saved, but nothing is said about the remaining 400 lives. However, people may understand pragmatically that *about* 200 will be saved, or that 200 people *or more* will be saved. Our work shows that, adding implicit, or deleting explicit, complements of the information changes framing effects: The addition of the hidden sure complement (e.g., 200 people will be saved and 400 people will not be saved), as well as the subtraction of a risky complement (e.g., not saying that with 2/3 probability nobody will be saved), makes the framing effect disappear. Arguing along similar lines, Mandel (2014) recently showed that most people spontaneously interpret the quantifiers in standard framing tasks as lower bounded (i.e., *at least* 200 will be saved) and that these interpretations of the numbers provided strongly moderate the framing effect. Taken together, the (cumulative) prospect-theory idea that framing effects follow from the specifics of outcome valuation (and probability weighting) can account for only a limited number of findings. Framing effects have more than just this single source.

Cognitive accounts

A paradigmatic model taking the peculiarities of information processing seriously is fuzzy-trace theory (Reyna & Brainerd, 1991). Fuzzy-trace theory holds that framing effects are the result of information processing strategies that operate on a superficial, simplified level. According to this approach, people have a preference for fuzzy processing at the lowest possible level. Thus, people reduce quantitative information to its qualitative gist. That is, they turn *200 people will be saved* into *some people will be saved*, and the risky option thus becomes *some people will be saved or no one will be saved*. *Some people will be saved* is common to both the sure and the risky option, and can therefore be ignored. The choice boils down on the simple contrast between *some people will be saved* and *no one will be saved*. Because saving some people is clearly better than saving none, the sure gain is selected. In the negative frame, the picture reverses: *No one will die* is the gist of the risky option, rendering this option more attractive (no one will die is better than some people will die). The fuzzy-trace

explanation can be applied in a straightforward way to risky-choice framing tasks; it is, however, not clear how it could be applied to attribute or to goal framing. Note also the contradiction to Mandel's (2014) idea of a lower-bound interpretation of the numbers.

Motivational accounts

Schneider (1992) argued that two motives, security (the motive to avoid failure) and potential (the motive to succeed), jointly determine risk attitude. In addition, aspiration level reflects the individual's hopes and needs and is dependent on the situation. In framing tasks, security seeking leads to the avoidance of worst outcomes, and potential seeking invites approaching the best outcomes. In the domain of gains, security seeking and the existence of an option exceeding the aspiration level (the sure gain) fuels the choice of the sure gain. The situation is different for losses, where security seeking favors the choice of the sure loss. This option, however, is below the aspiration level, rendering it unacceptable. Consequently, the risky loss is chosen.

Regulatory-focus theory explains risk attitude by similar motivational mechanisms of self-regulation. According to Higgins (1997, 2000), humans have two basic systems of self-regulation, the promotion focus and the prevention focus. Activation of these different foci affects the way in which goals are construed (i.e., as ideals or oughts), how strategies of pursuing these goals are selected (i.e., through approach or avoidance), and the resulting emotions after success and failure (cheerfulness and dejection versus quiescence and agitation). Initial research simply assumed that the promotion system is associated with using a maximax strategy (i.e., maximizing the maximum gain), leading to a bias toward accepting risky options, while the prevention system is associated with using a minimax strategy (i.e., minimizing the maximum loss), leading to a bias toward conservative options.

Recent research shows that the prediction of risk attitude from regulatory focus is more complex (Kühberger & Wiener, 2012; Scholer, Zou, Fujita, Stroessner, & Higgins, 2010). Take three outcomes: a loss (simply quantified as -1), a reference point (0), and a gain ($+1$). Promotion-focused individuals are sensitive to gains, but are not particularly sensitive to non-gains, a situation that we can describe as: $-1 \approx 0 < +1$. In contrast, prevention-focused individuals are sensitive to losses, but relatively insensitive to non-losses: $-1 < 0 \approx +1$. This simple model implies that, given that in prevention focus people are eager not to realize the loss, they will choose options that avoid the loss outcome. Thus, if an option falls short of some reference point, it cannot be chosen under a prevention focus. Because, in risky-choice framing tasks, it is the sure option which frequently falls short of the reference point, people in prevention focus will prefer the risky option (Scholer et al., 2010). Note that in the domain of gains, no specific preference exists, since promotion-focused individuals are motivated to make progress away from the current state, and the reference point has no special meaning as a goal.

This categorical coding of outcomes might look unrealistically simple. Note, however, that it is consistent with the representational ideas of fuzzy-trace theory

(Reyna & Brainerd, 1991, 2011): that decision making operates on simplified rather than on exact numerical information. Indeed, it is just a different form of gist representation.

Regulatory-focus theory is applicable to framing by a still different reasoning: Congruency between the regulatory focus and the perception of situations increases action motivation (Higgins, 2000). In a framing task, the focus can be on attaining gains, on avoiding non-gains, on avoiding losses, or on attaining non-losses. Lee and Aaker (2004) combined message framing and regulatory focus. They showed that increased persuasion occurred when the end state defined by desirability (i.e., gain versus loss) was compatible with regulatory focus (i.e., promotion versus prevention): A gain frame was more effective when it highlighted a promotion possibility, and a loss frame was more effective when it highlighted a prevention concern. Congruency of regulatory focus with salient outcomes is termed regulatory fit, and regulatory fit increases risk taking. Thus, for risky-choice framing tasks, risk seeking for losses under a prevention focus, and risk seeking for gains under a promotion focus follow (see also Kühberger & Wiener, 2012).

The pragmatics of language

Sher and McKenzie (2006, 2011) provided an instructive example of the role of the pragmatics of language for the framing effect. Their information-leakage account proposes that in order to understand the consequences of framing one has to take into account why a communicator chooses to express an opinion in one frame, or another. A speaker's choice of frame can convey (or "leak") choice-relevant information, for instance, about the quantity to be expected. Accordingly, speakers are more likely to describe an object in terms of some property when this property is above some expected reference point rather than below it: If I say that my cup is half full, I may signal that I expected it to be empty (rather than full) in the given circumstances. Similarly, if I say that 200 people were saved, this informs about my expectation that nobody would be saved otherwise.

A similar account, locating framing effects in the pragmatics of language, is related to the use of positive and negative natural language quantifiers (e.g., *many, a few, few*). Some quantifiers tend to focus attention on what Moxey and Sanford (2000) call the "reference set", while other focus attention on the "complement set". For instance, in the expression *a few passengers were killed in the crash, which is a terrible thing*, the terrible thing is that some passengers died (the quantifier relates to the set of passengers for whom the predicate is true). In contrast, in *few passengers were killed in the crash, which is a good thing,* the good thing is that some passengers did not die (the quantifier relates to the set of passengers for whom the predicate is false). The difference in framing that results from using positive or negative verbs like *save* or *die* can also stem from the difference in using positive or negative natural language quantifiers, by virtue of their ability to focus attention on different sets of quantities involved.

More generally, taking the communication perspective seriously shows that the human logical vocabulary is not disinterested. Rather, the choice of a particular

expression of a conditional, quantifier, or probability expression can implicitly convey a signal indicating the speaker's attitude for or against an action (Teigen & Nikolaisen, 2009; van Buiten & Keren, 2009).

Taken together, the story about the source of framing effects is quite complicated; it resolves around the notions of selection and salience of information. Some general processes seem to be that (i) framing can focus attention to different – explicit as well as implicit – aspects of the message; (ii) framing can lead to different degrees of elaboration; (iii) framing can lead to different subjective experiences when processing a message; (iv) framing can lead to different cognitive and motivational processes (Rothman & Updegraff, 2011). In all likelihood, framing effects are the consequence of a mixture of these processes, rather than of a single unitary process.

A broader construal of framing tasks

Our discussion of framing increasingly focused on specific tasks and findings. However, the framing phenomenon has a range of increasingly general applications; it has a hierarchy (Sher & McKenzie, 2011). At Level 1, two utterances are truly equivalent. This applies if they supply identical evidence, and description in communication is disinterested. To exemplify for attribute framing, Level 1 information equivalence requires tossing coins in order to select a description of ground beef as 25% fat or as 75% lean. This rarely is the case. Rather, communicators select descriptions on reason (however not intentional) and Level 1 information equivalence therefore is rarely met. Equivalent descriptions of proportions like in our meat example are Level 2 information equivalent, however, as they convey logically equivalent attribute descriptions. Although these descriptions are logically equivalent, they nevertheless can leak information about what a communicator expects, or prefers. At Level 3 lies the risky-choice framing task: These tasks are outcome-equivalent in the small world of economic analysis, where nothing matters beyond outcome and probability. However, even such a small thing as taking outcomes relative (e.g., by assuming reference points) suffices to introducing inequality. At Level 4, we leave the traditional domain of framing and turn to the universe of data summarized in lists, tables, or figures: Any two such summaries are equivalent as they represent the same observations. There are numerous examples of this. To give just a few disperse ones: (a) Presentation of the result of a framing experiment in terms of risk-seeking rather than risk-aversion may leak information about how researchers conceive the experiment. (b) Delay-discounting decisions change when delay information is framed either as a date, or by the number of days (Dshemuchadse, Scherbaum, & Goschke, 2013). (c) Another framing effect of this type follows from the way risk information is realized, namely, by description or by experience (Hertwig & Erev, 2009). Finally, at information Level 5, we find the emphasis-framing tasks used in sociological and political science studies, which highlight different aspects of topics. Needless to say, even more examples of framing of this type are existing. Taken together, "frames may be logically equivalent descriptions (Level 2), formally equivalent gambles (Level 3), observationally equivalent data digests (Level 4), or

substantively equivalent attempts at persuasion (Level 5). But frames equivalent at Levels 2–5 are sometimes information non-equivalent at Level 1" (Sher & McKenzie, 2011, p. 53). The verdict that the irrationality of human judgment and choice finds exquisite expression in framing effects stands on shaky ground: It is warranted only to the degree that Level 1 equivalence can be taken for granted.

Summary

- Framing exists in a communicative context. It involves the selection of some aspects of reality, making them more salient, thereby promoting a particular problem definition, causal interpretation, and/or treatment recommendation.
- Conceptually, framing research has two broad foundations: sociological (emphasis framing) and psychological (equivalency framing).
- Emphasis-framing effects are robust. In equivalency framing, attribute framing, and risky-choice framing, effects are robust and of medium effect size. Goal-framing tasks produce inconsistent findings.
- Framing effects are explained by a variety of models: formal, cognitive, motivational, and pragmatic. Their explanatory power varies and depends on the type of tasks.
- A framing effect does not show a fundamental flaw in human judgment and choice. Rather, framing effects testify to the complexity of communicative processes. Taking these into account shows that framing effects are entirely reasonable.

Further reading

Keren (2011) edited a useful book on framing offering broad and easily readable covering. An overview over the neurobiology of valuation and framing (not discussed here) is given by Louie and DeMartino (2014); a neuroscience analysis can be found in Levin et al. (2015). Kühberger (2002) offers a discussion on the relationship between framing and irrationality. Available meta-analyses are referred to in the text.

References

Borah, P. (2011). Conceptual issues in framing theory: A systematic examination of a decade's literature. *Journal of Communication, 61*, 246–263.

Chong, D., & Druckman, J. N. (2007a). Framing theory. *Annual Review of Political Science, 10,* 103–126.

Chong, D., & Druckman, J. N. (2007b). Framing public opinion in competitive democracies. *American Political Science Review, 10,* 637–655.

Druckman, J. N. (2001). Evaluating framing effects. *Journal of Economic Psychology, 22,* 91–101.

Druckman, J. N. (2004). Political preference formation: Competition, deliberation, and the (ir)relevance of framing effects. *American Political Science Review, 98*, 671–686.

Dshemuchadse, M., Scherbaum, S., & Goschke, T. (2013). How decisions emerge: Action dynamics in intertemporal decision making. *Journal of Experimental Psychology: General, 142,* 93–100.

Gallagher, K. M., & Updegraff, J. A. (2012). Health message framing effects on attitudes, intentions, and behavior: A meta-analytic review. *Annals of Behavioral Medicine, 43,* 101–116.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences, 13,* 517–523.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist, 52,* 1280–1300.

Higgins, E. T. (1998). Promotion and prevention: Regulatory focus as a motivational principle. *Advances in Experimental Social Psychology, 30*, 1–46.

Higgins, E. T. (2000). Making a good decision: Value from fit. *American Psychologist, 55*, 1217–1230.

Hogarth, R. M., & Einhorn, H. J. (1990). Venture theory: A model of decision weights. *Management Science, 36,* 780–803.

Huang, H. H., Su, H. J., & Chang, C. J. (2015). The moderating effect of a no-choice option on risky framing effect. *Asia Pacific Management Review, 20*, 18–23.

Kahneman, D., & Tversky, A. (1979). Prospect-theory: An analysis of decision under risk. *Econometrica, 47,* 263–291.

Keren, G. (Ed.). (2011). *Perspectives on framing.* New York: Psychology Press.

Kühberger, A. (1995). The framing of decisions: A new look at old problems. *Organizational Behavior and Human Decision Processes, 62,* 230–240.

Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes, 75*, 23–55.

Kühberger, A. (2002). The rationality of risky decisions: A changing message. *Theory & Psychology, 12,* 427–452.

Kühberger, A., & Gradl, P. (2013). Choice, rating, and ranking: Framing effects with different response modes. *Journal of Behavioral Decision Making, 26*, 109–117.

Kühberger, A., Schulte-Mecklenbeck, M., & Perner, J. (1999). The effects of framing, reflection, probability, and payoff on risk preference in choice tasks. *Organizational Behavior and Human Decision Processes, 78*, 204–231.

Kühberger, A., Schulte-Mecklenbeck, M., & Perner, J. (2002). Framing decisions: Real and hypothetical. *Organizational Behavior and Human Decision Processes, 89,* 1162–1175.

Kühberger, A., & Tanner, C. (2010). Risky choice framing: Task versions and a comparison of prospect theory and fuzzy-trace theory. *Journal of Behavioral Decision Making*, *23,* 314–329.

Kühberger, A., & Wiener, C. (2012). Explaining risk attitude in framing tasks by regulatory focus: A verbal protocol analysis and a simulation using fuzzy logic. *Decision Analysis, 9*, 359–372.

Lee, A. Y., & Aaker, J. L. (2004). Bringing the frame into focus: The influence of regulatory fit on processing fluency and persuasion. *Journal of Personality and Social Psychology, 86,* 205–218.

Levin, I. P. (1987). Associative effects of information framing. *Bulletin of the Psychonomic Society, 25*, 85–86.

Levin, I. P., & Gaeth, G. J. (1988). How consumers are affected by the framing of attribute information before and after consuming the product. *Journal of Consumer Research, 15,* 374–378.

Levin, I. P., McElroy, T., Gaeth, G. J., Hedgcock, W., Denburg, N. L., & Tranel, D. (2015). Studying decision processes through behavioral and neuroscience analyses of framing effects. In E. A. Wilhelms & V. F. Reyna (Eds.), *Neuroeconomics, judgment, and decision making* (pp. 131–156). New York: Psychology Press.

Levin, I. P., Schneider, S., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes, 76,* 149–188.

Louie, K., & DeMartino, B. (2014). The neurobiology of context-dependent valuation and choice. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics* (2nd. ed., pp. 455–476). New York: Elsevier.

Mandel, D. R. (2014). Do framing effects reveal irrational choice? *Journal of Experimental Psychology: General, 143,* 1185–1198.

Mann, T., Sherman, D., & Updegraff, J. (2004). Dispositional motivations and message framing: A test of the congruency hypothesis in college students. *Health Psychology, 23,* 330–334.

Meyerowitz, B. E., & Chaiken, S. (1987). The effect of message framing on breast self-examination: Attitudes, intentions, and behavior. *Journal of Personality and Social Psychology, 52,* 500–510.

Moxey, L. M., & Sanford, A. J. (2000). Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology, 14,* 237–255.

O'Keefe, D. J., & Jensen, J. D. (2008). Do loss-framed persuasive messages engender greater message processing than do gain-framed messages? A meta-analytic review. *Communication Studies, 59*, 51–67.

O'Keefe, D. J., & Jensen, J. D. (2009). The relative persuasiveness of gain-framed and loss-framed messages for encouraging disease detection behaviors: A meta-analytic review. *Journal of Communication, 59,* 296–316.

Pinon, A., & Gambara, H. (2005). A meta-analytic review of framing effect: Risky, attribute, and goal framing. *Psycothema, 17,* 325–331.

Reyna, V. F., & Brainerd, C. J. (1991). Fuzzy-trace theory and framing effects in choice: Gist extraction, truncation and conversion. *Journal of Behavioral Decision Making, 4*, 249–262.

Reyna, V. F., & Brainerd, C. J. (2011). Dual processes in decision making and developmental neuroscience: A fuzzy-trace model. *Developmental Review, 31,* 180–206.

Rosenthal, R., & Rubin, D. B. (1982). A simple general purpose display of magnitude and experimental effect. *Journal of Educational Psychology, 74,* 166–169.

Rothman, A. J., & Salovey, P. (1997). Shaping perceptions to motivate healthy behavior: The role of message framing. *Psychological Bulletin, 121,* 3–19.

Rothman, A. J., & Updegraff, J. A. (2011). Specifying when and how gain- and loss-framed messages motivate healthy behavior: An integrated approach. In G. Keren (Ed.), *Perspectives on framing* (pp. 257–277). New York: Psychology Press.

Rothman, A. J., Wlaschin, J., Bartels, R. D., Latimer, A., & Salovey, P. (2008). How persons and situations regulate message framing effects: The study of health behavior. In A. Elliot (Ed.), *Handbook of approach and avoidance motivation* (pp. 475–486). Mahaw, NJ: Erlbaum.

Schneider, S. L. (1992). Framing and conflict: Aspiration level contingency, the status quo, and current theories of risky choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1040–1057.

Scholer, A. A., Zou, X., Fujita, K., Stroessner, S. J., & Higgins, E. T. (2010). When risk seeking becomes a motivational necessity. *Journal of Personality and Social Psychology, 99,* 215–231.

Schulte-Mecklenbeck, M., & Kühberger, A. (2014). Out of sight – out of mind? Information acquisition patterns in risky choice framing. *Polish Psychological Bulletin*, *45*, 21–28.

Sher, S., & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition, 101,* 467–494.

Sher, S., & McKenzie, C. R. M. (2011). Levels of information: A framing hierarchy. In G. Keren (Ed.), *Perspectives on framing* (pp. 35–63). New York: Psychology Press.

Teigen, K. H., & Nikolaisen, M. I. (2009). Incorrect estimates and false reports: How framing modifies truth. *Thinking & Reasoning, 15,* 268–293.

Tombu, M., & Mandel, D. R. (2015). When does framing influence preferences, risk perceptions, and risk attitudes? The explicated valence account. *Journal of Behavioral Decision Making, 28,* 464–476.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453–458.

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business, 59,* 251–278.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5,* 297–323.

Van Buiten, M., & Keren, G. (2009). Speaker's choice of frame in binary choice: Effects of recommendation mode and option attractiveness. *Judgment and Decision Making, 4,* 51–63.

Wang, X. T. (1996). Framing-effects: Dynamics and task domains. *Organizational Behavior and Human Decision Processes, 68*, 145–157.

5

CONFIRMATION BIAS – MYSIDE BIAS

Hugo Mercier

The confirmation bias – "the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand" (Nickerson, 1998, p. 175) – has been described by contemporary researchers as "ubiquitous" (Nickerson, 1998), "perhaps the best known and most widely accepted notion of inferential error to come out of the literature on human reasoning" (Evans, 1989, p. 41). Earlier scholars had similar insights. Here is, for instance, Francis Bacon:

> The human understanding when it has once adopted an opinion . . . draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, or else by some distinction sets aside and rejects; in order that by this great and pernicious predetermination the authority of its former conclusions may remain inviolate.
>
> *(Bacon, Novum Organum, Book 1, Aphorism 46, 1620)*

Indeed, the existence of the confirmation bias might seem obvious to anyone who has talked politics with someone from the other side of the spectrum. As the journalist Jon Ronson put it: "ever since I learnt about confirmation bias I've started seeing it everywhere."[1]

As it turns out, this quip raises interesting questions: Why wouldn't the researchers who have claimed to prove the existence of the confirmation bias also be biased? Couldn't their experiments be designed in such ways that they are more likely to make participants look biased? Couldn't these scholars have mistaken rational behavior for proof of a confirmation bias?

In this chapter I will argue that most of the conventional wisdom about the confirmation bias is wrong – starting with its name. As we'll see, there is no such thing as a general tendency to confirm whatever one thinks about, only a tendency

to find arguments that support one's own views – a *myside bias* (see Perkins, 1989). Moreover, this bias is not as prevalent as is often thought: The best-known experiments that purport to demonstrate the wide scope of the confirmation bias are flawed. Finally, I will suggest that the myside bias is not an undesirable bug that we must strive to get rid of, but an adaptive feature that we can use to our advantage.

Demonstrations of myside bias

Hypothesis testing

The main source of evidence for the confirmation bias used to be two unrelated series of experiments on hypothesis testing. These experiments study what people do when they want to know whether one of their ideas is true or not. Both sets of experiments played a large role in shaping beliefs about the confirmation bias. It is thus worth going into some detail about their logic and their results.

In the 1950s, the English psychologist Peter Wason wanted to test whether people were able to test hypotheses the way scientists were supposed to do it, that is, through falsification – a good hypothesis being one that can withstand repeated attempts to show it is false. To this end, he developed a simple task that emulates the scientific process: the rule-discovery task (Wason, 1960; cf. Chapter 8). As the name indicates, the goal of the participants is to discover a rule. The rule is the equivalent of a law of nature, except that in this case it is the experimenter who has determined what the correct rule is (the arrangement of numbers in a triple). The experimenter starts by providing participants with a triple that conforms to the rule: 2, 4, 6 (which gave the task its moniker: the '2, 4, 6'). Participants can then form a hypothesis about the rule – for instance, "series of numbers to which 2 is added at each step." To test this hypothesis, they provide a triple, and the experimenter tells them whether it conforms to the rule or not. This is the equivalent of a scientist conducting an experiment to test her hypothesis. Participants can test as many triples as they like. When they think they know what the rule is, they lay out their hypothesis, and the experimenter tells them whether it is correct or not. If the participants have gotten it wrong, they can test more triples to evaluate another hypothesis until they get it right or give up.

In most cases, participants tested triples that were compatible with the hypothesis they had in mind. For instance, if a participant had made the hypothesis "series of numbers to which 2 is added at each step," she would be more likely to test it with "10, 12, 14" than with, say, "2, 4, 7." The difference between these two tests is that if the experimenter says that "10, 12, 14" conforms to the actual rule, the hypothesis is confirmed, whereas if he says that "2, 4, 7" conforms to the rule, the hypothesis is falsified. Because of this, Wason described the participants as having a tendency to confirm their hypotheses. Moreover, he found that this tendency either slowed down the participants' progression towards the correct rule, or precluded them from reaching it altogether.

It was soon pointed out, however, that Wason's reasoning was faulty (see Poletiek, 1996). In fact, both types of triples – those that conform to the hypothesis and those

that don't – can either confirm the hypothesis or falsify it. If "10, 12, 14" does not conform to the actual rule, then the participant's hypothesis is falsified. And if "2, 4, 7" does not conform, then the hypothesis is (weakly) confirmed. To tell whether a participant really wants to confirm her hypothesis, one would have to know how she expects her test to turn out. All we can infer, when she generates a triple that conforms to the hypothesis she has in mind, is that she engages in what has been called a 'positive test strategy.'

Whatever the participants' strategy is called, there seems to be a problem with it, because it hinders the discovery of the correct answer. As it turns out, positive test strategy is in fact the most rational course of action in the vast majority of the cases (Klayman & Ha, 1987). If it misfires in the case of Wason's rule discovery task, it is because the triple provided by the experimenter is deceptive: It leads participants to think of a very specific hypothesis – such as "series of number to which 2 is added at each step" – when in fact the correct rule is very general – "two ascending numbers." An analogy can explain why this is misleading, and why the positive test strategy is usually efficient.

Imagine that you suspect a mountain to hold a gold deposit. You have an idea of where the gold might be – in the north flank, say – and you can choose where to prospect first. The positive test strategy is to dig in the north flank. Using this strategy, not only do you find the gold if your hypothesis was correct, but you are also likely to quickly find out if your hypothesis is wrong. If the gold is on the south flank (i.e., your hypothesis does not overlap with the truth), you'll find out immediately that your hypothesis is wrong. If the gold is only on a specific ridge of the north flank (i.e., your hypothesis is too broad), you're also likely to find out quickly (unless you get lucky and then never prospect in other places you believe gold to be, which is unlikely). It's only if there happens to be gold not only in the north flank, but also somewhere else (i.e., your hypothesis was not broad enough), that the positive test strategy might mislead you. Such a situation, however, should be rare, since people are more likely to start from a hypothesis that is overly general rather than one that is overly specific.

Thus, Wason's discovery-rule task does not show that participants have a confirmation bias. Instead, participants rely on a rational strategy in an environment designed to trick them. The same conclusion applies to the other standard demonstration of confirmation bias in hypothesis testing.

This second paradigm, developed by Mark Snyder and his colleagues, aimed at understanding how people conduct interviews – more specifically, whether interviewers suffer from a confirmation bias in selecting the questions they ask interviewees. Participants were put in the position of interviewers, and they were given a hypothesis about an interviewee: That she was either introverted or extraverted, depending on the condition the participants were in (Snyder & Swann, 1978). The interviewers had to pick from a list some questions to ask the interviewee. The relevant questions on the list were designed so that they would be likely to confirm one of the hypotheses. For instance, when asked "In what situations do you wish you could be more outgoing?" most people, including most extroverts,

should provide an answer: Only an extreme extrovert would be unable to find a situation in which they do not wish they were more outgoing. As a result, these questions are not diagnostic when it comes to differentiating mild introverts from mild extroverts, since both of these groups are likely to answer them positively. The issue was that the interviewers tended to select questions that were likely to confirm their hypothesis. For instance, interviewers who had been given the hypothesis that the interviewee was introverted were more likely to ask questions such as "In what situations do you wish you could be more outgoing?" They would then receive answers suggesting that there are situations in which the interviewee wishes she were more outgoing, leading the interviewer to infer that the interviewee is indeed an introvert, when in fact she could also have been a mild extrovert.

This experiment suffers from flaws similar to those of Wason's rule discovery task. First, even though the questions might be more likely to confirm the participants' hypothesis, they still have the potential to falsify it. For instance, someone who is extremely outgoing could simply answer "None" to "In what situations do you wish you could be more outgoing?" – with such an answer, the hypothesis that the interviewee is an introvert would be quite falsified. If we do not know what answers participants expect to get, we cannot say that they are bent on confirming their hypothesis. Second, the participants' strategy might be quite rational: The experimenter might have led them to form the hypothesis that the interviewee was either an extreme introvert or an extreme extrovert, in which case the questions they were asking were quite appropriate. An experiment published a few years later showed that in fact participants are quite good at selecting diagnostic questions, and that they do not show much trace of a confirmation bias (Trope & Bassok, 1983).

The experiment Yaacov Trope and Miriam Bassok (1983) conducted was similar to that of Snyder and Swann. They asked participants (interviewers) to select questions so they could tell whether someone (the interviewee) was an introvert or an extrovert. But they made a couple of changes. First, they varied how diagnostic the questions were. Some questions were highly diagnostic ("Are you usually the initiator in forming new relationships?"). These questions could discriminate between mild introverts (who would typically say no) and mild extroverts (who would typically say yes). Other questions were less diagnostic ("After a test, do you compare your answers to those of other students?"). For these questions, only individuals from one extreme of the spectrum (here, extreme introverts) would be likely to answer differently from everyone else, while mild introverts and mild extroverts would give similar answers. The second change introduced by Trope and Bassok was that interviewers believed that the interviewee was either an extreme introvert, an extreme extrovert, a mild introvert, or a mild extrovert.

If the participants were keen on confirming their hypothesis, they could easily do so. They just had to pick questions that had low diagnosticity. This is not what the participants did: In every case they preferred to pick highly diagnostic questions. Moreover, when the participants picked the low diagnosticity questions, it was often appropriately, in order to test an extreme hypothesis. For instance, asking participants who they thought might be extreme introverts "After a test, do you

compare your answers to those of other students?" Such a question, which does not discriminate between mild introverts and mild extroverts, can discriminate extreme introverts, the only ones liable to answer "No." This strongly suggests that the participants in the original Snyder and Swann experiment only picked low diagnosticity questions because they had no alternative high diagnosticity questions and they had formed extreme hypotheses about the interviewees.

Two of the most important paradigms used to support the existence of the confirmation bias turn out to only reflect rational behavior. I now turn to another task – also designed by Peter Wason, and even more famous – which paints a more complex picture.

Wason selection task

The Wason selection task – or four cards selection task – is the most studied task in the psychology of reasoning (Wason, 1966; see Text box 5.1 and Chapter 8). In the standard version of the task, most participants answer either that the A card, or the

TEXT BOX 5.1 DETAILS OF THE CLASSROOM DEMONSTRATION

The experiment has two phases: an individual phase and a group phase.

Individual phase

Distribute to the students sheets with a standard Wason selection task such as the following:

> Here's a rule regarding the four cards below: "If there is a vowel on one side, then there is an even number on the other side."

FIGURE 5.1

> What card(s) need(s) to be turned over to determine whether the rule is true of these four cards? You can select as many cards as you want.

Give them 5 minutes to complete the task, then ask them to make, at the back of the sheet, a list of thoughts about the correct answer. Each thought should be put on a separate line (this should take another 5 minutes).

Group phase

Form groups of students, ideally at random and of 4 or 5 students. Distribute a new sheet with the same problem, one per group, and ask the students to reach a consensus regarding the correct answer. Give them at least 15 minutes to do so.

Gathering the results

First, give and explain the correct answer. Ask how many students gave the correct answer during the individual phase, and how many did so during the group phase (with a show of hands, for instance). Then ask the students to count the total number of thoughts they listed during the individual phase, as well as how many of these thoughts comprise arguments going against their answer, even if they are refuted (for instance, if they have not selected the 4 card: "the 4 card seemed relevant because the rule mentions even numbers, but in fact it doesn't matter"). The students announce their counts turn by turn while you make a running total for both numbers.

Analyzing the results

A binomial test can be used to show that there are fewer thoughts containing arguments that attack the students' answers than expected at chance. This suggests that participants suffered from a myside bias, which might explain why most of them stuck to the wrong answer. A chi-square can compare the number of correct and incorrect responses in the individual and group phases. The improvement between the individual and the group phases shows that in spite of their myside bias, students who had the wrong answer and were exposed to the correct answer were able to change their mind and to adopt a better answer.

FAQ

Could the Wason selection task be too difficult, or be already known by the students?

The experiment is much more likely to succeed if a few students are able to find the correct answer on their own. If there is a danger that this might not happen (e.g., the class is very small), then using a marginally easier logical task might be preferable. Also, many psychology students might already know of the Wason selection task. Several alternative problems might be considered, such as those of the Cognitive Reflection Test or, better, its recent extension by Toplak, West, and Stanovich (2014).

Aren't the statistics wrong?

Comparing the raw results of the individual and group phases is indeed not quite correct. When the N is large enough, other solutions can be implemented, such as considering each group as a single data point, or even comparing the scores of the real groups to that of nominal groups during the individual phase. Similarly, the tests demonstrating the existence of the myside bias should be performed on individual averages. In both cases, however, the effects tends to be very strong, so that all statistical methods yield similar results.

Does failing to give arguments against one's own answer really represent a myside bias?

Not necessarily. In the 'thought listing' subsection of the main text, I explain the limits of this method and give arguments suggesting that in spite of these limits, it is likely to reflect the operation of a myside bias in most cases.

Is the improvement in the group phase necessarily due to people accepting challenging arguments in spite of their myside bias?

Yes. Unfortunately there is no space here to cover this ground, but see, for instance, Section 2.3 of Mercier and Sperber (2011), and Trouche, Sander, and Mercier (2014). For instance, experiments have shown that the improvement is not simply due to the fact that the participants are asked the same question a second time, and thus given more time to think: If they are put in groups straight away, they also perform better than individuals (Moshman & Geil, 1998).

A and the 4 cards, should be turned over. The correct answer, however, is to turn over the A and the 7. The 4 card is not necessary, since the rule doesn't say what should be on the other side of even numbers; by contrast, the 7 is necessary, since it can falsify the rule – if a vowel is revealed on the other side, the rule is false.

At first, it was thought that the preference for the 4 over the 7 reflected a confirmation bias, a failure to pick a card that might falsify the rule. This interpretation, however, was promptly belied by two observations. First, the A card, picked by most participants, can falsify the rule, so that its choice cannot be explained by a conformation bias. Second, it's enough to add a negation to the consequent in the rule (i.e., "If there is a vowel on one side, then there is *not* an even number on the other side") for participants to provide the correct answer – in this case A and 4 (Evans & Lynch, 1973). Does the addition of a negation allow participants to shed their confirmation bias? No. Whether there is a negation or not, participants tackle

the task in exactly the same way. When they read the rule, the participants' attention is directed towards some cards by mechanisms of pragmatic inference (Sperber, Cara, & Girotto, 1995). These mechanisms allow us to understand utterances. They enrich the literal meaning so that it makes sense in the current context (for instance, going from "Can you pass me the salt?" to "Please pass me the salt"). In the present case, the cards that are made most relevant by the rule are simply those that the rule mentions – the vowel (A) and the even number (4). This does not change when a negation is introduced in the consequent. Thus, although participants use the same mechanisms both with the standard and with the negated rule, only in the latter case do they reach the correct answer. In neither case do they reach this initial answer because of a confirmation bias.

Mechanisms of pragmatic inference act rapidly: Participants' attention turns towards the relevant cards quickly after reading the rule (Ball, Lucas, Miles, & Gale, 2003). However, most participants take several minutes to complete the task. What happens in this lapse of time? Usually, not much. Most participants stick to their intuitive answer until the end. Yet they haven't been idle. Think aloud protocols reveal that they have considered many reasons . . . but they overwhelmingly thought of reasons why they should pick the cards they had intuitively selected. If they think about other cards at all, it is only to gather reasons why they are not relevant (Lucas & Ball, 2005). This looks like a confirmation bias: Participants seem bent on confirming their initial intuition. We'll see presently that this is more properly described as a myside bias, but a few observations are already worth making. First, there is no confirmation bias in the mechanisms that guide participants' initial, intuitive reaction. Instead, the bias affects reasoning; it makes them consider only a biased sample of reasons. Second, the bias can affect reasoning even in the absence of emotions or commitment (in the standard Wason selection task, the rule is abstract, and the participants do not publicly commit to their intuitive answer before reasoning about it).

Thought listing

The last demonstration of the confirmation bias we'll discuss here is also the simplest: asking people to list their thoughts. For instance, list your thoughts on a controversial issue – the death penalty, immigration, tax rates. Take your time, list as many thoughts as you'd like. Now go over your list and count how many thoughts mention arguments going against your own position on the issue (assuming you do have a position). Say, if you have listed thoughts on the death penalty, and you think it should be abolished, how many arguments did you list supporting its use? Even arguments that you went on to refute? A good bet is that you listed very few, if any, such arguments (see, e.g., Kuhn, 1991; Perkins, 1989). The ease with which we find arguments in support of our positions and the difficulty of finding arguments against our positions have been interpreted as proof of a confirmation bias (e.g., Edwards & Smith, 1996).

Some readers, however, might have noticed a flaw in this demonstration. Are people who give few arguments against their own position necessarily biased? On

the contrary, it might seem like the sane thing to do, by contrast with mostly providing arguments against the opinions we hold. A rational thinker might be expected to consult the arguments relevant to a given topic and, on this basis, form a position. If our thinker happens to mostly know of arguments supporting one position, then she should list these arguments and endorse the position they support. As a result, her list of arguments would be hard to distinguish from that of a biased thinker who starts with a set position and then looks for arguments in its favor. This is a sound objection, but more sophisticated experiments have shown beyond doubt that something like the confirmation bias affects how people gather their thoughts. But before moving on to these more sophisticated experiments, we should briefly review a simple one that makes an important point.

In the thought listing exercise above, you were asked to list thoughts and, if you are like most people, most of these thoughts could be said to 'confirm' your position – hence the inclusion under the 'confirmation bias' label. What would have happened if you had been asked instead to reflect on a position you disagree with? Would you have also been biased to find 'confirming' arguments, that is, arguments that confirm the position you disagree with? No, instead you would have been biased to find disconfirming arguments, arguments that refute the position you disagree with (Edwards & Smith, 1996). We do not have a general confirmation bias – a bias to confirm anything we think about – but a myside bias – a bias to find arguments that support our position (see Text box 5.2).

TEXT BOX 5.2 CONFIRMATION BIAS AND MYSIDE BIAS

Strictly speaking, a *confirmation bias* is a tendency to confirm whatever one thinks about. This should apply even when people disagree with what they are thinking about, which is implausible and has been proven to be false (Edwards & Smith, 1996). When people consider a thought they disagree with, they are biased towards falsification: they find exceptions, counter-arguments, counter-examples. It is thus wrong to speak of a general confirmation bias. Instead, people have a *myside bias*: a tendency to find arguments that defend their position, whether this entails supporting a position they agree with, or refuting a position they disagree with.

Edwards and Smith (1996) also showed that the myside bias was more forcefully expressed when the participants reflected on an argument they disagreed with. Participants were more motivated to refute such an argument than to support an argument they agreed with. This asymmetry makes sense in a social context. When we interact with people, we are only expected to offer justifications when we disagree with them, not when we agree. The myside bias might thus be helping us find arguments in anticipation of having to justify our position – a suggestion I will explore further below.

The most conclusive evidence for the myside bias was produced thanks to an original experimental paradigm called 'choice blindness.' In choice blindness experiments, participants answer a question, and are then tricked into believing that they have answered something different from their actual answer. In the most relevant choice blindness experiment, participants were asked to give their opinion, on a scale from "completely disagree" to "completely agree," on a series of moral statements such as "If an action might harm the innocent, it is morally reprehensible to perform it" (Hall, Johansson, & Strandberg, 2012). After a participant handed back her questionnaire, the experimenter performed a sleight of hand that inverted the meaning of some of the statements, so that the question above, for instance, would now read "If an action might harm the innocent, it is morally permissible to perform it." The answer scales, however, were not inverted. As a result, if a participant had agreed to the first statement, she now agreed with a statement meaning the exact opposite. Approximately half of the participants did not detect the manipulation. These participants then justified positions they thought they held, even though they had taken an exactly opposite stance a few minutes before. Indeed, the participants were as biased when they were justifying 'fake,' attributed positions as when they were justifying their genuine positions, overwhelmingly producing supporting arguments in both cases.

This is a perfect demonstration of the myside bias. It is impossible to claim that the participants were looking for reasons and then building a position on this basis. Instead, they were looking for reasons to justify whatever position they thought they held. Other experiments – for instance, in the large literature on motivated reasoning (Kunda, 1990) – offer further evidence of a myside bias, but none are quite as persuasive as these choice blindness experiments.

Limits of the myside bias

I have argued that there is no evidence of a general confirmation bias. Instead, the evidence points to a myside bias – a tendency to look for reasons that support one's opinions. This definition suggests two important limits to the myside bias. First, it only affects a process that deals with *reasons* – that is, reasoning. Second, it only affects how people *look* for reasons, not how they evaluate reasons. It's time to flesh out both of these claims.

Reasoning is sometimes understood as a general process of inference, including unconscious inferences. It is useful, however, to restrict the term 'reasoning' to more specific cognitive mechanisms, mechanisms that deal with reasons. The vast majority of the inferences we perform – for instance, going from "can you pass me the salt?" to "please pass me the salt" – are performed without paying attention to the reasons why we perform them – we do not have to go through a chain of reasoning such as 'it is obvious that I can pass the salt, and this information would be of no relevance to my interlocutor, therefore she must mean something else, probably that she wants me to pass her the salt' (Mercier & Sperber, 2011). In this perspective, reasoning is a cognitive mechanism that allows us to find reasons and to evaluate the support relationships between a reason and a conclusion. For instance,

reasoning allows a speaker to find a reason – "you can easily reach it from where you're sitting" – to support a request – "please pass me the salt" – and it allows her interlocutor to tell how much support the reason lends to the request.

Cognitive mechanisms – including reasoning – can have a variety of biases. For instance, it makes sense for food selection mechanisms to be biased towards judging that a substance is toxic, given the relative costs of ingesting a toxic substance (high) compared to passing over an edible one (low). However, there is no evidence of a confirmation or myside bias in cognitive mechanisms besides reasoning. This should not be surprising given that such a bias would be widely maladaptive. An animal eager to confirm mistaken beliefs – that there are no predators around, for instance – would not survive very long.

Even as a specific feature of reasoning, the myside bias is often believed to affect equally how people find reasons and how they evaluate them. In particular, people are supposed to evaluate other people's arguments very critically when they disagree with their conclusion. I mentioned earlier research by Edwards and Smith (1996) showing that participants were particularly keen to find thoughts refuting an argument whose conclusion they disagreed with. As a result, participants rated these arguments as being very poor, which suggests that argument evaluation is indeed biased. The issue with such experiments, however, is that they conflate the way people rate arguments after they have had time to gather counter-arguments, and the way they evaluate arguments as they read (or hear) them.

For instance, imagine a participant who believes that the death penalty should be abolished, and who reads the following argument:

> Sentencing a person to death ensures that he/she will never commit another crime. Therefore, the death penalty should not be abolished.
>
> *(Edwards & Smith, 1996, p. 9)*

Does she reject the argument as soon as she realizes that it challenges her beliefs? Or does she do so because she has been able to find counter-arguments – for instance, that there are more moral ways of ensuring that someone does not commit another crime? This is an important distinction. If there is a strong bias when people immediately evaluate the argument, such that arguments with unpalatable conclusions are likely to be immediately rejected, then argumentation should be largely pointless: People would not change their minds when exposed to challenging arguments. By contrast, if the bias mostly stems from the production of arguments that follows immediate evaluation, then there is hope: The counter-arguments raised at this stage can be addressed, in the course of a discussion, for instance.

Although we cannot rule out that evaluation is intrinsically biased, two pieces of evidence suggest that most of the bias stems from the production of counter-arguments that takes place after people have been confronted with a challenging argument. The first piece of evidence is that there is a relationship between how many counter-arguments people produce and how negatively they rate the argument (e.g., Edwards & Smith, 1996). The second is that when participants are

allowed to discuss issues at length, then they can be convinced to change their minds (see Mercier & Sperber, 2011). For instance, in a series of experiments participants were confronted with arguments defending the good answer to a logical task (the disjunctive reasoning task from the end of Text box 5.1; Trouche et al., 2014). When participants who had given the wrong answer were only exposed to one good argument, fewer than half of them changed their mind. By contrast, when participants were able to discuss the problem at length, so that their counter-arguments could be addressed, all of those exposed to the correct answer changed their mind – even if they had been extremely confident before the discussion.

These results suggest that when we read or hear an argument, we might not be biased in our evaluation of the argument. By contrast, we are biased in looking for arguments after this initial evaluation: If we are not swayed by the argument, we find arguments that counter it. The myside bias could thus affect argument evaluation indirectly only, through the biased production of counter-arguments.

Explaining the myside bias

When told about the myside bias, most people are not surprised. On the contrary, the myside bias seems to explain why they sometimes fail to convince others (not, obviously, why they sometimes fail to be convinced . . .). Some manifestations of the myside bias, however, should be quite puzzling. When people reason on their own, the myside bias often has dire epistemic consequences – piling up reasons that support our preconceived views is not the best way to correct them. Not only does the myside bias stop people from fixing mistaken beliefs – as in the Wason selection task – but it can even make things worse. When they reason on their own, people can become overconfident and strengthen their preexisting beliefs (for references, see Kunda, 1990; Mercier & Sperber, 2011). A similar phenomenon takes place when people reason with people who agree with them: Arguments for the side everyone agrees on pile up without being criticized, and the average opinion can become more extreme (Isenberg, 1986).

In light of its consequences, we are entitled to ask: Why do we have a myside bias? Two broad classes of explanations are usually offered, cognitive and motivational. The former attempts to explain the myside bias through other, more fundamental biases in the operation of cognitive mechanisms. For instance, here is the explanation ventured by Jonathan Evans, which is fairly representative:

> Subjects confirm, not because they want to, but because they cannot think of the way to falsify. The cognitive failure is caused by a form of selective processing which is very fundamental indeed in cognition – a bias to think about positive rather than negative information.
>
> *(Evans, 1989, p. 42)*

The issue with such an explanation is that it only applies to a putative confirmation bias, not to the myside bias. But we do not have a confirmation bias – when

it comes to ideas we disagree with, falsification comes easily. Thus there does not seem to be any hard cognitive constraints that would favor confirmation over falsification. And no cognitive constraint has been offered that would account for the myside bias.

Motivational explanation suggests that people are motivated to hold some beliefs, and that reasoning serves to prop up these beliefs, to insulate them against attacks (Kunda, 1990). But why would we want to hold some beliefs in particular? Beliefs are there to guide our actions, and mistaken beliefs lead to harmful behavior. Why wouldn't we want our beliefs to be the best approximation of reality we can come up with? Holding some beliefs can make us happy, but cognitive mechanisms do not evolve because they make us happy; they evolve because they increase our fitness. In any case, people are often biased to find arguments supporting beliefs that make them miserable – that their partner is cheating on them, that they are about to be fired, etc. It seems that neither cognitive nor motivational explanations can account for all the observed features of the myside bias.

Another way of thinking about the myside bias is in terms of its potential usefulness. What is the myside bias good for? Clearly, it doesn't help us reach sounder beliefs on our own. But when we have to defend our beliefs, to convince someone else of their worth, then a myside bias is desirable. It is easier to convince an audience with arguments that support our position than with arguments that challenge our views. This fits well with the argumentative theory of reasoning, which suggests that argumentation is the main function of reasoning (Mercier & Sperber, 2011). In a nutshell: Humans cooperate on an unprecedented scale among primates. To do so, they rely on equally unprecedented communication skills. Successful communication, however, raises an evolutionary challenge. For communication to be stable, senders must not be able to abuse receivers to the point at which communication stops being beneficial. In humans, receivers defend themselves against harmful communication by filtering messages. They are more likely to accept messages that fit with their prior beliefs, and messages sent by people they trust. Argumentation enables the transmission of messages when trust isn't sufficient, thereby enlarging the scope of what can be successfully communicated. For instance, if I had peremptorily asserted the position held here on the myside bias, few readers would have taken my word for it. By contrast, laying out arguments to support this position is (hopefully) more likely to convince. According to the argumentative theory, argumentation would be the main function of reasoning.

From the perspective of the argumentative theory of reasoning, the myside bias should be specific to reasoning – as it seems to be. Crucially, the myside bias should only apply to how we find arguments. When we evaluate others' arguments – at least when we disagree with their conclusion, the most common case – the main task of reasoning is to recognize good arguments and to make us change our mind accordingly, so that we end up with better beliefs on average. A strong myside bias in evaluation would make argumentation pointless. As argued above, a good case can be made that argumentation evaluation is either unbiased, or at the very least

not so biased as to preclude effective argumentation. Finally, as might be expected of an evolved, adaptive trait, no one seems to be exempt from the myside bias – intelligence, motivation, and open-mindedness do not protect against the myside bias (see, e.g., Stanovich, West, & Toplak, 2013).

If the argumentative theory of reasoning can account for the main features of the myside bias, one could still wonder why it also affects solitary reasoning, or even why people reason on their own altogether. Two lines of explanations can be sketched here. The first is that there are likely social benefits to reasoning on one's own. Solitary reasoning allows discarding beliefs that we can't justify (Kunda, 1990), and it makes us ready to defend those we can. The second is that the extent of solitary reasoning we observe in our cultures might be due to the huge differences between the environment in which we evolved – the environment for which reasoning and its myside bias were shaped – and the current environment. Denizens of modern societies are exposed to an unprecedented variety of opinions, along with the many disagreements this leads them to anticipate.

Debiasing the myside bias or making the best of it?

Many debiasing techniques have been tried to reduce the impact of the myside bias (for review, see Lilienfeld, Ammirati, & Landfield, 2009). They have met with limited and temporary success. One of the techniques that sometimes works is to ask participants to consider reasons for answers different from their own. However, in other cases participants have had trouble coming up with reasons why they might be wrong (Kuhn, 1991; Perkins, 1989), so the scope for the efficacy of this technique might be limited.

A more promising approach is to make the best of the myside bias. When people who disagree exchange arguments, the myside bias, from a potential flaw, turns into an efficient way to divide cognitive labor. Instead of having to consider the pros and cons of every answer, each discussant only has to look for the pros of his answer, and the cons of the others' answers. To the extent that people feel free to voice contrary opinions, and that they have no outstanding pressure to stick to their original ideas, the myside bias should not stop the best ideas from spreading in the group (for review, see Mercier & Sperber, 2011). For this to work, however, people must have different opinions to start with. If this is not the case, then substitutes can be used – devil's advocates, red teams – but they are not quite as efficient as genuine dissent.

Performance on the Wason selection task illustrates the stark contrast between the two approaches. Debiasing techniques that target individual performance – teaching people about conditionals, or explaining the good answer to another version of the task – have had very limited effects. By contrast, asking participants to solve the task in small groups creates, on average, a fivefold increase in performance (e.g., Moshman & Geil, 1998). Such results strongly suggest that psychologists and educators should focus less on trying to improve solitary reasoning at all costs, and teach instead how to make the best of interactions with others.

Summary

- The confirmation bias is a tendency to seek information that confirms a thought we currently consider.
- Famous experiments – the rule-discovery task, experiments on interrogation behavior, the four-cards selection task – purported to demonstrate the existence of a confirmation bias. Properly interpreted, these experiments in fact reveal rational behavior, not a confirmation bias.
- However, other experiments reveal a myside bias: A tendency to find arguments that defend our beliefs, whether they are supportive (if we agree with something) or refutational (if we disagree with something).
- The myside bias can have dire consequences when people reason on their own or with like-minded peers, leading to overconfidence and polarization.
- However, the myside bias does not preclude people from accepting arguments that challenge their position, provided their counter-arguments can be addressed.
- Standard cognitive and motivational accounts of the myside bias are unconvincing. Instead, the myside bias might play an adaptive role in making people better at finding arguments in order to convince others.
- The best way to deal with the myside bias might not be to fight it by trying to improve on solitary reasoning, but to make the best of it by making people exchange arguments with each other.

Further reading

Nickerson (1998) provides a good overview with many intriguing examples from outside the lab. Oswald and Grosjean (2004) provide an earlier review that details how to administer the interrogation behavior task. For specific paradigms, see Klayman and Ha (1987) and Poletiek (1996) for hypothesis testing; Sperber et al. (1995) for the Wason selection task; Hall et al. (2012) for choice blindness; Kunda (1990) for motivated reasoning; and Kuhn (1991) for informal arguments. Mercier and Sperber (2011) provide an introduction to the argumentative theory of reasoning, with a lengthy discussion of the role the confirmation bias (which should have been referred to as myside bias) plays in it. For a review of the debiasing literature, see Lilienfeld et al. (2009).

Note

1 http://jonronson.posterous.com/julie-burchill-and-psychopaths (retrieved February 25 2013).

References

Ball, L. J., Lucas, E. J., Miles, J. N., & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *The Quarterly Journal of Experimental Psychology*, *56*(6), 1053–1077.

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology*, *71*, 5–24.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ: Lawrence Erlbaum.

Evans, J. St. B. T., & Lynch, J. S. (1973). Matching bias in the selection task. *British Journal of Psychology*, *64*(3), 391–397.

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PloS One*, 7(9), e45457.

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, *50*(6), 1141–1151.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211–228.

Kuhn, D. (1991). *The skills of arguments*. Cambridge: Cambridge University Press.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498.

Lilienfeld, S. O., Ammirati, R., & Landfield, K. (2009). Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare? *Perspectives on Psychological Science*, *4*(4), 390–398.

Lucas, E. J., & Ball, L. J. (2005). Think-aloud protocols and the selection task: Evidence for relevance effects and rationalisation processes. *Thinking & Reasoning*, *11*, 35–66.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning*, *4*(3), 231–248.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomena in many guises. *Review of General Psychology*, *2*, 175–220.

Oswald, M. E., & Grosjean, S. (2004). Confirmation bias. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (pp. 79–96). Hove, UK: Psychology Press.

Perkins, D. N. (1989). Reasoning as it is and could be: An empirical perspective. In D. M. Topping, D. C. Crowell, & V. N. Kobayashi (Eds.), *Thinking across cultures: The third international conference on thinking* (pp. 175–194). Hillsdale, NJ: Erlbaum.

Poletiek, F. H. (1996). Paradoxes of falsification. *Quarterly Journal of Experimental Psychology*, *49A*, 447–462.

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, *36*(11), 1202–1212.

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, *57*, 31–95.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*(4), 259–264.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*(2), 147–168.

Trope, Y., & Bassok, M. (1983). Information-gathering strategies in hypothesis-testing. *Journal of Experimental Social Psychology*, *19*(6), 560–576.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129–137.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology: I* (pp. 106–137). Harmandsworth, England: Penguin.

6

ILLUSORY CORRELATION

Klaus Fiedler

As organisms learn to predict and control their environment through repeated observation, they have to assess the correlations that exist between important stimulus events. What signals come along with danger and safety? What behaviors are forbidden or allowed? Which traits characterize which social group? Or more generally, which causes precede which effects or consequences? The ability to figure out the correlations that hold between signals and their meanings, behaviors and reinforcements, groups and social attributes, or causes and effects constitutes a basic module of adaptive cognition.

The phenomenon of illusory correlation

If this central ability is impaired or distorted, organisms can be misled into erroneous predictions and decisions with detrimental consequences. For instance, the failure to learn which stimuli feel pleasant versus painful can cause much discomfort in a young child. Or, erroneously inferred correlations between symptoms and diseases can lead to false medical diagnosis. However, while the detection of environmental correlations appears to be crucial for survival and everyday problem solving, experimental evidence on the accuracy of correlation assessment is split. Although many findings testify to humans' and animals' high sensitivity to differential event frequencies (Alloy & Abramson, 1979; Malmi, 1986; Zacks, Hasher, & Sanft, 1982), an even larger body of evidence is concerned with biased subjective correlations that diverge from the actually encountered correlation (Crocker, 1981; Fiedler, 2000). The illusion of seeing a correlation that was not really there is termed an illusory correlation. More generally, the term applies not only to over-estimations of zero correlations but to all kinds of systematic biases in subjective correlation assessment.

Some "classical" examples help to illustrate the phenomenon. In a classical study on illusory correlations in diagnostic reasoning, Chapman and Chapman (1967)

showed their participants a series of draw-a-person test pictures, each with an indication of the problem that characterized the person who had allegedly drawn the picture. Participants persistently believed to have seen correlations that conformed to common diagnostic stereotypes. For instance, they reported that patients with worries about manliness had often produced drawings with pronounced shoulders, whereas patients characterized as suspicious would often highlight the eyes in the drawings, although in fact all combinations were equally often. In a similar vein, Hamilton and Rose (1980) described stimulus persons by vocational categories along with personality traits. Even though all vocational groups appeared equally often with all traits, participants believed to have seen mostly expected pairings, such as accountant/perfectionist or doctor/helpful.

While the illusions obtained in these two studies obviously originate in the participants' pre-experimental expectancies, or stereotypical knowledge, other variants of illusory correlations can be found when the use of neutral or meaningless stimulus materials rules out prior beliefs. Text box 6.1 presents one often cited study of this kind, as conducted by Kao and Wasserman (1993).

TEXT BOX 6.1 AN EXPERIMENT CONDUCTED BY KAO AND WASSERMAN (1993)

The experimental task referred to an unknown exotic plant, the Lanyu. Participants were asked to rate the "value of a fertilizer in promoting the Lanyu to bloom". They were fed with information about the frequencies with which the effect (blooming) occurred or did not occur when the cause (fertilizer) was given or not. Given an actual zero correlation because the relative blooming rate was the same in the presence as in the absence of the fertilizer, the fertilizer was nevertheless judged to have a positive causal impact on blooming when the absolute frequency of blooming (i.e., the outcome density) was high (e.g., when the frequencies of blooming and not-blooming were 19 vs. 7, respectively, both with and without the fertilizer). In contrast, when the absolute blooming frequency was low (e.g., 7 vs. 19), the fertilizer was judged to have a negative impact. Likewise, the fertilizer's causal impact was perceived to be positive (negative) when an equally high rate of blooming and not-blooming occurred more (less) frequently in the presence than in the absence of the fertilizer. Apparently, these findings indicate that one event combination, the co-occurrence of a present cause (fertilizer) with a present effect (blooming), receives a higher weight in correlation assessment than the events involving the absence of a cause or effect.

A third variant of illusory correlations, originally found by Hamilton and Gifford (1976), is explained in Text box 6.2. Hamilton and Gifford's (1976) seminal

TEXT BOX 6.2 AN EXPERIMENT CONDUCTED BY HAMILTON AND GIFFORD (1976)

Participants were presented with 39 descriptions of behaviors, each committed by a member of one of two groups. To rule out any preconceptions, the groups were simply denoted A and B. Group A was the majority and B was the minority, as a larger number of behaviors was associated with A (26) than with B (13). Within both groups, there were clearly more positive than negative behaviors, in accordance with the fact that in reality positive behavior is normal and negative behavior is norm-deviant and therefore unusual. The resulting stimulus distribution comprised 18 positive A behaviors, 8 negative A behaviors, 9 positive B behaviors, and 4 negative B behaviors. Note that the equal positivity rate for Group A (18/26) as for Group B (9/13) yielded a perfect zero correlation. Nevertheless, participants arrived at systematically less positive judgments of the minority than the majority. This was evident in various dependent measures, such as frequency estimates, evaluative group impression ratings, and cued-recall assignments of positive and negative behaviors to Groups A and B.

finding, which was replicated in countless other experiments, even with children (Primi & Agnoli, 2002), has obvious implications for the creation of minority devaluation and discrimination.

Definitions

From these various examples, it is evident that the term "illusory correlation" does not reflect a single phenomenon. Here we have introduced three different variants: *expectancy-based* illusory correlations, illusions arising from *unequal weighting of information*, and illusory correlations reflecting *unequal opportunity to learn* (e.g., more opportunity to learn about a majority than about a minority).

To delineate these different influences leading to illusory correlations more precisely, we introduce the following notation, assuming the simplest case of a correlation between two dichotomous variables x and y (see Table 6.1), though illusory correlations are not confined to dichotomous variables. To illustrate, let x denote a cause and y an effect; think of the causal influence of weather (x) on mood (y). The two levels on the first variable, $x+$ and $x-$, may represent good and bad weather, respectively, and $y+$ and $y-$ indicate good and bad mood, respectively. An elementary stimulus event s(x, y) in a stimulus series specifies the joint occurrence of one x-level with one y-level. For example, the stimulus series consist of pictures showing good or bad weather in the background and a human face expressing good or bad mood in the foreground. The frequency distribution of all four combinations yields

TABLE 6.1 Common notation to describe the correlation between two dichotomous variables

	Attribute y+ *Effect present*	*Attribute* y− *Effect absent*
Attribute *x*+ Cause present	**Cell A Frequency *a***	**Cell B Frequency *b***
Attribute *x*− Cause absent	**Cell C Frequency *c***	**Cell D Frequency *d***

a 2 × 2 table as in Table 6.1. Cell entries a, b, c, d indicate the joint frequencies with which good and bad weather co-occurs with good and bad mood. Various correlation coefficients can be defined as a function of a, b, c, d to measure the objectively existing correlation in the series. The degree of a causal relation can be quantified as the contingency

$$\Delta = a/(a + b) - c/(c + d),$$

which is the difference in the good-mood proportion given good weather, $a/(a + b)$, minus the good-mood proportion given bad weather, $c/(c + d)$. Another convenient measure is the coefficient

$$\phi = (ad - bc)/\sqrt{(a + b)(c + d)(a + c)(b + d)}.$$

Although the choice of an arbitrary normative model for correlation assessment presents a problem in its own right, many illusory-correlation findings are robust enough to generalize across different measures (McKenzie, 1994).

Experimental task and dependent measures

In the correlation-assessment paradigm, participants are exposed to a set of stimuli in which combinations of two attributes, x and y, occur with joint frequencies a, b, c, d. In some studies, frequencies are presented as statistical summary tables, as in Table 6.1, but richer insights into the cognitive process of correlation assessment come from experiments in which participants must extract the event frequencies from a more or less extended series of raw observations (e.g., from photos depicting weather and mood). The assessment task can be explicit in that the relevant attributes x and y are clearly identified from the beginning and participants are instructed to figure out the correlation. Or, the task may be implicit or incidental, such that stimuli are observed with another orienting task in mind (e.g., rating photos for pleasantness) and the call for retrospective correlation assessment may later come as a surprise. The amount and complexity of the task can further vary as a function of the total number of observations, their distribution across the four event

categories, the visibility of the variables *x* and *y* against the background of irrelevant context variables, and the pre-experimental knowledge about the relation of *x* and *y* and their meaning.

The cognitive process of correlation assessment encompasses several stages. Stimulus observations must be *classified* as either *x*+ or *x*–, either *y*+ or *y*–; observations have to be *perceived* and *encoded* attentively, and the distribution of the four event classes has to be somehow extracted and *integrated* in memory. Finally, the resulting memory *representation* must be mapped onto *responses, judgments* or *decisions*, which are indicative of the sign and size of the subjective correlation, the *accuracy* of which can then be assessed relative to the objectively presented correlation.

Explicit and implicit dependent measures are used in illusory-correlation experiments. The most common explicit measures include direct ratings of the size of the observed correlation on numerical or graphical scales, or estimates of the event frequencies *a*, *b*, *c*, *d*, from which the perceived correlation can then be computed (according to the chosen model, Δ, ϕ, etc.). Yates, McGahan, and Williamson (2000) suggested an alternative measure consisting of various ratings of propositions implying different correlations. Implicit measures of subjective correlations rely on choices or decisions that presuppose correlation knowledge without asking participants to express this knowledge directly on some quantitative scale. For example, having observed a series of weather-mood combinations, participants may be presented with a series of cards, drawn from the same pool as the stimulus series, that show a smiling or frowning face (symbolizing good or bad mood) on one side and their task is to predict the weather situation shown on the other side of the card.

Theoretical accounts of illusory correlations

Different theoretical explanations have been advanced to account for the three classes of illusory correlations depicted at the outset.

A major domain of *expectancy*-driven illusory correlations is the study of social and diagnostic stereotypes. The basic theoretical intention is to demonstrate the top-down impact of prior knowledge that can override the bottom-up processing of the stimulus data proper. For instance, participants may believe that mood generally improves when the weather is fine. This prior belief may then be used for guessing when they have to judge under uncertainty the number of smiling and frowning faces associated with good versus bad weather. In such an account, the illusion is attributed to guessing in the final judgment stage. The initial perception and encoding stages may also be biased, especially when the stimulus observations are ambiguous so that prior knowledge is required to classify observations according to *x* and *y*. It may be postulated, in addition, that expectancies also influence stimulus learning, giving an advantage to learning expected stimuli (smiling faces & sunny weather; frowning faces & rainy weather) as opposed to unexpected stimuli (smiling & rainy; frowning & sunny). However, this additional assumption is not necessary and, by the way, not patently supported empirically; there is indeed

evidence to suggest more effective encoding of *unexpected* rather than expected events at least under some conditions (Stangor & McMillan, 1992).

Expectancy-based illusory correlations are often confused, and actually equated, with *similarity*-based illusory correlations, which is not justified conceptually. Similarity is a stimulus property, whereas expectancies reside within the individual. One can increase the similarity of the stimulus display for good mood and sunny weather by adding some common perceptual features (e.g., same color, common symbols, smile in both the face and the sun, etc.) while holding expectancies constant. Such overlap in common features may enhance the experienced correlation (Fiedler, 2000; Plessner, Freytag, & Fiedler, 2000), but unlike expectancy effects, this reflects a stimulus-driven encoding influence.

Alternatively, biased correlation assessments may reflect the *unequal weighting* of the four cells in Table 6.1, that is, more weight given to present events or committed behaviors than to absent events and omitted behaviors. Thus, when the task focuses on the presence of the sun and the presence of good mood (in a slightly revised example), then the critical features are present when there is sunny weather and good mood but absent when there is rainy weather and bad mood. Due to the asymmetry of present and absent features, known as feature-positive effect (Newman, Wolff, & Hearst, 1980), a typical finding is that cell frequency a (i.e., the number of present effects & present cause) receives the highest weight in correlation assessment (Wasserman, Dorner, & Kao, 1990), followed by b (missing effect & present cause), and c (present effect & absent cause), while the least weight is given to d (missing effect & absent cause). As a consequence, two formally identical correlations can give rise to different subjective assessments. Observing $a = 20$ instances of good mood in sunny weather along with a constantly lower frequency in the other cells, $b = c = d = 10$, will be experienced as a stronger correlation than observing $a = b = c = 10$ and $d = 20$ (i.e., frequent experience of the absence of good mood & sunny weather). Such unequal weighting can be attributed to the enhanced salience of present features in the early stages of perception and encoding. Theoretically, however, unequal weighting can also affect the integration stage when observations from all four cells are combined to yield an overall judgment.

Just as present stimulus features are assumed to be more salient than absent features in the latter paradigm, enhanced salience and a memory advantage have also been attributed to rare or distinctive attribute levels, according to the famous von Restorff (1933) effect. This explanation mainly applies to illusory correlations of the third kind, due to *unequal opportunity to learn* about minority and majority cases. As explained earlier in Text box 6.2, the same high proportion of, say, 75% desirable behavior in a minority (defined by a small absolute sample size) leads to less positive impressions than the same proportion observed in a majority (large sample size), even though the constant proportion implies a zero correlation. The *distinctiveness* account (Hamilton & Sherman, 1989) of this frequently replicated finding states that the combination of the two infrequent attribute levels, that is, undesirable behavior by the minority, has a distinctiveness advantage, rendering these exceptional behaviors particularly salient and likely to be kept in memory.

However, the available evidence does not support an encoding and memory advantage of the most infrequent cell or stimulus combination (Fiedler, Russer, & Gramm, 1993; Klauer & Meiser, 2000; Kutzner & Fiedler, 2015). The phenomenon can be explained more simply as a *sample-size effect*. Learning theories predict that learning increases with the number of trials. There is more opportunity to learn for the majority than for the minority that most behaviors tend to be desirable, just due to unequal sample size (Fiedler, 1996).

The sample-size account can be set apart from the distinctiveness account when illusory correlations are studied in a hypothesis-testing paradigm (see Chapter 5 in this volume). Translating pertinent findings (cf. Fiedler, Walther, & Nickel, 1999) to the present example, we can ask participants to engage in active information search in order to test the hypothesis that sunny weather produces good mood. This can be accomplished by letting participants search, within a restricted time period, for relevant mood and weather entries in somebody's diary. The diary can be constructed such that the base rate of good-mood entries is 70%, and that this high good-mood rate is the same for days described as sunny and rainy, yielding an objective zero correlation. A common information search strategy in such a situation is *positive testing* (Klayman & Ha, 1987); the search focus should be mainly on features mentioned in the hypothesis, sunny weather and good mood. Despite the constant good-mood rate across weather conditions, positive testing should produce a skewed sample:

Good mood & sunny:	14
Bad mood & sunny:	6
Good mood & rainy:	7
Bad mood & rainy:	3

Note that the higher frequencies of sunny days and good mood reflect the participants' own self-determined focus of information search and encoding. It is therefore no longer possible to assume an encoding focus on the rarest events, which are by definition the least attended stimuli. As a consequence, if an illusory impression arises that good mood was more pronounced for sunny than for rainy days, this would be consistent with the enhanced opportunity to learn from a larger sample. However, it could be hardly attributed to the enhanced salience or attention given to the rarest events.

Parenthetically, the hypothesis-testing approach can be easily extended to investigate the joint influence of different sources of illusory correlations within the same experiment – a possibility largely neglected in previous research. For instance, the hypothesis to be tested might focus on an unexpected event (say, good mood on rainy days) while sample size is higher for bad mood and sunny days, thus pitting the sample size against the expectedness account. Or, the hypothesis might focus on an absent rather than a present feature, asking participants to find out whether good mood is most likely when the sun is absent. This amounts to pitting attention focus against the default focus on present features. Because integrating different sources

of illusory correlations within a single comprehensive approach is a major task for future research, this is the focus of the demonstration experiment elaborated below and in Text box 6.3.

TEXT BOX 6.3 PROCEDURAL DETAILS FOR A SAMPLE EXPERIMENT

Participants and design

To keep the experiment simple, the basic design includes only one between-participants factor, *sample size of expectancy-consistent versus -inconsistent behaviors*, along with a repeated-measures factor for *measures of expectancy-consistent versus inconsistent information*. Prior research suggests that 30 participants per condition should be sufficient to demonstrate illusory correlations. Or, starting from an estimate of the expected effect size, one might use G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) to determine the sample size required to reach a certain level of statistical power ($1 - \beta$).

Materials

Each learning trial starts with a pairing of one item of overt or covert aggression (drawn randomly from the list provided in the Appendix) and a male or female face, framed as a question: Has this behavior been observed in this target person? A feedback presented a few seconds later then indicates whether the data-base from a systematic assessment of the targets' behavior confirms or disconfirms the pair, that is, whether the behavior is "present" or "absent". Across all trials, participants can thus observe the frequencies with which overt and covert aggression is present or absent in male and female targets.

The confirmation rate is 75%, for both types of aggression (overt and covert) in both targets (male and female). However, the absolute sample size of the four combinations of aggression type and target gender is manipulated. In one condition, sample size (and hence the opportunity to learn) is higher for expectancy-consistent combinations. The learning list includes:

12 present and 4 absent cases of male overt aggression,
12 present and 4 absent cases of female covert aggression, but only
6 present and 2 absent cases of male covert aggression,
6 present and 2 absent cases of female overt aggression.

Note that twice as many learning trials refer to expected combinations (male & overt; female & covert) than to unexpected combinations. The unequal opportunity to learn should thus reinforce the expectancy-based correlation

bias. Despite the zero correlation (i.e., the constant 75% rate), participants should come to believe that aggression tends to be overt in males but covert in females.

In another experimental condition, however, sample sizes are higher for unexpected combinations. Here the learning list includes:

6 present and 2 absent cases of male overt aggression,
6 present and 2 absent cases of female covert aggression, but as many as
12 present and 4 absent cases of male covert aggression,
12 present and 4 absent cases of female overt aggression.

In this condition, the unequal opportunity to learn should counteract the expectancy-biased correlation effect. Consequently, expectancy-driven illusory correlations should be reduced, eliminated or even reversed.

Procedure

Prior to the experiment, the proposed gender stereotype should be tested by asking participants to rate the degree to which they believe that males and females tend to exhibit overt and covert aggression. Of course, the participants' own gender will also be assessed together with some demographic attributes (e.g., age and profession).

A cover story should be constructed to render the task interesting and meaningful. For example, the hypothesis to be tested can be embedded in a diagnostic task setting: The male and female target may be two individuals living together in a close relationship. In the course of a partner therapy, it is important to figure out whether different aggression styles characterize the male or female partner. In this context, a constant male and female face is required. Alternatively, the cover story might ask participants to imagine that they start working on a new job and that their task is to find out whether an aggression type occurs more frequently among male or female colleagues. In this case, variable faces representing different individuals would be needed.

Further instructions should be as explicit as possible regarding the hypothesis to be tested: "The crucial question you are supposed to answer is whether, across all behaviors, overt aggression is more likely to be paired with male than female persons, whereas covert aggression is more likely to be paired with female than male persons."

Instructions should clarify that the task refers to the relationship between gender and aggression *in the stimulus list*, as distinguished from the participants' general beliefs about the correlation in reality. As soon as participants have read the instructions, stimuli are presented at a rate of 8 or 10 seconds per item with an inter-trial interval of one or two seconds between items. The presentation order should be randomized. Given a stimulus distribution of 12

and 4, or 6 and 2 present and absent aggression items, respectively, the first and second half of the series should reflect the same distribution (i.e., 6 and 2 and 3 and 1), to rule out ordinal position effects.

Three dependent measures are used to capture illusory differences between male and female aggression: (a) frequency estimates of present and absent cases of overt and covert aggression in male versus female targets; (b) impression ratings of the male and female target on several (5 to 10) adjective scales related to overt and covert aggression; and (c) a cued-recall test whereby all observed stimulus behaviors are presented once more and participants have to recall whether the behavior had been paired with a male or a female target. The order of the three dependent measures may vary, but frequency estimates should not follow after the cued-recall test, because the repeated presentation of items for cued recall would obscure the frequency estimates.

Finally, participants should be thoroughly debriefed concerning the study purpose.

Illusory-correlation experiment

The demo experiment is intended to elucidate the interplay of top-down expectancies and bottom-up stimulus effects in correlation assessment. It was never conducted exactly as described here but closely resembles earlier research by Fiedler et al. (1999).

Correlation assessment is embedded in a hypothesis-testing task. Participants are asked to test the hypothesis that "Overt aggression is more likely among males than females, whereas covert aggression is more likely among females than males." Accordingly, they are presented with a series of photographs, showing either a male or female person, coupled with a verbal behavior description, which entails either the presence or the absence of (overt or covert) aggression. Thus, the stimulus list constitutes a twofold 2 × 2 distribution (male vs. female × present vs. absent), one for overt aggression and one for covert aggression. Both correlations are actually zero, as in both domains the proportion of present aggression is the same for males and females. The constant proportion is high (viz., 75%) so that aggression is more often present than absent both in males and females, and both in the domain of overt and covert aggression.

An *expectancy-based illusory correlation* should lead participants to report having seen more behaviors consistent with the stereotypical expectancies (overt aggression in males, covert aggression in females) than expectancy-inconsistent behaviors (overt aggression in females, covert aggression in males). At the same time, however, correlation assessment should depend on another source of bias, which is stimulus-driven rather than expectancy-driven. Thus, although the relative proportion of present aggression is constantly high (i.e., 75%), the absolute frequency or sample size varies. In one condition, in which larger samples provide more opportunities

to learn the 75% confirmation rate for male overt aggression and female covert aggression than for male covert and female overt aggression, the *stimulus-driven learning effect* should reinforce the expectancy-driven illusion. Both expectancies and learning opportunities should reinforce the link from male gender to overt aggression and from female gender to covert aggression. In contrast, when absolute sample size is larger for male covert and female overt aggression than for male overt and female covert aggression, then the impact of opportunity to learn should be in conflict with stereotypical expectancies. Depending on which tendency is stronger, the impact of sample size should markedly reduce or eliminate or even reverse the expectancy bias. Procedural details are given in Text box 6.3.

Results

Let us first summarize the theoretical predictions before we turn to the analysis of the experimental results. Two possible findings can be expected, reflecting different sources of illusory correlation. On one hand, expectancy-based illusory correlations should be manifested in a general tendency to confirm the gender stereotype: Responses to all dependent measures should on average reflect the erroneous belief that male aggression tends to be overt, whereas female aggression tends to be covert. Inter-individual differences in the gender-stereotypical belief (as assessed at the start) should predict the strength of illusory correlation. On the other hand, the sample-size manipulation should produce a different kind of illusory-correlation effect that may reinforce or counteract the expectancy-based illusion. If sample size is larger for expectancy-consistent combinations of target gender and aggression types, then the learning-based illusion should strengthen the expectancy-based illusion. If, however, the sample is larger for expectancy-inconsistent combinations, the learning effect should reduce, eliminate or even reverse the expectancy effect.

For a statistical test of these predictions, summary scores for expectancy-congruent and -incongruent responses are computed for all dependent measures. For the frequency estimates, this amounts to calculating for each participant the estimated proportion of present aggression (i.e., the estimate of present divided by the summed estimates of present and absent aggression) for each combination of gender and aggression type. The two scores required for the data analysis are then the average estimated proportions for expected aggression (overt male and covert female) and for unexpected aggression (covert male and overt female). The two corresponding scores from the impression ratings are the average ratings of male targets on overt aggression scales and female targets on covert-aggression scales (expected) and the average ratings of male targets on covert-aggression scales and female targets on overt-aggression scales (unexpected). On the cued recall test, too, when the task is to recall the gender of the target associated with each behavior, the two scores are computed as the proportions of cued-recall responses consistent or inconsistent with stereotypical expectancies.

The results reported below are borrowed from the Fiedler et al. (1999) studies, which slightly differed from the present experiment in terms of the specific sample

sizes. First, a check on whether most participants really shared the stereotypical expectancy that aggression tends to be overt in males and covert in females provided support for this premise. The average rating on a graphical scale from –21 to +21 (with positive values indicating an increasing degree of belief that aggression is overt in males and covert in females) was +5.54 and clearly above 0, $t(73) = 4.18$, $p = 0.00006$.

Turning to the dependent measures proper, the results are considered separately for the two experimental groups (i.e., sample-size conditions). When sample size was larger for expectancy-consistent events (see left pair of bars in both charts of Figure 6.1), the mean estimated proportion of expectancy-consistent behaviors (pooling over male-overt and female-covert) was 66.7% as compared with an estimated proportion of 43.5% expectancy-inconsistent behaviors, $F(1, 31) = 16.98$, $p = 0.0003$. The effect size here amounted to approximately two thirds of a standard deviation.

Likewise, on the impression-rating measure, the mean rating for targets on expectancy-consistent traits (e.g., male-brutal, female-shining) was higher (+4.58) than for inconsistent traits (–1.27), $F(1, 31) = 8.13$, $p = 0.008$. The effect size here was about one standard deviation. As evident from Figure 6.1, both the estimated percentage of confirmed aggression (left chart) and the average impression rating on relevant adjective scales (right chart) were higher for expected (black bars) than for unexpected combinations (shaded bars).

On the cued-recall test, which is not reported in detail, the proportion of aggressive behaviors recalled as belonging to expected pairings (i.e., overt aggression assigned to male person and covert aggression assigned to females) was also higher than cued recall of unexpected pairings.

FIGURE 6.1 Mean percentage estimates (left chart) and mean impression ratings (right chart) for expectancy-consistent (black bars) and expectancy-inconsistent (shaded bars) attributes as a function of sample-size conditions.

However, despite the gender-stereotypic expectancies and the instruction focus on gender-typical aggression, a strong reversal was obtained when sample size was larger for expectancy-inconsistent pairings (see right pairs of bars in each chart). Thus, when there was more opportunity to learn the constantly high confirmation rate of 75% aggression in the female-overt and the male-covert domain than in the expected domains, the mean estimated percentage of confirmed aggression was higher for expectancy-inconsistent (52.4%) than for expectancy-consistent behaviors (43.6%), $F(1, 41) = 4.18$, $p = .045$. Similarly, mean impression ratings for targets were higher on unexpected traits (+4.29) than on expected traits (+1.14), $F(1, 41) = 8.83$, $p = 0.005$.

Discussion

Thus, frequency estimates and impression ratings were dominated by sample-size effects, which largely overrode the impact of prior expectancies. This finding highlights the need to take the so-far neglected impact of learning opportunities into account.

Suffice it to mention briefly that in the original investigation by Fiedler et al. (1999), the impact of sample size also dominated another manipulation, namely, the focus of hypothesis testing. When participants were asked to test the anti-stereotypical hypothesis that aggression tends to be covert in males and overt in females, this also led to illusory correlations consistent with the focal hypothesis, but only if sample size enhanced the opportunity to learn about male covert and female overt aggression. The illusion was reversed if unequal sample size enhanced the opportunity to learn about overt-male and covert-female aggression. Thus, the sample-size effect dominated the impact of both stereotypical expectancies and experimentally induced attention focus.

Although the opportunity to learn overrode the influence of stereotypical expectancies in the experiment depicted here (as anticipated by Fiedler et al., 1999), this is of course not to say that analogous findings can be generalized to illusory correlations studied in many other task settings. The ultimate purpose of an integrative, multi-factor approach is to study systematically how the relative impact of different sources of bias depends on such boundary conditions as the degree of uncertainty and noise in the stimulus materials, the motivational pay-off structure of the task, the stimulus-presentation mode and precise encoding conditions, the degree of memory load and decay, and the presence of meta-cognitive monitoring and correction processes.

Conclusions

Detecting and estimating correlations between attributes of significant environmental objects is an important module of adaptive intelligence and behavior. Although humans as well as animals seem to have the competence to assess environmental correlations quite accurately, their performance is often impaired under less than

optimal conditions. Numerous experiments on illusory correlations converge in demonstrating that subjective correlation estimates are often distorted as a function of prior knowledge, attention, asymmetric representation of variable levels, sample size, similarity, and motivational factors. The degree of distortion can be quite severe and some types of illusory correlations can be reproduced easily and can hardly be eliminated through training. In some domains, the existence and size of these illusions are interesting in their own right. For instance, economists and consumer researchers are interested in the perceived correlation between price and quality of consumer products. In social psychology, the perceived correlation between trait attributes and group membership provides a basic building block for theories of stereotyping.

However, although evidence for illusory correlations is strong and uncontested, one should be cautious in drawing ideological conclusions about human irrationality. From a broader theoretical perspective, illusory correlations can also be considered as indicators of adaptive intelligence. Many illusions, in perception and cognition, represent the backside of otherwise adaptive and prudent algorithms. For instance, the higher weight given to present than to absent attributes might be indispensable for survival. An absent traffic sign would be a less useful guide for safety than the presence of signs. Similarly, the effect of sample size makes sense if information is unreliable and organisms have to check for reliability before inferring correlations. An observed proportion of 8 out of 12 provides stronger evidence than 2 out of 3.

Conversely, a well-adapted organism not only has to be accurate but also quick and must not waste too many resources on each and any task. Using simplified algorithms that produce errors some of the time may be preferable to more demanding algorithms in the long run. In this regard, McKenzie (1994) has shown through Monte Carlo simulations across most reasonable distributions that primitive algorithms of correlation assessment are strongly correlated with more refined correlation measures. For instance, the sum of diagonal cell frequencies ($a + d$ in Table 6.1) is highly correlated with the full-fledged ϕ coefficient, especially when the marginal frequencies of the two levels on x and y are approximately equal. Thus, if the cut-off points that distinguish between "good versus bad weather" and between "good versus bad mood" are chosen such that each variable level occurs at roughly 50%, then merely counting the diagonal sum (i.e., the relative frequency of good weather & good mood or bad weather & bad mood) provides a perfect estimate of the correlation. The diagonal sum will only be misleading when marginal frequencies are unequal. Indeed, highly skewed distributions can render correlation assessment obsolete. Given over 80% good mood, the best strategy to predict mood is to always predict the high base-rate event (good mood), rather than trying to infer mood from the weather (Kareev & Fiedler, 2006; see Chapter 2 in this volume).

Thus, although many illusory-correlation experiments provide cogent evidence for erroneous and fallacious reasoning on the specific task, one should refrain from premature pessimistic conclusions about human intelligence and rationality.

Summary

- The detection and assessment of environmental correlations is an important module of adaptive intelligence.
- The phenomenon of illusory correlations refers to partly severe failure and inaccuracy in correlation assessment.
- Different types of illusory correlations can be distinguished in terms of underlying cognitive processes: Illusions can be based on expectancies, unequal weighting of present and absent events, and different opportunities to learn about frequent and infrequent targets.
- New insights can be gained from studies designed to pit different sources of illusory correlation against each other. The demo experiment outlined here is designed to pit the knowledge-driven impact of prior expectancies against the stimulus-driven influence of unequal opportunities to learn.
- The most prominent applied settings for illusory-correlations research include diagnostics, economical decisions, evaluation, hypothesis testing, and the social psychological domain of group stereotypes.

Further reading

More comprehensive reviews and advanced readings on illusory correlations and correlation assessment are available. Allan (1993) conceptualizes correlation assessment in an associative-learning framework. A more recent approach by Allan, Hannah, Crump, and Siegel (2008) complements associative learning with a psychophysical signal-detection framework that takes payoff structures and response strategies into account. Alloy and Tabachnik (1984) offer an intriguing view on human and animal performance on correlation-assessment tasks. Fiedler (2000) provides a review of different variants of illusory correlations within the same connectionist framework. And, recent research on pseudocontingencies (Fiedler, Kutzner, & Vogel, 2013) suggests a completely new perspective on the cognitive process underlying illusory correlations.

References

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin*, *114*(3), 435–448.

Allan, L. G., Hannah, S. D., Crump, M. C., & Siegel, S. (2008). The psychophysics of contingency assessment. *Journal of Experimental Psychology: General*, *137*(2), 226–243.

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General, 108*, 441–485.

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review, 91*(1), 112–149.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72*, 193–204.

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin, 90*(2), 272–292.

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic, multiple-cue environments. *Psychological Review, 103*, 193–214.

Fiedler, K. (2000). Illusory correlations: A simple associative algorithm provides a convergent account of seemingly divergent paradigms. *Review of General Psychology, 4*, 25–58.

Fiedler, K., Kutzner, F., & Vogel, T. (2013). Pseudocontingencies: Logically unwarranted but smart inferences. *Current Directions in Psychological Science*, *22*(4), 324–329.

Fiedler, K., Russer, S., & Gramm, K. (1993). Illusory correlations and memory performance. *Journal of Experimental Social Psychology, 29,* 111–136.

Fiedler, K., Walther, E., & Nickel, S. (1999). The autoverification of social hypothesis: Stereotyping and the power of sample-size. *Journal of Personality and Social Psychology,* 77, 5–18.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology, 12*, 392–407.

Hamilton, D. L., & Rose, R. L. (1980). Illusory correlation and the maintenance of stereotypic beliefs. *Journal of Personality and Social Psychology, 39*, 832–845.

Hamilton, D. L., & Sherman, S. J. (1989). Illusory correlations: Implications for stereotype theory and research. In D. Bar-Tal, C. F. Graumann, A. W. Kruglanski, & W. Stroebe (Eds.), *Stereotype and prejudice: Changing conceptions* (pp. 59–82). New York: Springer.

Kao, S.-F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1363–1386.

Kareev, Y., & Fiedler, K. (2006). Nonproportional sampling and the amplification of correlations. *Psychological Science*, *17*(8), 715–720.

Klauer, K. C., & Meiser, T. (2000). A source-monitoring analyses of illusory correlations. *Personality and Social Psychology Bulletin, 26*, 1074–1093.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211–228.

Kutzner, F. L., & Fiedler, K. (2015). No correlation, no evidence for attention shift in category learning: Different mechanisms behind illusory correlations and the inverse base-rate effect. *Journal of Experimental Psychology: General*, *144*(1), 58–75.

Malmi, R. A. (1986). Intuitive covariation estimation. *Memory & Cognition, 14*, 501–508.

McKenzie, C. R. M. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology, 26*, 209–239.

Newman, J., Wolff, W. T., & Hearst, E. (1980). The feature-positive effect in adult human subjects. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 630–650.

Plessner, H., Freytag, P., & Fiedler, K. (2000). Expectancy-effects without expectancies: Illusory correlations based on cue-overlap. *European Journal of Social Psychology, 30*, 837–851.

Primi, C., & Agnoli, F. (2002). Children correlate infrequent behaviors with minority groups: A case of illusory correlation. *Cognitive Development*, *17*(1), 1105–1131.

Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin, 111*, 42–61.

von Restorff, H. (1933). Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychologische Forschung, 18*, 299–342.

Wasserman, E. A., Dorner, W. W., & Kao, S.-F. (1990). Contribution of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 509–521.

Yates, M. C., McGahan, J. R., & Williamson, J. D. (2000). Intuitive covariation assessment of the illusory correlation. *Journal of General Psychology*, *127*(4), 397–411.

Zacks, R. T., Hasher, L., & Sanft, H. (1982). Automatic encoding of event frequency: Further findings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(2), 106–116.

APPENDIX

TABLE 6.2 Stimulus items representing overt and covert aggression

	Overt aggression		*Covert aggression*
1.	Tends to use violence	1.	Becomes unfair in arguments
2.	Quickly goes too far with language	2.	Acts as though others were not there
3.	Shouts in arguments	3.	Enjoys disparaging others
4.	Threatens with violence	4.	Lies to get an advantage
5.	Screams when s/he doesn't like something	5.	Makes a pretense of being friendly with everyone
6.	Shakes people when angry	6.	Makes others feel sorry for him/her
7.	Quickly gets into a temper	7.	Hangs up when fed up
8.	Shouts others down	8.	Simply walks out of an argument
9.	Doesn't stop short of hitting people	9.	Plays with others' feelings
10.	Kicks things	10.	Gossips about people s/he doesn't like
11.	Was involved in a fistfight	11.	Cuts others after an argument
12.	Throws things around the room	12.	Pretends to be unforgiving
13.	Gets out of control quickly	13.	Schemes
14.	Sometimes smashes dishes	14.	Sets traps for others
15.	Defends his/her rights with violence	15.	Sets people against each other
16.	Easily gets into a rage	16.	Manipulates others to fight each other
17.	Likes to argue with people	17.	Denigrates others
18.	May slap someone's face when in rage	18.	Puts up a show
19.	Sometimes wants to smash something	19.	Pointedly ignores others
20.	Quickly has a fit after insults	20.	Flatters others
21.	Smashes pencil on the ground	21.	Alludes to others' weaknesses
22.	Insults others with swearwords	22.	Makes embarrassing gestures
23.	Bangs fist on the table	23.	Mocks other people
24.	Pokes tongue out at someone	24.	Ridicules others

Note. English translations of originally German items. (Items 21–24 were not included originally.)

Author's note

The author's own research on illusory correlations has been supported by various grants from the Deutsche Forschungsgemeinschaft. Thanks for helpful comments on a draft of the chapter are due to Florian Kutzner and Max Ihmels.

7

ILLUSIONS OF CONTROL

Suzanne C. Thompson

One of the enduring themes of psychological theory and research is that human beings are motivated to have control over the events of their lives (Rodin, 1986; White, 1959). Extensive research has demonstrated that perceived control is associated with many positive outcomes, including successful coping with stressful events, making health-improving lifestyle changes, and better performance on tasks (Thompson & Spacapan, 1991).

The central role of perceived control in many areas of functioning has led to a focus on the accuracy of personal control judgments. Illusions of control occur when individuals overestimate their personal influence over an outcome. For example, Peter takes an herbal supplement, echinacea, with the goal of avoiding colds and the flu. He is likely to attribute a period of good health to the supplement even if, in fact, it has only a minimal effect or perhaps no effect at all. At times individuals may judge that they have control even over a obviously chance process: People who play slot machines have been known to act as if their choice of machine or style of pulling the handle can affect their chances of winning.

Studies of illusions of control have taken three different approaches to demonstrating the existence of control illusions (see Text box 7.1). Ellen Langer conducted the first programmatic study of illusory control. Her approach was to examine people's perceptions of the likelihood of getting a desired outcome when the task involves chance situations with skill elements. In a series of studies, she showed that in chance situations with elements such as familiarity or choice, participants gave higher estimates of getting the outcomes they desire (Langer, 1975). For example, in one study, lottery tickets were decorated with familiar or novel symbols. The participants who received the familiar symbols were less likely to exchange their tickets for new ones even though the probability of winning was higher with a new ticket. It was assumed that the unwillingness to exchange the ticket indicated that participants believed they had control, that is, they chose a ticket that was more

TEXT BOX 7.1 CLASSROOM DEMONSTRATIONS

Each of the three ways of researching illusions of control can be used to demonstrate illusions of control in a classroom context. They are discussed below in ascending order of difficulty of preparation and time needed to complete the demonstration.

Demonstration 1: Illusions of control over driving

The questions used by McKenna (1993) can easily be adapted for classroom use. One set of questions focuses on the likelihood of an automobile accident when the participant is the driver vs. the passenger of the car (see Q1 and Q2 in Table 7.1). The other four questions focus on the circumstances of an accident with the participant as driver: For two questions (Q3 and Q4), the circumstances are low control; for the other two (Q5 and Q6), the circumstances are high control. These questions are rated on a –5 to 5 scale with 0 = average as the midpoint.

In the classroom experiment, students receive a handout with these six questions that they answer anonymously and turn in. A number of analyses should be done:

1. Compare answers to Q1 and Q2, with the independent variable of self as driver or passenger (paired *t*-test).
2. Average the two low-control questions and the two high-control questions and compare mean differences in those (paired *t*-test).
3. See if the mean for Q1 differs significantly from zero (one-sample *t*-test).

TABLE 7.1 Questions used to demonstrate illusory control over driving (adapted from McKenna, 1993)

Q1	Compared to other drivers, how likely do you think you are to be involved in an automobile accident when you are driving?
Q2	Compared to other drivers, how likely do you think you are to be involved in an automobile accident when you are a passenger?
Q3	Compared to the average driver, how likely do you feel you are to be involved in an accident that is caused by another vehicle hitting you from behind?
Q4	Compared to the average driver, how likely do you feel you are to be involved in an accident that is caused by an unexpected tire blow-out?
Q5	Compared to the average driver, how likely do you feel you are to be involved in an accident in which the vehicle you are in is driven into the rear of another vehicle?
Q6	Compared to the average driver, how likely do you feel you are to be involved in an accident in which the vehicle you are in is changing traffic lanes?

Note. Response scale from –5 (much less likely) through 0 (average) to 5 (much more likely).

The results can be graphed and presented at the next class session. The class is asked to guess the results of the comparison between Q1 and Q2 and they will accurately predict that ratings of the likelihood of an accident will be higher when the participant is listed as the passenger. The discussion of why this would be brings out the idea of illusory control. The issue of accuracy is often raised with some students protesting that they (or some of the respondents) ARE better drivers than someone they might ride with as a passenger. This is a good time to cover the difference between individual and group prediction. It is also useful to point out the items that have an absolute value of less than zero and have the class consider what that implies if, on average, the class ratings of being the driver in an accident is less than the mid-point of the scale.

Another issue this raises is that of being able to accurately assess one's own capabilities. In addition, some especially perceptive students will comment on alternative explanations for the results, in particular, that lower ratings of accidents when one is a driver could be a "better than average" effect (cf. Chapter 17), but not necessarily one that is due to overestimating one's control. At that point, the graphs that show the comparison between the low-control and high-control accident circumstances can be examined. Typically, the high-control accident circumstances are rated as significantly less likely to lead to an accident than the low-control circumstances. The comparison suggests that personal perceptions of control make a contribution to the effect.

There are a number of ways that this demonstration can be expanded. For example, number of traffic tickets received, gender, or self-esteem can be added to the questionnaire and analyzed to see if experience or personality measures predict the amount of illusory control.

Overall, this demonstration is easy to prepare and administer and very likely to yield results that demonstrate illusions of control. Using similar materials, McKenna (1993) found a mean of –1.41 for the driver condition as opposed to 0.01 for the passenger condition. Thus, participants judged that they had less likelihood of getting in an accident when driving than the average person, but not less likelihood when they were passengers.

Demonstration 2: Illusions of control in a gambling game

Demonstration 2 is based on Langer's (1975) research on illusory control, using a simple gambling game. The class is divided into pairs and one student in each pair is randomly assigned to be the participant or the observer. Each pair is given a pair of dice and a sheet for recording the outcomes of dice throws. The dice will be rolled 20 times by the participant and the results of each roll recorded and then summed across all 20 throws. There is a prize for the pair that gets the highest total. Before the dice throwing begins, each participant and observer gets a piece of paper asking them to separately and

anonymously rate the chances of getting the prize on a 0–10 scale from "no chance at all" to "an excellent chance". These measures are collected, the dice rolling is done, and the prize distributed. Before the next class, the analysis is done with role as the independent variable (participant vs. observer) and ratings of chance as the dependent variable (paired *t*-test).

Even though both the participants and the observers have no control over their dice-throwing score, there will be a slight tendency for the participants to rate the chances of getting the prize higher than the observers will. The results may not be significant with a smaller class, but the data can be saved and aggregated over several classes for stronger effects.

This demonstration uses a fairly realistic context (at least for those who play games of chance). If the demonstration works as proposed, it is an excellent experience for students to analyze the causes and effects of control illusions. If the results are not consistent with Langer's work, then the discussion can focus on the differences in the research set-up used by Langer and that of the demonstration. For instance, students may have felt pressure to make their ratings of their chances of getting the prize in a "rational" way, given the class context. Previous research has found that circumstances that highlight the right or "rational" way to estimate likelihoods reduce or eliminate illusions of control (Thompson, Armstrong, & Thomas, 1998).

Demonstration 3: Illusions of control on the computer

This demonstration requires a more elaborate set-up and most likely would need to be done outside of class time for later discussion. Thompson et al. (2004) adapted for computer use the Alloy and Abramson (1979) light onset task. In the original task, participants pushed a button to see if they could control the onset of a light. In our adaptation, experiment-presentation software (SuperLab) was used to set up a similar task. The software was used to present a series of screens, each of which displayed either a red "O" or a green "X". Participants were told that for each of 40 trials, they could choose to press or not press the space bar to get the green "X" screen to appear. Their job was to judge how much control they had over the appearance of the green "X" on the screen. The level of reinforcement was manipulated by the number of times the desired green "X" appeared (25% or 75% of the time). This was easy to set up by randomly assigning each of the 40 screens that participants will see to be either an X or O. At the end of 40 trials, participants judged their control on a 100 point scale, labeled "0 = no control", "50 = intermediate control", and "100 = complete control". Although participants had no control over the onset of the screens, estimates of control were high, especially in the 75% reinforcement (M = 43) compared to the 25% reinforcement (M = 11) condition.

For use as a demonstration, participants could be randomly assigned to the high or low reinforcement condition and complete the task of judging their control outside of class time. The analysis is then done with reinforcement level (25% vs. 75%) as the independent between-subjects variable and judgments of control as the dependent variable (unpaired *t*-test). When the results are presented in class, they can provoke the discussion of several issues: Why do illusions of control occur? What are their "real world" implications? Why are the overestimations of control so much lower in the low reinforcement condition?

likely to win. This approach to studying illusory control does not measure control perceptions directly, but relies on preferences for options to infer that participants believe they have control.

A second approach to research on illusions of control has participants work on laboratory tasks where the researcher can set the level of actual control that can be exercised on the task. Typically, participants are given no control over the occurrence of a particular outcome. Then, after working on the task, participants rate the amount of control they believe they had. For example, Alloy and Abramson (1979) used a light-onset task to explore illusions of control. Participants tried to get a light to come on by pressing a button. After each button press, the light either did or did not come on. In actuality there was no relationship between their actions and onset of the light: The light was programmed to come on either on 25% or 75% of the trials. However, when the light came on more frequently (75% of the time), estimates of personal control over onset of the light were high. This work clearly demonstrates that even when people have no control, control judgments can be high.

A third way of researching illusory control asks participants to report on their behavior under various circumstances. For example, McKenna (1993) used the issue of driving safety and asked participants to rate the likelihoods that, compared to other drivers, they would be involved in a road accident when they are driving and when they are passengers. Participants rated the likelihood of an accident to be lower when they were the driver. In a second study, high and low driver-control scenarios for an accident were used. Participants were particularly likely to judge that they could avoid an accident that involved high driver control (e.g., driving your vehicle into the rear of another car) as opposed to low driver control (e.g., being hit from behind). Thus people show illusory control over avoiding an accident by assuming that they will be able to exert control that others cannot.

These three approaches to researching illusory control have strengths and weaknesses as research strategies. The Langer approach has the advantage of using realistic situations that people are likely to face in everyday life (lotteries and competitive games). In addition, the strategy of using an indirect measure of control allows people to express their feeling of control when they may be reluctant to admit that they believe they can control a chance process. At the same time, it has the disadvantage

of not demonstrating whether control per se is the critical factor. The laboratory manipulations of control such as Alloy and Abramson (1979) used employ a dependent measure that is clearly tapping control judgments, but typically the studies do not use tasks with good external validity. The self-report measures used by McKenna (1993) have good external validity but suffer the disadvantages of self-report methodology. One of the strengths of the illusions of control research as a whole is that studies using these diverse methodologies, with their attendant advantages and disadvantages, have reached similar conclusions.

When do illusions of control occur?

Thompson et al. (1998) reviewed five conditions that have been found to influence control judgments: (1) skill-related factors, (2) success or failure emphasis, (3) need or desire for outcome, (4) mood, and (5) the intrusion of reality. More recent research has identified additional conditions that affect illusory control: (6) power and (7) regulatory focus.

1. Skill-related factors

Skill-related factors are attributes associated with situations where skill is an issue, including familiarity, making choices, active engagement with the material, competition, and foreknowledge. According to Langer (1975), when a chance situation contains these elements, people mistakenly think that skill is involved. Hence, they judge that they have some control over the outcomes. For example, the act of choosing options is associated with skilled tasks. Therefore when lottery participants are allowed to choose their own numbers, the game has the feel of a task involving skill and one that is controllable.

Numerous studies have shown that situations with skill-associated features such as familiarity with the materials, personal involvement, competition, and foreknowledge of the possible outcomes lead to overestimations of personal control (see Ayeroff & Abelson, 1976; Dunn & Wilson, 1990; Langer, 1975, for representative studies.) For example, in one study, Ayeroff and Abelson (1976) used an extrasensory perception task to examine illusions of control. Two factors were manipulated – choice of symbols used on the task (participants choose vs. were assigned symbols) and personal involvement (participants shuffled the symbol deck vs. someone else did). Both choice and involvement led to higher estimates of success on the extrasensory perception task, presumably because choice and involvement enhanced illusions of control.

2. Success or failure emphasis

Success or failure emphasis refers to the extent to which the task or the context highlights expectations or perceptions of success vs. failure. An emphasis on success enhances illusions of control, whereas failure emphasis undermines control illusions.

For example, Langer and Roth (1975) showed that a pattern of early successes on a coin toss task led to higher illusions of control than a pattern of early failures, despite the fact that the overall number of wins was constant. The early successes focused participants' attention and expectations on successes, thereby raising control illusions.

Success emphasis is the likely reason why the frequency of reinforcement has a strong effect on illusory control in the Alloy and Abramson (1979) light-onset task. When the light comes on frequently (regardless of what participants do), participants' actions are frequently followed by the desired outcome and they receive a strong message of "success". With infrequent onset of the light, it appears that their actions result in "failure". In their analysis of the mediators of illusory control effects, Thompson et al. (2004) found that a high level of reinforcement was associated with higher estimates that one's attempts to exert control were successful and higher control judgments.

In a further explication of the effects of failure, Langens (2007) showed that only explicit failure undermines illusory control and its effects on mood and persistence. Participants in his study received either explicit or ambiguous failure feedback on an illusory control task. When the failure was ambiguous (i.e., no direct statement of failure from the experimenter) and thus could be interpreted as a possible success, illusory control on the task was associated with more positive mood and greater persistence on a subsequent task. In contrast, when failure feedback was explicit, there was no boost in mood and persistence associated with illusory control. Thus, ambiguous, non-explicit feedback regarding failure may be open to interpretation as a success and not undermine control. Explicit failure feedback, however, makes it difficult to maintain a sense of control.

3. Need or desire for the outcome

Need or desire for the outcome refers to situations where people are motivated to believe that they have personal control. In one test of desire for the outcome, Thompson et al. (2004) manipulated the motivation to have control by paying participants for each success at a computer screen onset task or having no payments for success. Illusions of control were considerably higher when participants were paid for successes and, presumably, motivated to have control over the onset of the screen. In research done by Biner, Angle, Park, Mellinger, and Barber (1995), the motive to have control was not monetary, but the added appeal of food when one is hungry. In their study 1, half the participants were motivated to have control over obtaining a hamburger meal because they fasted from solid food on the day they reported for the study; the other half did not fast. Those who had fasted were significantly more confident that they would win the hamburger meal through participation in a drawing than were those who were not hungry. In later research, Biner and colleagues (Biner, Johnston, Summers, & Chudzynski, 2009) examined the desire to avoid an aversive event, such as giving a speech or immersing one's arm in cold water. Illusory control over the task that determined their fate was

greater when the outcome was expected to be the more aversive of two conditions (e.g., cold vs. warm water). In addition, illusory control can increase in situations that are associated with stress even if performance on the task is not linked to the stress. Friedland, Keinan, and Regev (1992) asked Israeli Air Force cadets to complete an illusions of control measure at a low-stress time or one-half hour before they were tested during a critical training flight (high stress condition). Illusions of control were higher immediately prior to the stressful flight test. Thus illusory control can be prompted by motivations to obtain a desired outcome, to avoid an aversive one, and also to reap the stress-reducing benefits of a sense of control even if the control is not related to the stress at hand.

4. Mood

A number of studies find that illusions of control are higher when people are in a positive mood. For instance, Alloy, Abramson, and Viscusi (1981) manipulated mood states (positive, negative, or neutral) in depressed and nondepressed individuals. Those participants whose mood was temporarily induced to be positive showed higher illusions of control; those with a temporarily more negative mood showed lower illusions of control.

5. Intrusion of reality

Research has found that situations that focus people on a realistic or rational assessment of control may reduce or entirely eliminate illusory control thinking. In one study to test this idea, Bouts and Van Avermaet (1992) had individuals focus on the objective probabilities of winning a gambling game either before or after placing a bet on a card-drawing gamble. Those who considered the probabilities before placing the bet showed considerably lower illusions of control (i.e., made a lower bet).

6. Power

Recent research has focused on the effects of power on illusory control. A series of experiments manipulated power through either assigned roles or priming past experience in a high or low power position (Fast, Gruenfeld, Sivanathan, & Galinsky, 2009). Illusory control was measured by control-indicating judgments such as rolling the dice oneself in a betting situation or perceiving a high likelihood of controlling future events. The research showed that higher power conditions enhanced three correlates of power (optimism, self-esteem, and action orientation). This effect was mediated by higher illusory control in the higher power conditions. Those who are in power most often have more control than those who are subordinate to them. In addition to this, the possession of power can increase illusory control, which further enhances the optimism, self-regard, and tendency to action associated with power.

7. Regulatory focus

Regulatory focus theory distinguishes between a promotion focus, which involves getting desired outcomes, and a prevention focus, which involves avoiding undesirable consequences (Higgins, 1998). Langens (2007) reasoned that promotion-focused individuals should be more prone to illusory control because they have a greater sensitivity to matches between action and desired outcomes. In a series of three studies, he found that dispositional differences in promotion focus, as well as situational factors that enhance a promotion focus, led to greater illusions of control.

Theories of illusory control: Why does it occur?

Langer (1975) offered the earliest theory to explain why people often overestimate their influence even in situations where there is no actual control. According to Langer, illusions of control occur because people confuse skill and chance situations, especially when chance situations contain elements that are usually associated with skill-based tasks. This theory can be used to explain why the presence of skill-based elements such as familiarity, involvement, and competition lead people to overestimate their control on chance-based tasks. However, in their critique of this theory, Thompson et al. (1998) pointed out several flaws with this explanation, including that (1) all situations contain both skill and chance elements, so it seems likely that people are used to sorting out these influences, and (2) this theory cannot explain why non-skill-based elements such as success-focus, need for the outcome, or regulatory focus also influence illusions of control.

Thompson et al. (1998) offered a more comprehensive explanation of illusions of control based on a *control heuristic*, a shortcut that people use to judge the extent of their personal influence. The control heuristic involves two elements: one's intention to achieve the outcome and the perceived connection between one's action and the desired outcome. When one acts with the intention of obtaining a particular outcome and there is a relationship (temporal, common meaning, or predictive) between one's action and the outcome, people judge that they had control over the outcome.

Like most heuristics, this shortcut to estimating control often leads to accurate judgments. For example, when people have the ability to influence whether or not they obtain an outcome, they often act with the intention of getting that outcome and there is a connection between their actions and the receipt of the desired event. However, individuals can also act with the intention of getting a desired outcome and see a connection between their actions and the outcome in situations where they do not have control. For example, gamblers at slot machines may pull the handle with the intention of getting a winning combination. If the winning items appear, there is a temporal connection between the gambler's action and the appearance of the winning items. Thus, using the control heuristic to judge their personal influence can lead gamblers to judge that they have control over getting a winning combination.

In a test of this theory, Thompson et al. (2004) manipulated reinforcement and motive for control in a computer screen onset task. They found that judgments of intentionality mediated the relationship between motives and judgments of control. That is, as would be predicted by the control heuristic theory, the motive to have personal control resulted in higher illusory judgments of control because it affected an element of the control heuristic – intentions to get the outcome. Although judgments of connection were correlated with control judgments, they did not mediate the relationship between motives and illusory control, perhaps because the perceptions of connection were fairly accurate in this situation (i.e., people did not overestimate the number of hits they received).

The Langens (2007) study finding described above that a regulatory promotion focus enhanced illusions of control also supports the control heuristic explanation. Those dispositionally or situationally inclined to a promotion focus attend more closely to the the desired outcomes of their actions, which can heighten their awareness of intending to obtain those outcomes and the contingency between their action and the outcome.

Implications of illusions of control

Most of the research in this area investigates situations that lead to control overestimation rather than the frequency of illusory control, so there is little information about how common illusions of control are. However, they do appear to be fairly easy to elicit in psychological studies (e.g., Alloy & Abramson, 1979; Langer, 1975) which may say something about how often they naturally occur. In addition, the strong effects obtained in McKenna's (1993) research on people's perceptions that they can avoid motor vehicle accidents in "controllable" situations suggests that illusory control is a common phenomenon.

Although illusory control may be common, not everyone overestimates their personal control. For example, moderately depressed individuals tend to have a realistic sense of how much they are contributing to an outcome (Alloy & Abramson, 1982). Does that mean that we are better off if we overestimate our personal control? Overestimating one's control might have a number of consequences, including positive ones (enhanced self-esteem, better motivation for attempting difficult tasks) and negative ones (failure to protect oneself against harm, disappointment when control is disconfirmed, pursuing unrealistic goals, and blaming others for their misfortune). Far less research has focused on this question and the few studies that have been done indicate that both positive and negative consequences can follow from control overestimation.

Positive consequences of control illusions

Some studies have found positive effects of illusory control. Alloy and Clements (1992) used the light-onset task to assess the extent to which college students developed/experienced illusory control. Students who displayed greater illusions of

control had less negative mood after a failure on a lab task, were less likely to become discouraged when they subsequently experienced negative life stressors, and were less likely to get depressed a month later, given the occurrence of a high number of negative life stressors. Thus, individuals who are more susceptible to an illusion of control may be at a decreased risk for depression and discouragement in comparison to those individuals who are not. The idea that "positive illusions" (in this case, illusory control) are associated with adaptive outcomes is consistent with Taylor and Brown's (1988) thesis that positive illusions provide motivation and the confidence to engage in positive action (cf. Chapter 17).

The Fast et al. (2009) research described above that showed that individuals in high power positions have higher illusory control, which in turn augments their self-esteem, optimism, and action orientation, is another example of the benefits that can be derived from overestimating one's control. Illusory control can help foster positive emotions and action orientations that result in a higher probability of success.

Negative consequences of control illusions

In contrast to these positive findings, there is evidence that overestimating one's control has a number of costs and disadvantages in areas as diverse as child care, driving behavior, protective health behaviors, financial investment, and problem gambling. Donovan, Leavitt, and Walsh (1990) investigated the influence of illusory control on performance demands associated with child care. The degree of illusory control was measured by having mothers try to terminate the noncontingent crying of an audio-taped baby. A subsequent simulation assessed the mothers' ability to learn effective responses in ceasing an infant's cry. Mothers with a high illusion of control on the first task showed a depressive attribution style, aversive physiological responses to impending infant cries, and less proactive coping.

As mentioned above, illusions of control are common when people are asked to self-assess their driving behavior (McKenna, 1993). The consequences of illusory control in this area were investigated by Schlehofer et al. (2010). Their study measured general illusions of control and actual ability to drive safely while using a cell phone with a driving simulator. Participants who overestimated their control, in general, and who overestimated their performance on the simulated driving task while using a cell phone were more likely to drive while using a cell phone in everyday life and had poorer driving records. This suggests that an illusory sense that one can compensate for driving distractions is one cause of distracted driving.

Health behaviors are another area of risk where illusory control can reduce effective protection. Reduced self-protection against risk of disease has been linked to illusory control thinking. College students and gay men who had higher scores on general illusory control thinking felt less vulnerable to HIV disease and used less effective protection against HIV (Thompson, Kent, Thomas, & Vrungos, 1999).

Fenton-O'Creevy, Nicholson, Soane, and Willman (2003) studied the illusory control beliefs of financial investment traders. A propensity toward illusory control

was measured with a computer task in which participants attempted to raise the value of an index through key strokes and then rated their control over the index. In fact, the index moved up or down irrespective of participants' actions. Those traders who thought they had more control over the index were rated by their managers as less successful and received a lower annual salary, which was based on their overall track record. Thus, a propensity to overestimate one's control was associated with lower career success. Presumably, the misperception of control leads one to take risks that lead to less favorable outcomes.

Gambling is an area in which the negative consequences of illusory control have received the most research attention. A number of studies found that a tendency toward illusory control thinking and behavior are associated with more involvement in gambling (Lim, Bowden-Jones, & Rogers, 2014; Toneatto, Blitz-Miller, Calderwood, Dragonetti, & Tsanos, 1997). There are a number of characteristics of gambling situations that can help foster an illusion of control, including choice, familiarity, and desirability of the outcome. In addition, it has been suggested that the gambling industry helps enhance the perception of control though advertising slogans ("You can't win if you don't play"); the design of games that incorporate more choices, such as multi-line slot machine games that allow gamblers to choose the number of paylines (Harrigan, MacLaren, Brown, Dixon, & Livingstone, 2014); and video-lottery terminal-stopping devices (Ladouceur & Sevigny, 2005). These designs give the illusion of control by emphasizing winning and offering choices, but do not decrease the probability of loss.

Recent research has used magnetoencephalography to compare the neural patterns of problem and non-problem gamblers while making decisions in gambling situations that do (e.g., poker) or do not (e.g., roulette) offer some degree of control (Hudgens-Haney et al., 2013). Non-problem gamblers showed decreased visual and prefrontal cortex processing in the no-control situations, whereas problem gamblers did not. It appears that problem gamblers do not make a distinction between gambling situations with and without control, presumably because of their greater propensity for illusory control.

Two promising interventions for reducing cognitive illusions that encourage problem gambling were tested by Larimer et al. (2012). One of the interventions, a cognitive behavioral approach that challenged cognitive illusions in problem gamblers, was found to reduce illusions of control and gambling consequences. Thus, there is evidence that gamblers' illusions of control can be modified with intensive work.

In a final area related to negative consequences of control overestimation, Mortiz, Thompson, and Andreou (2014) examined illusory control in patients with schizophrenia and in a nonclinical control group using a task similar to the light-onset paradigm of Alloy and Abramson (1979). Both groups showed an illusion of control, but illusory control was higher in the clinical group with positive symptoms, particularly those who reported experiencing hallucinations. It is not clear, though, whether the higher levels of illusory control contributed to the symptoms or were merely a side effect of a pathological process such as the propensity to see patterns in random events.

Reconciling positive and negative effects

Which is the correct view: that illusory thinking is generally useful because it leads to positive emotions and motivates people to try challenging tasks or that people are better off if they have an accurate assessment of themselves and their situation? Another possibility is that sometimes illusory control is adaptive and at other times it is not. For example, illusions of control may be reassuring in stressful situations, and can contribute to confidence and positive self-views, especially in power positions. This boost to motivation and confidence can help maintain positive moods and protect against bouts of depression, but some of this confidence and motivation may come at a cost in other areas of life. It is clear that illusory control can lead people to take unnecessary risks with their health, to make overly risky financial decisions, and to engage in problem gambling with financial and social costs.

It is not known whether some people are able to select their areas of control estimation such that they receive the benefits, but avoid the costs, of illusory control. The challenge for researchers is to examine how individuals manage their use of illusory control across a variety of circumstances to find optimal effects.

Neural research on illusory control

Neuroscience offers a promising area for further understanding of illusory control. For example, Kool, Getz, and Botvinick (2013) examined the neural correlates of control overestimation, using a task in which participants made a choice of gambling "wheels" and were either allowed to keep their chosen wheel or forced to use a non-chosen one. Their first study confirmed that the task aroused illusions of control: Estimated reward probabilities were higher for the chosen wheel. In the second study, brain imagery was used to examine the correlates of illusory control. It was expected that a reward prediction error (a stronger neural reaction to rewards when they are unexpected) would negatively correlate with illusory control. That is, illusory control would lead to weaker reward prediction errors because illusory control implies that rewards are expected. Contrary to prediction, prediction error rates were not related to the choice manipulation. Instead, two unexpected, but interesting, results were found. One result was activation in response to having a choice, regardless of reward outcome. A second result was a heightened response to the revealing of outcomes when allowed to keep the chosen wheel. This outcome saliency effect was "unsigned", that is, it occurred for *both* wins and losses.

From these two results, the authors suggest two routes for illusory control. One is that illusions of control could arise from an intrinsic value placed on having a choice, which leads to an optimistic assessment of future events. The second is that personal choice leads to outcome saliency for the results, which makes the outcome more self-relevant, thereby heightening illusory control.

These are intriguing ideas, but given what we know from past research on illusions of control, the "intrinsic value" explanation cannot explain all instances of illusory control. For example, in the "light onset" paradigm used by many researchers, reinforcement rates are varied while choice is kept constant. Strong illusory control, related to reinforcement rate is found in these studies (e.g., Alloy & Abramson, 1979) which cannot be explained by the intrinsic value of choice.

The "outcome saliency" explanation is more promising because it fits nicely with the control heuristic idea that illusory control is based on both intention and connection. Outcome saliency for outcomes on choice trials could increase the perception of connection between one's actions and the outcome. Further neural imaging work that examines the effects of intended vs. unintended outcomes would be valuable to confirm the validity of these explanations.

Summary

- Illusions of control occur when individuals overestimate their personal influence over an outcome.
- Three approaches to researching illusions of control are preference analysis, experimental laboratory studies that directly measure control perceptions, and self-reports of control-related behaviors.
- Illusions of control are affected by skill-related factors, success or failure emphasis, need or desire for the outcome, mood, intrusion of reality, power, and regulatory focus.
- Langer (1975) originally proposed that illusions of control occur because people confuse skill and chance situations. Thompson et al. (1998) have presented a more comprehensive explanation of illusory control based on a control heuristic that can account for more of the findings.
- According to the control heuristic explanation, people use both connection and intention to judge their control. Because both can be present even when control is not, personal control is often overestimated.
- Depending on the circumstances, illusions of control can be positive or negative in everyday life.

Further reading

Langer's (1975) classic studies of illusory control in the selection of lottery tickets and games of chance are a good place to start. For a different and equally influential approach to control overestimation, the original Alloy and Abramson (1979) set of studies provides a systematic exploration of this topic and the beginnings of the depressive realism concept. For a comprehensive review of illusions of control research and the control heuristic explanation, see Thompson et al. (1998). A condensed review is also available (Thompson, 1999).

References

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General, 108*, 441–485.

Alloy, L. B., & Abramson, L. Y. (1982). Learned helplessness, depression, and the illusion of control. *Journal of Personality and Social Psychology, 42*, 1114–1126.

Alloy, L. B., Abramson, L. Y., & Viscusi, D. (1981). Induced mood and the illusions of control. *Journal of Personality and Social Psychology, 41*, 1129–1140.

Alloy, L. B., & Clements, C. M. (1992). Illusion of control: Invulnerability to negative affect and depressive symptoms after laboratory and natural stressors. *Journal of Abnormal Psychology, 101*, 234–245.

Ayeroff, F., & Abelson, R. P. (1976). ESP and ESB: Belief in personal success at mental telepathy. *Journal of Personality and Social Psychology*, *34,* 240–247.

Biner, P. M., Angle, S. T., Park, J. H., Mellinger, A. E., & Barber, B. C. (1995). Need and the illusion of control. *Personality and Social Psychology Bulletin, 21,* 899–907.

Biner, P. M., Johnston, B. C., Summers, A. D., & Chudzynski, E. N. (2009). Illusory control as a function of the motivation to avoid randomly determined aversive outcomes. *Motivation and Emotion, 33*, 32–41.

Bouts, P., & Van Avermaet, E. (1992). Drawing familiar or unfamiliar cards: Stimulus familiarity, chance orientation, and the illusion of control. *Personality and Social Psychology Bulletin*, *18,* 331–335.

Donovan, W. L., Leavitt, L. A., & Walsh, R. O. (1990). Maternal self-efficacy: Illusory control and its effect on susceptibility to learned helplessness. *Child Development, 61*, 1638–1647.

Dunn, D. S., & Wilson, T. D. (1990). When the stakes are high: A limit to the illusion-of-control effect. *Social Cognition, 8,* 305–323.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Brighton: Erlbaum.

Fast, N. J., Gruenfeld, D. H., Sivanathan, N., & Galinsky, A. D. (2009). Illusory control: A generative force behind power's far-reaching effects. *Psychological Science, 20,* 502–508.

Fenton-O'Creevy, M., Nicholson, N., Soane, E., & Willman, P. (2003). Trading on illusions: Unrealistic perceptions of control and trading performance. *Journal of Occupational and Organizational Psychology, 76*, 53–68.

Friedland, N., Keinan, G., & Regev, Y. (1992). Controlling the uncontrollable: Effects of stress on perceptions of controllability. *Journal of Personality and Social Psychology, 63*, 923–931.

Harrigan, K., MacLaren, V., Brown, D., Dixon, M. J., & Livingstone, C. (2014). Games of chance or masters of illusion: Multiline slots design may promote cognitive distortions. *International Gambling Studies, 14*, 301–317.

Higgins, E. T. (1998). Promotion and prevention: Regulatory focus as a motivational principle. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 30, pp. 1–46). New York: Academic Press.

Hudgens-Haney, M. E., Hamm, J. P., Goodie, A. S., Krusemark, E. A., McDowell, J. E., & Clementz, B. A. (2013). Neural correlates of the impact of control on decision making in pathological gambling. *Biological Psychology, 92*, 365–372.

Keller, J., & Bless, H. (2008). Predicting future affective states: How ease-of-retrieval and faith in intuition moderate the impact of activated content. *European Journal of Social Psychology, 38*, 1–10.

Kool, W., Getz, S. J., & Botvinick, M. M. (2013). Neural representation of reward probability: Evidence from the illusion of control. *Journal of Cognitive Neuroscience, 25*, 852–861.

Ladouceur, R., & Sévigny, S. (2005). Structural characteristics of video lotteries: Effects of a stopping device on illusion of control and gambling persistence. *Journal of Gambling Studies, 21*, 117–131.

Langens, T. A. (2007). Emotional and motivational reactions to failure: The role of illusions of control and explicitness of feedback. *Motivation and Emotion, 31*, 105–114.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32,* 311–328.

Langer, E. J., & Roth, J. (1975). Heads I win, tails it's chance: The illusion of control as a function of the sequence of outcomes in a purely chance task. *Journal of Personality and Social Psychology, 32*, 951–955.

Larimer, M. E., Neighbors, C., Lostutter, T. W., Whiteside, U., Cronce, J. M., Kaysen, D., & Walker, D. D. (2012). Brief motivational feedback and cognitive behavioral interventions for prevention of disordered gambling: A randomized clinical trial. *Addiction, 107*, 1148–1158.

Lim, M. S. M., Bowden-Jones, H., & Rogers, R. D. (2014). Expressing gambling-related cognitive biases in motor behaviour: Rolling dice to win prizes. *Journal of Gambling Studies, 30*, 625–637.

McKenna, F. P. (1993). It won't happen to me: Unrealistic optimism or illusion of control? *British Journal of Psychology, 84*, 39–50.

Moritz, S., Thompson, S. C., & Andreou, C. (2014). Illusory control in schizophrenia. *Journal of Experimental Psychopathology, 5*, 113–122.

Rodin, J. (1986). Aging and health: Effects of the sense of control. *Science*, *233*, 1271–1276.

Schlehofer, M. M., Thompson, S. C., Ting, S., Ostermann, S., Nierman, A., & Skenderian, J. (2010). Psychological predictors of college students' cell phone use while driving. *Accident Analysis and Prevention, 42*, 1107–1112.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*, 193–210.

Thompson, S. C. (1999). Illusions of control: How we overestimate our personal influence. *Current Directions in Psychological Science, 8*, 187–190.

Thompson, S. C., Armstrong, W., & Thomas, C. (1998). Illusions of control, underestimations, and accuracy: A control heuristic explanation. *Psychological Bulletin*, *123*, 143–161.

Thompson, S. C., Kent, D. R., Thomas, C., & Vrungos, S. (1999). Real and illusory control over exposure to HIV in college students and gay men. *Journal of Applied Social Psychology, 29*, 1128–1150.

Thompson, S. C., Kyle, D., Osgood, A., Quist, R. M., Phillips, D. J., & McClure, M. (2004). Illusory control and motives for control: The role of connection and intentionality. *Motivation and Emotion, 28*, 315–330.

Thompson, S. C., & Spacapan, S. (1991). Perceptions of control in vulnerable populations. *Journal of Social Issues*, *47*, 1–21.

Toneatto, T., Blitz-Miller, T., Calderwood, K., Dragonetti, R., & Tsanos, A. (1997). Cognitive distortions in heavy gambling. *Journal of Gambling Studies*, *13*, 253–265.

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review*, *66*, 297–333.

8

WASON SELECTION TASK

Jonathan St. B. T. Evans

Peter Wason is regarded by many as the founder of the modern psychology of reasoning. He was an unorthodox but highly creative researcher who invented several reasoning tasks, in particular, the selection task first published in 1966, on which this chapter is focussed, and the 2–4–6 task first published in 1960 (see Wason & Johnson-Laird, 1972, for Wason's early interpretation of these tasks). Work continues on both tasks to the current day and although his last paper was published more than 20 years ago, citations of Wason's work are still increasing (Evans, in press). In the context of this book, it is also worth noting that Peter Wason was an early leader in emphasising bias and irrationality (as he saw it) in the psychology of reasoning. To understand the origin of this, I will briefly consider his work on the 2–4–6 problem before focusing on the selection task (for more detailed coverage, see my recent review of both Wason's early work on the problem and the later development of the task within the reasoning literature; Evans, in press).

Wason's early thinking and the 2–4–6 problem

Wason's thinking about bias and irrationality was influenced by a combination of a deep interest in Freudian theory and the observations of his participants on his fiendish reasoning tasks. There were no computers to run experiments in those days, and Wason neither employed research assistants nor handed out booklets. He ran experiments himself, one to one, carefully watching his 'subjects', as they were then called. His papers contain direct observations about their behaviour, not just statistical analyses of their responding.

The 2–4–6 task was designed as a test of induction and rule learning. Participants were told that the experimenter had in mind a rule that classified triples of three whole numbers. An example that conformed to the rule was 2–4–6. Their

task was to discover the rule by generating triples of their own and to receive feedback: The triple did or did not conform. They were required to record their triple together with the current hypothesis and the feedback given. When they were sure, they could announce the rule. If they were wrong, they were invited to continue.

What made the task difficult was the fact that the actual rule was *any ascending sequence*. Hence, the 2–4–6 example was deliberately biased and induced participants to formulate more specific hypotheses such as 'ascending in equal intervals'. Wason found that they typically tested positive examples of their hypothesis (e.g., 3–6–9 or 10–20–30) and, of course, in each case the experimenter gave positive feedback – it conformed to the rule. In the original experiment, only six out of 29 participants solved the problem at the first attempt, and eight were unable to solve it at all, even given several chances to continue testing triples. Wason believed that participants had a verification bias, later better known as *confirmation bias* (see Chapter 5). However, he may have been wrong in this interpretation, as many authors now believe that the 2–4–6 problem shows instead a *positive-testing bias* (Evans, in press; Klayman & Ha, 1987; Poletiek, 2001). That is, people strongly prefer to test positive examples of their hypotheses, but this does not in general lead to confirmation of hypotheses. It does on Wason's task with its particular design, but it is certainly not clear that people are motivated by an attempt to verify their hypotheses on this task.

Equally important to Wason, however, was the observation that participants often persisted with the same hypothesis after being told that it was wrong! They did this by reformulating it in different words. For example, one participant persisted in testing triples such as 2–6–10, 1–50–99 and 3–10–17. On being told that the announcement 'the rule is that the difference between two numbers next to each other is the same' was wrong, she then announced instead that it was 'the rule is adding a number, always the same, to form the next number'. There were many such examples in Wason's protocols, which he found perhaps of more interest than the quantitative data. He described the behaviour as fixated and obsessive, viewing the verbal formulations as rationalisations of an unconscious bias. These ideas were later to link up with his observations on the selection task and underlie the first published formulation of a dual-process theory in the psychology of reasoning (Wason & Evans, 1975), discussed below.

The abstract Wason selection task

One of the best known tasks in the study of deductive reasoning is the Wason selection task, or four-card problem (see also Chapter 5). Invented by Peter Wason in the 1960s, the task became well known after a series of studies was described in the early textbook on reasoning published by Wason and Johnson-Laird (1972). Studies of this task are generally divided between those using abstract problem materials and those using concrete or thematic material. A typical abstract version of the task is the following.

There are four cards lying on a table. Each has a capital letter on one side and a single digit on the other side. The visible sides of the cards are as follows:

A D 3 7

The following statement applies to these four cards and may be true or false:

> If there is an A on one side of the card, then there is a 3 on the other side of the card.

Your task is to decide which cards, and only which cards, would need to be turned over in order to check whether the rule is true or false.

Most people give the answer A alone, or A and 3. Neither is logically correct according to the analysis given by Peter Wason and accepted by most later authors in the field. Logically, the statement can only be false if there is a card with an A on one side and without a 3 on the other. For example, if the A is turned over and a 5 is on the back, we know the statement is false. Because turning the A card *could* discover such a case, it is logically necessary to turn it. But, by the same argument, the 7 card must be turned as well. 7 is a number that is not a 3, and discovering an A on the back would similarly disprove the statement. Very few people select this card, however. What they often do instead is to choose the 3 card, which is *not* logically necessary. The statement says that As must have 3s on the back, but it does not say that 3s must have As on the back. So, whether you turn over the 3 and find an A or find a B, it will be consistent with the statement either way. In fact, you cannot prove the statement true except by eliminating any possibility that would make it false.

Why do people make these logical errors on the Wason selection task? Wason and Johnson-Laird (1972) again suggested that people have a *confirmation bias*. Their idea was that people tend to look for information that will confirm their hypotheses rather than information that could refute or falsify them. Such a bias could be important in science, since scientists generally agree that they should try to disprove theories in order to test them thoroughly. So how might confirmation bias explain the selection task findings? Wason suggested that people think that the statement would be true if a card were found with an A and a 3 on it. Because they have a confirmation bias, they turn over the A and the 3 cards trying to find this confirming case. They overlook the 7 card because they are not focussed on trying to find the disconfirming card that has an A and *not* a 3.

Matching bias

Whilst plausible, the confirmation bias account was abandoned by Wason and others shortly after the publication of his 1972 book with Johnson-Laird. The reason was an experiment reported by Evans and Lynch (1973) that provided strong evidence for an alternative account, known as *matching bias*. Note that the cards people

tend to choose on the standard task, A and 3, are those that are the values explicitly named in the conditional statement (If there is an A, then there is a 3). What if people are simply matching their card choices to these named values? How could we tell if they were doing this, rather than looking for confirmation as Wason suggested? The answer requires a change to the presentation of the task. Suppose we introduce a negative into the conditional statement as follows:

> If there is an A on one side of the card, then there is NOT a 3 on the other side of the card.

The instructions are the same as before as are the cards displayed. Now what will people choose? If they have a confirmation bias, they should choose the A and the 7 cards, in order to discover a card that has an A on one side and does not have a 3 on the other. If they have a matching bias, on the other hand, they should choose A and 3 in order to match the cards to the named items. Note that this is now the logically correct answer, as an A3 card would disprove the statement. The results of the Evans and Lynch study were decisively in favour of matching bias. In fact, once the effects of matching were controlled, there was no evidence of confirmation bias at all in their study. The effect has been replicated many times in subsequent studies using a variety of tasks and linguistic formats (Evans, 1998).

Many researchers in the field were quite disconcerted by this finding when it appeared. Matching bias seemed to make participants in these experiments look rather foolish. How could they ignore the logical reasoning instructions and make such a superficial response? Nevertheless, there is strong evidence to suggest that people think mostly about the matching cards. If people are asked, in a computer presentation of the task, to point with a mouse to cards they are thinking of choosing, for example, most point little if at all at the 7 card. It is as though the matching bias acts as a kind of preconscious filter, drawing people's attention to the A and 3 cards. (Of course, the actual letters and numbers given varies for different participants.) The same effect has been shown with the improved methodology of eye-movement tracking (Ball, Lucas, Miles, & Gale, 2003). People spend more time looking at matching cards and, more generally, the cards they end up choosing.

When people are asked to provide verbal justifications on the selection task, it becomes apparent that they are engaged in reasoning and that they do think about the hidden sides of the cards. But once again they focus their attention on the *matching values* that might be on these hidden sides (Lucas & Ball, 2005; Wason & Evans, 1975). With the affirmative conditional – if A, then 3 – for example, they might well say that they are turning over the A card because a 3 on the back would prove the statement true. With the negative statement – if A, then not 3 – they say they need to turn the A card because a 3 on the back would prove the statement *false*. In either case, they think only about the matching cards and end up finding a justification for choosing them.

It appears that matching bias is strongly linked to problems in understanding implicit negation. Suppose we give people a similar scenario to that introducing the

selection task above. There are cards with letters on one side and numbers on the other, and once more the following rule applies:

If the letter is a not an A, then the number is a 3.

If we tell people that a card has a letter which is not an A and ask them what follows, then everyone says that the number must be a 3. Obvious. Suppose, however, that we tell people instead that the card has the letter D. Now what follows? A surprisingly large number of people say that nothing follows (Evans & Handley, 1999). Of course since there is only one letter on the card and it is a D, then it cannot be an A – and if it is not an A, then the number must be 3. The difficulty here is that D is an implicit negation of an A, as compared with the explicit negation 'not an A'.

Returning now to the standard selection task, we can see that the cards that do not match the items in the statement are implicit negations of them. The key logical error on the selection task is failing to choose the 7 card (or its equivalent for the actual statement presented). Recall that the reason this card is needed is to check that there is no card that has an A on one side but does not have a 3 on the other. So the 7 card has to represent, by implicit negation, 'not 3'. What happens if the task is reformulated so that this card is described using explicit negation as a card which has a number that is not 3? If implicit negation is the cause of the matching bias then it should no longer operate when negations are described explicitly in this way. The relevant experiments (discussed by Evans, 1998) have been run and the results again are very clear. Use of explicit negation *removes* the matching bias effect.

There is another explanation of matching bias favoured by some theorists in the area (Oaksford & Chater, 1998). The argument here is that people are prone to choose information that is generally informative in everyday life. Negative information is generally less informative than positive. Suppose I tell you that there is a letter on the card and the letter is a D. This is highly informative because I have eliminated 25 other possibilities of what the letter might be. Suppose instead I tell you that the letter is *not* D. This is much less informative because I have only eliminated one possibility and 25 still remain. For this reason, Oaksford and Chater suggested that people are generally biased towards seeking positive information. Evidence for a general *positivity bias* in thinking and hypothesis testing is quite widespread (Evans, 1989) and other theorists also have argued such a bias reflects a process that would normally be adaptive in everyday life (Klayman & Ha, 1987). These accounts tend to portray findings of error with laboratory reasoning problems, such as the selection task, as cognitive illusions that are unrepresentative of everyday reasoning.

Dual-process theory

Dual-process theories of reasoning and decision making have become increasingly popular in cognitive psychology and are often used to interpret evidence about cognitive biases in reasoning and decision making (Evans, 2007, 2008; Kahneman, 2011; Stanovich, 2011). Biases are typically attributed to Type 1 processes, which operate

in a fast and autonomous manner. Type 2 processing is generally considered to be slower, loading on working memory and allows abstract and rule based reasoning. In the kinds of tasks that reasoning and decision researchers study, such Type 2 thinking is often required to solve the problems and avoid biases. For this reason, solution rates are also reliably correlated with cognitive ability (Stanovich, 2011).

In the psychology of reasoning, dual-process theory first arose from Wason's early research. It has been mentioned that he was a Freudian and interested in unconscious thinking, fixation, rationalisation and so on. He first developed these ideas in his work on the 2–4–6 problem, but their explication as dual-process theory came in the collaboration with myself mentioned above (Wason & Evans, 1975). We found that people usually chose matching cards both on the standard task and on a negated rule where it leads to the correct answer. However, they typically justified the former in terms of verifying the conditional statement and the latter in terms of falsifying it. Hence, we suggested that matching bias reflected an unconscious Type 1 process and that conscious Type 2 thinking was used to provide a rationalisation.

Of course, dual-process theory has been developed much since this paper, and on other tasks Type 2 thinking is often thought to be responsible for a rational choice rather than simply a means of rationalising biased responses. Actually, Type 2 thinking may do both of these things according to a popular version of dual-process theory in which Type 1 processes provide default intuitive answers, which may be changed by the intervention of higher effort Type 2 reasoning. The general view is that people do not intervene as much as they should and are by nature 'cognitive misers' (Kahneman, 2011; Stanovich, 2011). Hence, people may routinely seek to justify an intuitive answer but do so in a cursory manner and are too easily satisfied with it. Recent research has shown that some intuitions come to mind easily and have a strong *feeling of rightness* associated with them (Thompson, Prowse Turner, & Pennycook, 2011). In such cases, people spend little time thinking about their answers. A recent paper showed that matching bias has precisely this quality (Thompson, Evans, & Campbell, 2013). That is, matching cards *feel* right and are often chosen quickly as a result.

The realistic selection task

Do realistic materials facilitate reasoning?

Psychologists use the rather ugly word 'debias' to refer to factors that remove cognitive biases. Experiments described by Wason and Johnson-Laird (1972) led to a popular hypothesis (now seen as greatly oversimplified) that realistic problem materials facilitate reasoning performance. This stands in contrast with many claims that prior beliefs are a major cause of bias in reasoning (see Chapter 9). Nevertheless, this hypothesis was the starting point for a very productive set of experiments on the effects of content and context on the selection task.

Let us start by examining a version of the Wason selection task that is known to be very easy: the 'drinking age problem' first reported by Griggs and Cox (1982).

> Imagine you are a police officer observing people drinking in a bar. You need to check that they are obeying the following rule:
>
> If a person is drinking beer, then that person must be over 18 years of age.
>
> There are four cards, each representing an individual drinking in the bar. One side shows what beverage they are drinking and the other side shows their age. The four exposed sides of the cards are as follows:
>
> | | | | |
> |---|---|---|---|
> | Drinking beer | Drinking coke | 22 years of age | 16 years of age |
>
> Which cards would you need to turn over in order to find out whether or not the rule is being obeyed?

These cards are laid out in the same logical order as for the abstract selection task discussed earlier. Hence, the first and last cards are again the correct choices. You should check the person drinking beer and the person who is 16 years of age. The great majority of participants do precisely that. They get this problem right and they show no evidence of matching bias.

What is the difference between this problem and the original selection task? Actually, there are several. The problem is 'realistic', using familiar content that people can relate to rather than abstract content. It also has a context – the police-officer scenario. Brief though it is, this context is critical to the facilitation. If the task is presented without an introductory context, performance is little better than on the abstract version. The logic of the problem is also subtly changed. The standard task asks you to decide whether the rule is true or false. In the drinking age problem, you have to decide whether or not the rule is obeyed. Technically, this changes it from an indicative to a *deontic* conditional which uses a different form of logic (Manktelow & Over, 1991). Most realistic forms of the selection task that reliably facilitate correct choices have this deontic form, but this is not enough in itself – for example, when abstract materials are used with the request to check if a rule has been obeyed.

You might think that the drinking age problem facilitates because people have direct real-world knowledge of drinking age laws and simply know from experience that underage drinkers are the ones to worry about. However, this is not the correct explanation. The conditional statement in the drinking age problem is a permission rule. You need to fulfil a condition (over 18) in order to have permission to do something (drink beer). Other problems with permission rules work equally well, even where people have no direct experience of these rules.

An example is Condition A in the classroom demonstration shown in Text box 8.1. This was adapted from an experiment reported by Manktelow and Over (1991). The italicised text provides the cue that the store may be cheating the customers, which would occur if they spent over £100 but did not receive a free gift.

TEXT BOX 8.1 CLASSROOM DEMONSTRATION OF THE SELECTION TASK AS DECISION MAKING

Design and participants

This is a between-participant design with two groups. I would suggest a minimum of 20 participants in each group.

Materials and procedure

Two sets of materials should be prepared printed on a single sheet of paper. These should be randomly intermixed and then given out to the students face down. On instruction, all participants should then turn over the paper and work on the problem at the same time. In this way, they are unlikely to notice the subtle difference between the two conditions if their neighbours have the other version.

Condition A materials

You are a company manager. Your firm is trying to increase business by offering free gifts to people who spend money in its shops. The firm's offer is:

If you spend more than £100, then you may take a free gift.

You have been brought in because you have been told that in one of the firm's shops the offer has run into a problem: *You suspect that the store has not given some customers what they were entitled to*.

You have four receipts in front of you showing on one side how much a customer spent and on the other whether they took a gift. The exposed sides show:

Spent £120 Spent £85 Took a gift Did not take a gift

Which of the receipts must you turn over to see whether the store's rule has been followed?

Condition B materials

These are exactly the same as in Condition A except that the italicised words are replaced with *'You suspect some customers in the store might have taken more than they were entitled to'*.

Statistical analysis

The expectation is that the frequency of choice of the four cases will be different between the two groups (see main text). The simplest way to analyse this is with a series of 2 × 2 chi-squared tests for each case. For example, we might find that in Group A 18 people select 'Spent £120' and 6 do not, but in Group B 10 people select this receipt and 14 do not. The chi-squared test will show us whether this difference is significant. Repeat the test for the other three cases.

This is not a rule with which most people have direct experience, but Manktelow and Over showed that people will examine Receipts 1 and 4, which could reveal such a case of cheating and correspond to the logically correct selections. What is particularly interesting, however, is Condition B in the demonstration. Here, the italicised text is changed so that the participant is alerted to the possibility that the customers are cheating the store. This could happen if they spent less than £120 but took the free gift. With this perspective, we expect a shift towards choosing Receipts 2 and 3, which could reveal this form of cheating, again reported by Manktelow and Over in a similar experiment. This study was followed by others (for example, Gigerenzer & Hug, 1992) that also presented scenarios which altered the perspective of the decision maker and changed participants' choices accordingly.

These findings are very important because they show that realistic materials are *not* facilitating logical reasoning. When the perspective is shifted, people tend to choose the two cards that are not normally considered the logical choices. They nevertheless appear to make appropriate choices given the goals that the context elicits. This argument was supported by Oaksford and Chater (1994), who extended it to the abstract form of the selection task. In the latter case, they made a technical argument that the normal choices on the standard task, considered by Wason and most other authors to be logically incorrect, are appropriate on the basis of expected information gain. This paper was influential but also controversial, with several objections being published (e.g., Laming, 1996).

Theories developed with the realistic selection task

Realistic versions of the selection task were heavily investigated during the 1980s and 1990s, in particular, leading to several bold and novel theoretical claims. I will consider some of these here. The first, proposed by Cheng and Holyoak (1985) was that people reason on this version of the task using *pragmatic reasoning schemas*. The idea is that people have learned to reason in certain contexts such as those involving permission and obligation rules. While context-dependent, such schemas are abstract within the context, containing a rule such as:

If you wish to perform action A, then you must have permission P.

Such rules would be instantiated for particular contexts, for example, A is drinking beer and P is over 18 years of age. Cheng and Holyoak suggested that such schemas contain production rules such as:

> If you do not have permission P, then you should not perform action A.

With this retrieved rule, people can see immediately that they must check underage drinkers to make sure they are not drinking beer. Thus, they can solve the task without need for any difficult reasoning. The abstract task does not prompt retrieval of such a schema and so remains very difficult.

Perhaps the most radical and surprising theory was that of Cosmides (1989), in the most-cited paper on the Wason selection task. She brought an entirely new perspective to the psychology of reasoning, that of evolutionary psychology. Her view was that the mind consisted of a number of special-purpose and domain-specific cognitive modules without a general facility for logical reasoning (Cosmides & Tooby, 1994). In the case of the selection task, she proposed that a social-contract module would have evolved with a built-in facility for cheater detection. Contexts such as that of the drinking age rule would facilitate correct choices because they are social contracts: A cheater would be an underage drinker in this case. The theory can also explain perspective shifts, as in the Manktelow and Over (1991) study. The context signals in one case that the cheater may be the store failing to deliver the free gift when obliged to do so, and in the other case the customer taking a free gift to which he or she is not entitled. The cheater-detection algorithm then leads to the choices observed in both cases. In a later paper, it was argued that other facilitation effects were due to a hazard-avoidance module (Fiddick, Cosmides, & Tooby, 2000).

The idea of cognitive modules was first proposed by Fodor (1983) but in conjunction *with* a general reasoning system. Fodor was actually highly critical of the 'massive modularity' approach and of Cosmides' theory of the selection task (Fodor, 2000). In fact, despite its fame, the Cosmides paper was heavily criticised, especially by researchers working in the psychology of reasoning (Manktelow & Over, 1991; Over, 2003; Sperber & Girotto, 2002). The general view was that more parsimonious explanations could account for the data, such as the decision-theoretic approach already mentioned, or the theory of pragmatic relevance (Sperber, Cara, & Girotto, 1995). Cosmides also used standard abstract selection tasks as controls, despite the problem discussed here that they involve indicative and not deontic conditionals.

Bias and rationality in human reasoning

Research on deductive reasoning was developed during a period in which psychologists were content to follow many philosophers in regarding logic as a model for rational thinking (Evans, 2002). Although the Wason selection task involves hypothesis testing as well as deduction, it is generally regarded as a key paradigm in this literature and – among other things – a test of logical reasoning. If logic is indeed the basis for rational thinking, then results on the selection task and on many

other deductive problems should make us very concerned. Although error rates are exceptionally high on the selection task, other typical methods used to study reasoning, including syllogistic and conditional inference tasks, also show evidence of frequent logical errors and biases (see Chapter 9). As shown here, introducing realistic content does not necessarily induce better logical reasoning as once thought, although it often changes the answers given. People's reasoning is highly influenced by content and context that are logically irrelevant to the task set, and the knowledge and belief introduced is just as likely to bias as to debias responses from a strictly logical point of view.

Faced with these findings psychologists (and to a lesser extent philosophers) have felt it necessary to resolve what Evans and Over (1996) termed the 'paradox of rationality'. The human species is highly successful and has succeeded in adapting the environment to its own needs, inventing science, technology and so on. We seem to be a very intelligent species. So why are representatives of the human race generally so poor at solving reasoning tasks set in the psychological laboratory? Discussions of rationality in reasoning experiments have turned on three major issues first observed by Cohen (1981). The first of these is the *normative-system* problem. Perhaps people seem illogical because formal logic provides a poor framework for assessing the rationality of everyday reasoning. Psychologists have in the past been somewhat naïve in adopting standard textbook logic as a normative reference and have lacked awareness that such systems have been rejected by contemporary philosophical logicians precisely because they cannot be mapped on to natural language and everyday reasoning (Evans, 2002). The majority still favours use of normative systems, however, even if based on alternatives to the standard, traditional norms (see Elqayam & Evans, 2011).

As an example, many psychologists have treated the conditional statement 'If p, then q' as though it represented a relationship of material implication between p and q, as described in elementary textbooks on formal logic. Such a representation means that the conditional is true unless we have a case of p and not-q and, hence, is logically equivalent to the statement 'Either not-p or q'. Suppose I make this statement:

If I am in London, then I am in France.

If I am in Plymouth when I make this statement and the conditional is material, then you would have to say the statement is true. A material conditional is equivalent to:

Either I am not in London or I am in France.

Since I am in Plymouth, and therefore not in London, the first part of the disjunction is confirmed, so the statement is true. However, it is self-evident that the conditional statement is false. Many philosophers and psychologists consequently reject the material conditional. What we actually do is to imagine the world in which I am in London and ask whether I would be in France. Evidently we would

be in England, so the statement is clearly false as defined by an alternative *suppositional* theory of the conditional (see Evans & Over, 2004). This could not be used to explain away the typical choices on the selection task, however, as the same correct choices also follow from the suppositional or probabilistic conditional, which is now widely adopted in the psychology of reasoning. All theories of the conditional statement 'if p, then q' preclude the case of p and not-q.

A second issue is known as the *interpretation problem*. Perhaps participants construe the task differently than the experimenter intended. Consider again the problems presented in the classroom presentation (Text box 8.1). In Condition B we could say that participants are making a logical error if they choose to investigate those who spent less than £100 and those who took a gift. But interpreted as a decision-making task with the context cued, we can easily argue that these are the correct choices to make.

The third issue is the *external-validity* problem. This is the argument that many of the reasoning problems used in the laboratory are artificial and unrepresentative of real-world reasoning. Cohen (1981) described the selection task as a 'cognitive illusion' of little importance. But Cohen wanted to argue that everyone is rational and was having trouble explaining the task by his other arguments! There has, however, been considerable debate as to whether Wason's 2–4–6 problem provides a valid test of scientific thinking as originally claimed (Evans, 2007; Poletiek, 2001). The selection task has also been debated in this context. Whilst used as a centre piece for testing theories of reasoning, as seen above, it can also be regarded as an unfortunate choice for this purpose. The task itself arguably involves very little reasoning, except perhaps for the small minority of high-IQ people who manage to solve it. If, say, most people just focus on matching cards and often convince themselves to accept this, what does this have to do with logical reasoning? The standard task is just too difficult to provide much variance of correct responding in all but the most elite populations.

The force of these three arguments, taken together, makes it difficult to argue that the biases observed in the Wason selection task and other reasoning experiments are necessarily indicative of irrationality in human beings. Moreover, there is a deeper question as to whether normative assessment is all that relevant to the psychology of reasoning and decision making, anyway. Elqayam and Evans (2011) have argued that psychologists should be trying to describe reasoning and not to judge it, and that emphasis on normative systems creates systematic biases in the research process itself. It must be said, however, that the majority of commentaries published with this article took a contrary view. Also, one of the leading researchers in this field, Keith Stanovich, has placed concerns with rationality at the forefront of his research programme. In the latest of several books on the topic that examine individual differences (Stanovich, 2011), he has argued that normative solutions are closely related to both (a) cognitive ability (e.g., IQ) and (b) a disposition to think rationally rather than rely on intuition. This leads to yet another controversy, as other authors claimed that intuitions or 'gut feelings' are often a good basis for decision making (Gigerenzer, 2007; Gladwell, 2005).

Summary

- Peter Wason was the first researcher in the psychology of reasoning to place emphasis on cognitive biases and to seriously question the rationality of his participants.
- He invented two important reasoning tasks: the selection task, which is the main focus of the present chapter, and the 2–4–6 task, which is covered briefly.
- Wason's work on bias and irrationality was revolutionary, starting in the period dominated by Piaget and logicist treatments of human thinking. It preceded the major heuristics and biases programme launched by Tversky and Kahneman (1974).
- Wason claimed evidence for a confirmation bias on both the 2–4–6 and selection tasks, although later researchers have suggested alternative accounts in terms of matching bias and positive-testing bias.
- Early investigations of thematic content effects on the selection task led to major discoveries using the realistic form of the selection task. The task became a focus for the study of pragmatic effects, showing that a reasoning task is much influenced by the prior beliefs that a particular context can elicit.
- Study of pragmatically rich versions of the selection task also led to a series of major theoretical developments, including pragmatic reasoning schemas, evolutionary theory, relevance theory and decision-theoretical approaches.
- The selection task has been the focus of much debate about human rationality. While published studies of the task have decreased somewhat in recent years, it remains one of the most important tools for the study of human reasoning yet to be invented.

Further reading

Publication of papers on the selection task has fallen off in recent years, so the review offered by Evans and Over (2004, Chapter 5) still covers the bulk of work and most of the famous papers on this problem. More recent reviews and discussions are to be found in Evans (2007, Chapters 2 and 4) and Manktelow (2012, Chapters 3 and 4).

References

Ball, L. J., Lucas, E. J., Miles, J. N. V., & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *Quarterly Journal of Experimental Psychology, 56A*(6), 1053–1077.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*, 391–416.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences, 4*, 317–370.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? *Cognition, 31*, 187–276.

Cosmides, L., & Tooby, J. (1994). Beyond intuition and instinct blindness: Toward an evolutionary rigorous cognitive science. *Cognition, 50*, 41–77.

Elqayam, S., & Evans, J. St. B. T. (2011). Subtracting 'ought' from 'is': Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences, 34*, 233–290.

Evans, J. St. B. T. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking & Reasoning, 4*, 45–82.

Evans, J. St. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin, 128*, 978–996.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove, UK: Psychology Press.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology, 59*, 255–278.

Evans, J. St. B. T. (in press). Reasoning, biases and dual processes: The lasting impact of Wason (1960). *Quarterly Journal of Experimental Psychology, 68*. DOI: 10.1080/17470218.2014.914547.

Evans, J. St. B. T., & Handley, S. J. (1999). The role of negation in conditional inference. *Quarterly Journal of Experimental Psychology, 52A*, 739–769.

Evans, J. St. B. T., & Lynch, J. St. (1973). Matching bias in the selection task. *British Journal of Psychology, 64*, 391–397.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove: Psychology Press.

Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford: Oxford University Press.

Fiddick, L., Cosmides, L., & Tooby, J. (2000). No interpretation without representation: The role of domain-specific representations and inferences in the Wason selection task. *Cognition, 77*, 1–79.

Fodor, J. (1983). *The modularity of mind*. Scranton, PA: Crowell.

Fodor, J. (2000). Why we are so good at catching cheaters? *Cognition, 75*, 29–32.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. London: Penguin.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating and perspective change. *Cognition, 43*, 127–171.

Gladwell, M. (2005). *Blink*. London: Penguin.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic materials effect in the Wason selection task. *British Journal of Psychology, 73*, 407–420.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review, 94*, 211–228.

Laming, D. (1996). On the analysis of irrational data selection: A critique of Oaksford & Chater (1994). *Psychological Review, 103*, 364–373.

Lucas, E. J., & Ball, L. J. (2005). Think-aloud protocols and the selection task: Evidence for relevance effects and rationalisation processes. *Thinking & Reasoning, 11*(1), 35–66.

Manktelow, K. I. (2012). *Thinking and reasoning*. Hove, UK: Psychology Press.

Manktelow, K. I., & Over, D. E. (1991). Social roles and utilities in reasoning with deontic conditionals. *Cognition, 39*, 85–105.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608–631.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world*. Hove, UK: Psychology Press.

Over, D. E. (2003). From massive modularity to metarepresentation: The evolution of higher cognition. In D. E. Over (Ed.), *Evolution and the psychology of thinking: The debate* (pp. 121–144). Hove, UK: Psychology Press.

Poletiek, F. (2001). *Hypothesis-testing behaviour.* Hove, UK: Psychology Press.

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition, 57*, 31–95.

Sperber, D., & Girotto, V. (2002). Use or misuse of the selection task? Rejoinder to Fiddick, Cosmides and Tooby. *Cognition, 85*, 277–290.

Stanovich, K. E. (2011). *Rationality and the reflective mind.* New York: Oxford University Press.

Thompson, V. A., Evans, J. St. T., & Campbell, J. I. D. (2013). Matching bias on the selection task: It's fast and feels good. *Thinking & Reasoning, 19*(3–4), 431–452.

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*(3), 107–140.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Wason, P. C., & Evans, J. St. B. T. (1975). Dual processes in reasoning? *Cognition, 3*, 141–154.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content.* London: Batsford.

9

BELIEF BIAS IN DEDUCTIVE REASONING

Jonathan St. B. T. Evans

In Chapter 8, I described work on the Wason selection task, which has been intensively studied by reasoning researchers and provided evidence of a number of biases. In this chapter, I will look at deductive-reasoning tasks proper, especially those most commonly studied: syllogistic reasoning and conditional inference. These tasks have given rise to widespread claims of belief biases that influence reasoning. One researcher (Stanovich, 1999) even went so far as to claim that the tendency to contextualise all problems is the fundamental computational bias in human cognition, responsible for many errors in human reasoning and decision making. Before examining these claims, let us look at the nature of deduction itself.

Deductive reasoning involves drawing conclusions that necessarily follow from some given information. For example, if I told you that Sally is shorter than Mary and that Mary is taller than Joan, you could safely conclude that Mary is the tallest of the three. However, if I asked you who was taller, Joan or Sally, you could not infer the answer from the information given. This is because the information given is consistent with two possible situations that you might represent in *mental models* (Johnson-Laird & Byrne, 1991) as follows:

Mary > Sally > Joan (A)
Mary > Joan > Sally (B)

These models allow us to deduce who is tallest, but not, for example, who is shortest. Most people can solve this kind of problem, although they might need to think about it for a few seconds before answering. Consider a more complex reasoning problem, like the following from the study of Handley and Evans (2000):

> You urgently need to get hold of your friend Jane. Jane is on holiday somewhere in Britain. You know she is staying in a youth hostel, but you do not

know in which city. Jane, being somewhat mysterious, gives you the following information about her whereabouts:

If Jane is in Hastings, then Sam is in Brighton.
Either Jane is in Hastings or Sam is in Brighton, but not both.

Based on this information, does it follow that:

a Jane is in Hastings.
b Jane is not in Hastings.
c It is impossible to tell whether or not Jane is in Hastings.

The reader may care to give this problem some thought before reading on. It is possible to draw a definite conclusion that follows logically from the stated information, although it is hard to see. It involves what is called suppositional reasoning, where you need to suppose a possibility for the sake of argument. In this case, let us suppose that Jane is in Hastings and see what follows. Clearly, we can conclude from the first piece of information that Sam is in Brighton. However, the second statement tells us that either Jane is in Hastings or Sam is in Brighton *but not both*. So if Jane is in Hastings, by the second statement it follows that Sam is not in Brighton. So we have a contradiction. Our supposition that Jane is in Hastings has led to us to conclude both that Sam is in Brighton and that she is not in Brighton. Since this is an impossible state of affairs, it follows logically that our supposition is false. Hence, we can conclude that Jane is not in Hastings.

This kind of indirect reasoning is very hard for people who are not trained in logic, so don't worry if you didn't get the right answer. Handley and Evans gave this problem, among other similar ones, to undergraduate students as a pencil and paper task in class setting. Only 9.5% offered the correct answer (b). Of the remainder, 48% said (c), impossible to tell, and an astonishing 42.5% gave answer (a), which is the *opposite* of the correct answer. The authors offered an explanation in terms of mental-model theory. According to this theory, people try to imagine states of affairs, or mental models, that are suggested by the information given. We know that the first statement will suggest the model:

Jane is in Hastings; Sam is in Brighton.

even though people may realise that there are other possibilities. When they try to integrate the information in the second statement, they ought to reject this model because it is inconsistent. Those concluding that Jane is in Hastings must have overlooked the significance of the phrase 'but not both' in the second statement. However, it is likely that those who did notice the inconsistency mostly moved to the other incorrect answer: that it is impossible to tell. Although people may acknowledge that the statement 'if p, then q' allows possibilities other than p and q to be the case, no other state of affairs comes easily to mind for most people, so it seems to them that no conclusion is possible.

The small number who can solve the above task must either reason by supposition, as shown above, or by examination of all logical possibilities. If you list them out as follows:

1 Jane is Hastings and Sam is in Brighton.
2 Jane is in Hastings and Sam is not in Brighton.
3 Jane is not in Hastings and Sam is in Brighton.
4 Jane is not in Hastings and Sam is not in Brighton.

and check each in turn against the two statements, you will discover that there is just one state of affairs that is possible, Number 3. So you can conclude that Jane is not in Hastings *and* that Sam is in Brighton. However, contrary to proposals once made popular by the Swiss psychologist Jean Piaget, most adults do not approach deductive reasoning by this kind of exhaustive logical analysis.

Why is it important to study deductive reasoning in psychology? The answer to this question has changed quite radically over the past 40 years or so. The paradigm was developed at a time when Piaget's views were very influential and when most psychologists and philosophers saw logic as the basis for rational thinking (Evans, 2002). Hence it seemed a good idea to give people logical problems and see whether they could solve them, in order to determine how rationally people can reason. This meant giving people problems where they must assume the premises are true, introduce none of their real-world knowledge and draw only conclusions that strictly and necessarily follow. A large number of experiments of this kind have been conducted from the 1960s onwards, with two general findings (Evans, 2007; Manktelow, 2012). First of all, people make many errors when their answers are compared with a logical analysis of what is right and wrong. Second, they are highly influenced by the content and context in which the problem is framed, even though that is irrelevant to the logical task they are set. This led to a big debate about human rationality (see Chapter 8).

The study of reasoning biases is, however, of considerable psychological interest in its own right. In this chapter, I will focus mostly on belief bias, a topic which arose originally in the study of syllogistic reasoning.

Syllogistic reasoning

A popular form of reasoning task that is used in the study of deductive reasoning is the syllogism, originally invented by Aristotle. A syllogism consists of two premises and a conclusion, which are always in one of the following four forms:

All A are B
No A are B
Some A are B
Some A are not B

The syllogism always links three terms together. I will use the convention of assuming that the conclusion links two terms A and C and that each premise therefore connects with a middle term, B. A complete syllogism might be of the form:

All A are B
All C are B
Therefore, all A are C

This argument is a fallacy; that is, its conclusion does not necessarily follow. This can be seen easily if I substitute some realistic terms for A, B and C as follows:

All dogs are animals
All cats are animals
Therefore, all dogs are cats

The fallacy would be much harder to see if I substituted some different terms as follows:

All tigers are animals
All cats are animals
Therefore, all tigers are cats

People are strongly influenced by whether they agree with conclusions (the belief bias effect, discussed below). However, the task they are set is to say whether or not the conclusion follows in light of the logical structure of the argument. When syllogisms are presented in abstract form – say, using letters as in the earlier examples above – error rates are very high indeed. Hundreds of different syllogisms can be formed by varying the terms (technically, the mood) used in each premise and conclusion (all, no, some, some not) and by varying the order of reference to the terms A, B, and C. With the conclusion in the form A–C the premises can take four different arrangements (or figures, as they are technically known): A–B, B–C; A–B, C–B; B–A, B–C; B–A, C–B. With three statements (two premises and a conclusion) in each of four moods and four different figures, you can make 256 logically distinct syllogisms.

Psychological experiments on abstract syllogistic reasoning are reviewed in detail by Evans et al. (1993, Chapter 7) and a study in which people evaluated every possible logical form was reported by Evans, Handley, Harper, and Johnson-Laird (1999). When syllogisms are given to people to evaluate, they frequently say that the conclusion follows, even though the great majority are actually invalid, or fallacious. The endorsement of fallacies is a strong bias in deductive reasoning research as a whole. With syllogisms, people are also biased by both the mood and the figure. For example, they are more likely to say that the conclusion follows if it is of a similar mood to the premises. The first syllogism described above, although invalid, has

two 'all' premises and an 'all' conclusion. This type of fallacy is much more often endorsed than one where the conclusion is incongruent, such as:

All A are B
All C are B
Therefore, no A are C

Note that the conclusion here is possible given the premises, just as it was in the 'all' form. People are also biased by the figure of the syllogism. They would, for example, be much more likely to agree with the following argument:

Some A are B
Some B are C
Therefore, some A are C

Than this one:

Some A are B
Some B are C
Therefore, some C are A

even though both are actually fallacies. In the first case, the terms seem to follow in a natural order. These fallacies and biases of abstract syllogistic reasoning bear a similarity to the matching bias effect discussed in Chapter 8. They suggest very superficial processing by most of the participants. They also suggest that the typical populations of undergraduate students find abstract logical reasoning very difficult. What, however, if problems are made more realistic and easier to relate to everyday knowledge and thinking? How will that affect people's ability to reason deductively? We consider relevant studies in the following sections.

Belief bias in syllogistic reasoning

One of the major phenomena studied in deductive reasoning research is that of *belief bias*. Belief bias is typically described as a tendency to endorse arguments whose conclusions you believe, regardless of whether they are valid or not. This is not very accurate because, as we have already seen, people tend to endorse fallacies when syllogisms have abstract or neutral content. The belief bias effect is really a suppression of fallacies when conclusions are unbelievable, and so might be better called a *debiasing* effect, at least for invalid syllogisms.

The usual method by which belief bias is studied involves giving people syllogisms and asking them whether the conclusion necessarily follows (full details of the experimental method can be found in Text box 9.1). Some of the conclusions are logically valid deductions and some are not, and some have believable conclusions

TEXT BOX 9.1 CLASSROOM DEMONSTRATION OF BELIEF BIAS IN SYLLOGISTIC REASONING

Participants

The effects are typically quite large and have been demonstrated with small samples. I recommend a minimum of 32 participants drawn from a population of average or above average intelligence.

Materials

The material consists of syllogisms like those in Table 9.1. Note that there are two logical forms used. The valid form is

No C are B
Some A are B
Therefore, some A are not C

and the invalid form is

No A are B
Some C are B
Therefore, some A are not C

It is necessary to keep these same forms, which were carefully chosen, but the design calls for two syllogisms of each type to be presented. The four shown in Table 9.1 can be used, but at least one other set is needed. In order to get a powerful effect, the experimenter needs to make sure that the conclusions follow (believable) or violate (unbelievable) a *class-inclusive* relationship, as do those in the table. For example, all cigarettes are addictive but not all addictive things are cigarettes. So while it is believable, as in the example shown, to say that 'some addictive things are not cigarettes', it is unbelievable if it is turned around as 'some cigarettes are not addictive'. Ideally, the experimenter also checks the believability of the conclusions by asking a separate group of participants to rate them on a five-point scale from 'Highly Unbelievable' to 'Highly Believable' and working out the average ratings. About 16 participants is sufficient for the rating study.

Design

The design is within-participants. Each participant will be asked to solve all eight problems, presented in an independently randomised order.

Procedure

A booklet with 8 pages is given to each participant. Each page contains one of the problems, with a layout like this:

> Given
>
> No millionaires are hard workers
> Some rich people are hard workers
>
> Does it follow that
>
> Some millionaires are not rich people
> YES NO

The experimenter assigns each problem a number from 1 to 8 and uses a random-number table or spreadsheet program to work out a separate random sequence for each participant. The problem sheets are numbered in the corner before they are copied and then put together in right order for each participant. The front cover of the booklet can have the written instructions. The participants are instructed to reason deductively, as otherwise the influence of belief could not be considered a bias. Typical instructions (adapted from Evans et al., 1983) would be:

> This is a test of reasoning ability. You will be given eight problems. In each case you are given two statements that you should assume to be true. You will also be asked if a further statement – that is, a conclusion – follows from these two statements. If you think the conclusion *necessarily* follows, then mark the word 'YES'; otherwise, mark 'NO'. Take your time and make sure you have the right answer. Do not go back to a previous problem once you have left it.

Analysis and Results

There are three effects of interest: an effect of logic, an effect of belief, and an *interaction* between the two. There are too few measures for parametric analysis, but there is a way around this. First, a table can be made up with a row for each participant and a column for each condition: VB, VU, IB, and IU. For each participant, the number of Yes answers he or she gave for each type are recorded – these must be between 0 and 2. The next step is to compute for each participant the value of three indices in a further three columns. These indices are computed by adding two different pairs of columns and then subtracting the totals as follows:

Logic Index: (VB + VU) – (IB + IU)
Belief Index: (VB + IB) – (VU + IU)
Interaction Index: (VU + IB) – (VB + IU)

The first two are fairly obvious. The Logic Index is the number of valid conclusions accepted minus the number of invalid conclusion accepted with belief balanced. Conversely, the Belief Index measures the difference between the acceptance of believable and unbelievable conclusions with logic controlled. The Interaction Index is designed to measure whether (as is usually found) the belief-bias effect is larger for invalid than for valid syllogisms. One might want to think of this as (IB – IU) – (VB – VU), although it is algebraically equivalent to the above. Tests for statistical significance are simple. For each index, it can be determined whether it is significantly above zero by using the sign test. For each index, the number of participants with a score above zero is counted and compared with the number at or below zero, using the binomial test. Since these effects are well known, one-tailed tests can be used in each case.

and some do not. A well-known study by Evans, Barston, and Pollard (1983) is often cited as showing clearly the basic phenomena with the necessary controls, although there are much earlier reports of belief bias to be found. Evans et al. presented four categories of syllogisms classified as Valid-Believable (VB), Valid-Unbelievable (VU), Invalid-Believable (IB), and Invalid-Unbelievable (IU). Examples of these are shown in Table 9.1 together with the rates at which people accepted them as valid arguments over three experiments.

TABLE 9.1 Examples of four kinds of syllogisms presented by Evans et al. (1983) together with the percentage rates of acceptance of each argument as valid over three experiments

Type	*Example*	*Rate*
Valid-Believable	No police dogs are vicious Some highly trained dogs are vicious Therefore, some highly trained dogs are not police dogs	89%
Valid-Unbelievable	No nutritional things are expensive Some vitamin tablets are expensive Therefore, some vitamin tablets are not nutritional	56%
Invalid-Believable	No addictive things are inexpensive Some cigarettes are inexpensive Therefore, some addictive things are not cigarettes	71%
Invalid-Unbelievable	No millionaires are hard workers Some rich people are hard workers Therefore, some millionaires are not rich people	10%

The higher acceptance rate for believable syllogisms compared to the unbelievable ones, both for valid and invalid cases, indicates the presence of a strong belief bias. Evans et al. (1983) also showed that people accepted significantly more valid than invalid arguments in line with the logical task set. In addition, they observed a belief by logic interaction such that the belief bias effect was significantly larger for invalid than valid arguments. All three of these effects have been replicated in a number of subsequent studies (see Evans, Newstead, & Byrne, 1993; Klauer, Musch, & Naumer, 2000; Newstead, Pollard, Evans, & Allen, 1992).

In the literature, a lot of interest has been focussed on the cause of the belief by logic interaction. One explanation offered by Evans et al. (1983) has become known as the Selective Scrutiny Model. According to this model, people tend to accept believable conclusions uncritically and check the logic more thoroughly when it is unbelievable. Hence, they are more likely to detect a fallacy when it leads to a conclusion with which they disagree. A mental-models theory version of this proposes that people form an initial mental model of the premises that supports typically fallacious conclusions that can only be refuted by searching for a counterexample model, that is, one in which the premises do not support the conclusion. Such search for counterexamples is more likely to occur when the conclusion is unbelievable (Oakhill, Johnson-Laird, & Garnham, 1989). The Selective Scrutiny Model is supported by evidence that when people are forced to respond rapidly, the belief bias effect increases and the belief-logic interaction disappears (Evans & Curtis-Holmes, 2005). From a dual-process theory viewpoint (see Chapter 8) we can suppose that there are two kinds of belief bias: A Type 1 belief bias is a response bias that should favour believable conclusions across the board; a Type 2 belief bias involves motivated reasoning and accounts for the interaction. Since standard dual-process theory claims that Type 1 processing is quicker than Type 2, it follows that the Type 2 belief bias should be suppressed by rapid responding, leading to loss of the interaction.

Some findings in the literature, however, appear inconsistent with the Selective Scrutiny Model. For example, people may spend most time reasoning about the invalid-believable problem, which should be rapidly accepted, according to the model (Thompson, Newstead, & Morley, 2011), although other authors have interpreted response latencies as consistent with the model (Stupple, Ball, Evans, & Kamal-Smith, 2011). A slightly different account is the Selective Processing Model (Evans, Handley, & Harper, 2001; Klauer et al., 2000) in which people are supposed only to form one model of the premises, without a subsequent search for counterexamples. In this model, people will try to form a mental model that is consistent with the premises unless the conclusion is unbelievable; in the latter case they will try to find a model which refutes the conclusion, that is, they search for a counterexample model from the start. This is more difficult, however, leading to the interaction usually observed.

It has been claimed recently that the belief bias effects in syllogistic reasoning, including the interaction, can be accounted for as a response-bias effect only when analysed using the methods of signal detection theory (Dube, Rotello, & Heit,

2010). If this is correct, then belief bias would all be Type 1. However, the argument involves technical issues and has been disputed on theoretical and empirical grounds (Klauer & Kellen, 2011; Trippas, Handley, & Verde, 2013). Overall, there is no doubt about the reliability of the belief bias effect, and my view remains that the balance of evidence supports the view that both response biases (Type 1) and motivated reasoning (Type 2) are involved. But is this effect restricted to the world of syllogisms? I examine below evidence for belief biases in two other domains: conditional inference, which is the major paradigm used currently in the psychology of deductive reasoning, and also informal inference.

Conditional inference

The conditional inference task has been widely employed in the psychology of reasoning for many years, and in recent times has taken over in popularity from the Wason selection task as the most used paradigm. Participants are invited to endorse or reject each of four inferences that can be drawn from a conditional statement. I will illustrate this first with letter-number rules, similar to those used for the abstract version of the Wason selection task (Chapter 8). The conditional statement relates the letter on one side of the card to a number on the other side. The statement might be

> If there is an A on one side of the card, then there is a 3 on the other side of the card.

The four inferences are the following:

Modus Ponens (MP)	The letter is an A; therefore, the number is a 3.
Denial of the Antecedent (DA)	The letter is not an A; therefore, the number is not a 3.
Affirmation of the Consequent (AC)	The number is a 3; therefore, the letter is an A.
Modus Tollens (MT)	The number is not a 3; therefore, the letter is not an A.

Although there are different theories of what the conditional statement means (Evans & Over, 2004) they all agree that only MP and MT are valid inferences. MP is pretty obvious but MT requires a little reflection. If a card does not have a 3 on it, it cannot have an A because if it did have an A, the number would be 3. With abstract problems like these, most people endorse MP but MT is endorsed significantly less frequently (Evans & Over, 2004, Chapter 3). The DA and AC inferences are *fallacies*, meaning that they do not necessarily follow: It is consistent with the rule to have a 3 paired with some letter other than A. Studies with abstract conditionals, however, show that these two inferences are quite often endorsed. There is no

logical interpretation of the conditional (e.g., as a biconditional) that is consistent with the overall pattern of responding and it is likely that several cognitive biases could operate on particular inferences. As with matching bias (Chapter 8), some of these can be illustrated by introducing negative components. Consider these two forms of MT:

1 If the letter is A, then the number is 3. The number is not 3; therefore, the letter is not A.
2 If the letter is not A, then the number is 3. The number is not 3; therefore, the letter is A.

Both inferences are valid, but 2 requires an extra step: Strictly speaking, the conclusion of MT is that 'the letter is not not an A'. Logic allows us to remove this double negation as it means the same as 'the letter is A'. Perhaps not surprisingly, fewer people endorse 2 than 1. Now compare these:

3 If the letter is A, then the number is 3. The number is not 3; therefore, the letter is not A.
4 If the letter is A, then the number is 3. The number is 4; therefore, the letter is not A.

In 4, the negation of the second premise is implicit: The number is a 4 and therefore, by implication, not a 3. In such cases, people draw MT less frequently (Evans & Handley, 1999), an effect analogous to the matching bias discussed in Chapter 8.

The main interest in this chapter is the effect that prior beliefs about realistic conditional statements have on the inferences people draw. Before discussing relevant studies, however, I need briefly to explain the difference between the old and new paradigms in the psychology of reasoning.

Old and new paradigms

The psychology of deductive reasoning has traditionally used the *deduction paradigm* founded in classical binary logic (Evans, 2002). In classical logic, all statements are either true or false and all inferences either valid or invalid. A valid inference is one whose conclusion can never be false if all of its premises are true. Thus, deduction is classically truth preserving: True premises guarantee true conclusions. The experimental method in the deduction (or old) paradigm reflects this kind of logic. Participants are asked (a) to assume that the premises given are true and (b) to endorse only those conclusions which necessarily follow. Research with this paradigm led to the discovery of many errors and biases in human reasoning. However, a number of researchers began to question whether binary logic and the deduction paradigm provided the right approach for the psychology of reasoning (Evans, 2002; Oaksford & Chater, 1998), leading to the development of a *new paradigm* that has become popular in recent years.

In making this shift, psychologists were influenced by the writings of some philosophers (e.g., Edgington, 1995) who exposed the problems with conditional statements in the binary-logic paradigm. In order for a conditional always to be true or false, 'if p, then q' has to mean the same thing as 'not-p or q'. This leads to some silly and unacceptable inferences (see Chapter 8). For example, it must be true that 'If David Cameron is president of the USA, then he will declare war on Russia', simply because David Cameron is not actually president of the USA. In fact, any conditional statement is true in classical logic when the antecedent is false or the consequent is true. So given that 2 + 2 = 4, then we can infer that 'If the moon is made of blue cheese, then 2 + 2 = 4' and so on.

The solution favoured by Edgington and by many psychologists now (e.g., Evans & Over, 2004) is that a conditional is not always true or false. Instead, it has a probability or degree of belief that can only be evaluated when the antecedent is true. To decide if we believe a conditional statement, we perform a thought experiment. We imagine that the antecedent is true and then decide the extent to which we believe the consequent. So if I say to you 'If the Democrats are elected, then taxes will rise', you imagine the premise of this scenario and use your knowledge of the past behaviour of the Democratic party, their current policies, the state of the economy etc. to determine how likely it is that taxes will rise. This determines your belief in the conditional statements, so that the P(if p, then q) = P(q|p), often referred to as the Equation. There is now considerable evidence that people's actual beliefs about conditional statements does conform to the Equation (e.g., Over, Hadjichristidis, Evans, Handley, & Sloman, 2007). People believe conditional statements to the extent that they believe q when they imagine p.

In the new paradigm, participants are no longer asked to assume the truth of premises and are often asked to give degrees of belief in conclusions. With these changes, use of the conditional inference paradigm continues.

Belief bias in conditional inference

Beliefs bias conditional inference but not in the same way as they bias syllogistic reasoning. In syllogistic reasoning, inferences are biased by the believability of the conclusion. With conditional inference, however, it is belief in the conditional statement itself – the major premise – that is important. It was discovered some time ago that the valid inferences, MP and MT, can be suppressed if participants do not believe the conditional statement (Byrne, 1991; George, 1995; Stevenson & Over, 1995). Here is an example, using the method employed by Byrne.

Argument 1

If Ruth meets her friend, they will go to a play.
Ruth meets her friend.
What follows?

Argument 2

If Ruth meets her friend, they will go to a play.
If Ruth has enough money, she will go to a play.
Ruth meets her friend.
What follows?

With Argument 1 most people will say that Ruth goes to the play – Modus Ponens. With Argument 2, however, far fewer people give this conclusion, even though it is as logically necessary as it was in the first argument. The reason is that the second conditional has created doubt about the first one. We cannot be sure that Ruth will go to the play just because she meets her friend; she must have enough money as well. These kinds of *pragmatic* effects (beliefs induced by context) can be readily demonstrated. It does not need an active suppressor as in the above method. We can simply measure people's *a priori* belief in conditional statements and observe that fewer inferences are made from the ones they disbelieve. In fact, with this method, all four inferences, valid and invalid, are suppressed by such disbelief (see, e.g., Evans, Handley, Neilens, Bacon, & Over, 2010).

The suppression of conditional inferences can be demonstrated using methods of both old and new paradigms. For example, in the classical paradigm people will less frequently say that an inference follows when it is suppressed by disbelief, and in the new paradigm they will give a lower probability rating to the conclusion. One study that used both methods (Evans et al., 2010) found that when people were just given arguments and asked to rate probabilities of conclusions (new method), they endorsed fewer inferences of all four kinds, regardless of cognitive ability. However, when given standard deductive reasoning instructions (assume premises, decide necessity of conclusion) this belief effect was suppressed by high- but not low-ability participants. This is a result consistent with dual-process theory (see Chapter 8) particularly of the kind proposed by Stanovich (2011). According to Stanovich, people need a combination of cognitive capacity and motivation in order to engage in effective formal reasoning. Unless both are present, people will rely on belief-based reasoning.

The old-paradigm interpretation of these belief effects was that inferences could be blocked by the availability of counterexamples. This permits the binary approach (e.g., mental-model theory). Take the case of AC: 'If it is a cat, it has four legs; it has four legs, therefore it is a cat'. We would expect few people to endorse this because a counterexample case, for example, a dog with four legs, is easy to think of. The new paradigm approach would be to say that this conditional statement has a low degree of belief. An important difference, however, is that belief effects are not necessarily regarded as biases in the new paradigm. In this approach, we do not expect people to engage in assumption-based reasoning except as a special kind of problem solving, which only high-IQ people seem to be good at. Everyday reasoning, such as that supporting decision making, can and should take account of the belief we have in premises. The new paradigm still deals with deduction but allows that confidence in conclusions can rightly be affected by confidence in premises (see Evans & Over, 2012).

Belief bias in informal reasoning

In addition to syllogistic and conditional inference studies, psychologists have investigated belief biases in informal reasoning and argumentation. Stanovich and West (1997) devised what they call the Argument Evaluation Task, or AET for short. Here is an example item:

Dale's belief: Seventeen-year-olds should have the legal right to drink alcoholic beverages.

Dale's premise or justification for belief: Seventeen-year-olds are just as responsible as 19-year-olds, so they ought to be granted the same drinking rights as other adults.

Critics' counter-argument: Seventeen-year-olds are three times more likely to be involved in automobile accident while under the influence of alcohol than 19-year-olds (assume statement factually correct).

Dale's rebuttal to critics' counter-argument: Seventeen-year-olds will drink no matter what the law says (assume statement factually correct), so it is useless to try to legislate that they not drink.

Indicate the strength of Dale's rebuttal to critics' counter-argument:

A = Very Weak B = Weak C = Strong D = Very Strong

Stanovich and West used a panel of expert judges to decide how strong Dale's arguments were (in this case, judged to be weak). Of course, participants will also vary in the extent to which they believe the conclusion for which Dale is arguing. If asked to judge the strength of the argument, they should disregard such beliefs: failure to do so is seen as a form of belief bias. Stanovich and West have used the AET in many studies since (Stanovich, 1999), showing that individuals who score well (i.e., resist belief) are both higher in cognitive ability and in the disposition to think rationally, as measured by a number of psychometric scales. Another informal reasoning task measures people's tendency to take account of the law of large numbers when assessing evidence for propositions they do or do not agree with (Klaczynski & Robinson, 2000). These authors also found evidence of belief bias, for example, disregarding a small sample when the study favoured a conclusion with which they agreed. However, the effect was much stronger in older than in younger adults.

Although it is tempting to assume that belief biases in formal and informal reasoning tasks must be essentially the same thing, there is little evidence to support this. Thompson and Evans (2012) ran tests of syllogistic belief bias, AET, and the law of large numbers on the same participants. There was little evidence of any correlation between the different measures of belief bias, nor did the tasks respond in the same way to relevant experimental manipulations. As Thompson (2000) has demonstrated previously, many reliable effects in reasoning tasks turn out to be task-specific.

Conclusions

People's reasoning is undoubtedly and strongly affected by pragmatic factors. When realistic and thematically rich materials are employed, the beliefs that they evoke change the answers that are given to the questions. From the viewpoint of the old paradigm in the psychology of reasoning, all such effects are biases. This is because the old paradigm tests assumption-based reasoning, in which participants must always evaluate the form of a logical argument and assume its premises to be true. Belief effects demonstrate how difficult this is for ordinary people to do. In fact, the only people at all good at this have unusually high IQs.

These findings should not concern us if we believe that only a minority of the population need to be expert in abstract thinking, for example, mathematicians, scientists, designers of information systems and so on. We can expect this minority both to have high intelligence and a large amount of specialist training. The problem with the old paradigm, however, was that researchers assumed that ordinary people without training should be able to reason logically. This is clearly not the case, and most psychologists now believe that this approach was mistaken. Most people are intelligent enough to make reasonably good decisions and achieve their everyday goals. It therefore seems likely that they achieve this by belief-based rather than abstract reasoning. Thus, we should not be defining all influences of belief as bias.

The new paradigm is in its infancy and we have much to learn as yet about whether people generally can take beliefs into account in ways which are effective and appropriate. The likely truth is that they do to quite an extent – hence, effective everyday decision making – but not all the time. Studies of informal reasoning show that belief can bias judgements on tasks that are not formal and do not require very abstract reasoning, as do many related studies in social psychology. Beliefs can be a source of bias, but they are also essential to human rationality from a new-paradigm perspective. We can only make the best decision for ourselves by acquiring and using all relevant beliefs.

Summary

- Belief bias in deductive reasoning is the tendency to draw or approve conclusions that conform with prior beliefs, whether or not logically supported by the premises.
- The classical effect is demonstrated using three-term syllogisms. Many studies have shown three reliable effects: (a) people accept more valid than invalid arguments, (b) people accept more believable than unbelievable conclusions (belief bias), and (c) the belief bias is stronger for invalid arguments.
- These effects are often interpreted within dual-process theory, in which belief bias is attributed to intuitive Type 1 processing, while the validity effect reflects effortful Type 2 reasoning. The belief by logic interaction is interpreted as a Type 2, motivated reasoning bias.

- Belief bias of a different kind can be shown in conditional inference. In this case, it is belief in the conditional statement itself that causes a bias. People are more inclined to accept inferences drawn from believable conditional statements, regardless of their logical validity.
- In the new paradigm of reasoning, influences of belief are not necessarily considered biases. The new approach allows that confidence in conclusions can reflect the degree of belief that people hold in premises. Assumption-based reasoning, required by the old paradigm, is important in formal problem solving but less relevant to everyday decision making.
- Belief bias has also been demonstrated in informal reasoning tasks, where arguments may be perceived as stronger simply because someone agrees with their conclusions. No clear link with belief bias in formal reasoning tasks has been established, however.

Further reading

Detailed review and discussion of belief-bias models of syllogistic reasoning is provided by Klauer et al. (2000). For a less formal and broader treatment of the role of knowledge and belief in reasoning, see Evans (2007, Chapter 4).

References

Byrne, R. M. J. (1991). Can valid inferences be suppressed? *Cognition, 39*, 71–78.

Dube, C., Rotello, C., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review, 117*, 831–863.

Edgington, D. (1995). On conditionals. *Mind, 104*, 235–329.

Evans, J. St. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin, 128*, 978–996.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove, UK: Psychology Press.

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11*, 295–306.

Evans, J. St. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning, 11*(4), 382–389.

Evans, J. St. B. T., & Handley, S. J. (1999). The role of negation in conditional inference. *Quarterly Journal of Experimental Psychology, 52A*, 739–769.

Evans, J. St. B. T., Handley, S. J., & Harper, C. (2001). Necessity, possibility and belief: A study of syllogistic reasoning. *Quarterly Journal of Experimental Psychology, 54A*(3), 935–958.

Evans, J. St. B. T., Handley, S. J., Harper, C., & Johnson-Laird, P. N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1495–1513.

Evans, J. St. B. T., Handley, S., Neilens, H., Bacon, A. M., & Over, D. E. (2010). The influence of cognitive ability and instructional set on causal conditional inference. *Quarterly Journal of Experimental Psychology, 63*(5), 892–909.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove, UK: Erlbaum.

Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford: Oxford University Press.

Evans, J. St. B. T., & Over, D. E. (2012). Reasoning to and from belief: Deduction and induction are still distinct. *Thinking & Reasoning, 19*(3–4), 267–283.

George, C. (1995). The endorsement of the premises: Assumption-based or belief-based reasoning. *British Journal of Psychology, 86*, 93–111.

Handley, S. J., & Evans, J. St. B. T. (2000). Supposition and representation in human reasoning. *Thinking & Reasoning, 6*, 273–312.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove & London: Erlbaum.

Klaczynski, P. A., & Robinson, B. (2000). Personal theories, intellectual ability and epistemological beliefs: Adult age differences in everyday reasoning biases. *Psychology and Aging, 15*, 400–416.

Klauer, K. C., & Kellen, D. (2011). Assessing the belief bias effect with ROCs: Reply to Dube, Rotello & Heit (2010). *Psychological Review, 118*(1), 164–173.

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107*, 852–884.

Manktelow, K. I. (2012). *Thinking and reasoning*. Hove, UK: Psychology Press.

Newstead, S. E., Pollard, P., Evans, J. St. B. T., & Allen, J. L. (1992). The source of belief bias effects in syllogistic reasoning. *Cognition, 45*, 257–284.

Oakhill, J., Johnson-Laird, P. N., & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition, 31*, 117–140.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world*. Hove, UK: Psychology Press.

Over, D. E., Hadjichristidis, C., Evans, J. St. B. T., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology, 54*, 62–97.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mawah, NJ: Lawrence Erlbaum Associates.

Stanovich, K. E. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*(2), 342–357.

Stevenson, R. J., & Over, D. E. (1995). Deduction from uncertain premises. *Quarterly Journal of Experimental Psychology, 48A*, 613–643.

Stupple, E. J. N., Ball, L. J., Evans, J. St. B. T., & Kamal-Smith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. *Journal of Cognitive Psychology, 23*(8), 931–941.

Thompson, V. A. (2000). The task-specific nature of domain-general reasoning. *Cognition, 76*, 209–268.

Thompson, V. A., & Evans, J. St. B. T. (2012). Belief bias in informal reasoning. *Thinking & Reasoning, 18*(3), 278–310.

Thompson, V. A., Newstead, S. E., & Morley, N. J. (2011). Methodological and theoretical issues in belief bias: Implications for dual-process theories. In K. I. Manktelow, D. E. Over, & S. Elqayam (Eds.), *The science of reason* (pp. 309–338). Hove, UK: Psychology Press.

Trippas, D., Handley, S. J., & Verde, M. F. (2013). The SDT model of belief bias: Complexity, time, and cognitive ability mediate the effects of believability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1393–1402.

This page intentionally left blank

PART II

Judgment

This page intentionally left blank

10

AVAILABILITY

Rolf Reber

When you ask each spouse of a married couple to estimate the percentage of their own contribution to the housework, chances are high that each spouse overestimates one's own contribution so that the sum exceeds one hundred percent. People normally overestimate their own contribution to the joint product of a group (Ross & Sicoly, 1979). These authors found this effect in naturally occurring discussion groups, basketball players, groups assembled in the laboratory, and married couples. Why do people behave this way? One explanation would be that people are motivated to see themselves in a positive light and therefore overestimate their contribution to a joint product. It is, however, possible that cognitive processes alone, without any involvement of motivational processes, account for the observed overestimation. Let us take a closer look at the married couple. If the husband is asked about his contribution to the housework, he retrieves information that is relevant to the question. He recalls instances of preparing meals and cleaning the house. Moreover, he recalls instances of his wife doing the same work. However, this retrieval is biased: He is better at retrieving instances of his own housework than on instances of his wife's work. He remembers in some detail how he prepared a tiramisu. He may have forgotten, however, that his wife prepared paella, a work that needed about the same time and effort. Even if he remembers the paella, his memories of his own efforts expended for the tiramisu are probably more vivid than the memories of his wife's work. Other instances are remembered in an analogous way so that in general, he remembers more easily instances of his own contribution to the housework than of his wife's contribution. If he now has to estimate his own contribution, he compares the ease with which he can retrieve instances of his own work with the ease with which he can retrieve instances of his wife's work. As he can more easily remember his own contributions, he overestimates his share of the housework. Of course, his wife proceeds in the same manner, with the consequence that she can retrieve instances of her housework with greater ease; this results in an overestimation of her contribution.

The responsible mechanism leading to these overestimations might be "availability." This is one of the famous heuristics proposed by Tversky and Kahneman (1973), along with the representativeness heuristic (see Chapter 11 in this volume) and anchoring and adjustment (see Chapter 12). Text box 10.1 provides a definition of availability.

TEXT BOX 10.1 DEFINITION OF AVAILABILITY

Availability is the ease with which relevant instances of a class (in our example, housework) come to mind (Tversky & Kahneman, 1973).

Alternative terms to *availability* have been proposed. Higgins (1996), for example, distinguished between *availability* and *accessibility* in accordance to Tulving, who used "the term 'availability' to refer to the hypothetical presence of information in the memory store [. . .]. That part of the available information that could be recalled was said to be accessible" (Tulving, 1983, p. 203). Note that the term *availability* is used differently from the way Tversky and Kahneman (1973) used it. Higgins (1996) referred to *accessibility experiences* to refer to the ease with which instances come to mind. In this chapter, I use the term *availability* as a general heuristic relying on ease or amount of recall, and the term *accessibility experiences* or *ease of recall* when discussing the specific mechanisms behind the availability heuristic. In order to circumvent the vagueness of the concept (Betsch & Pohl, 2002), some scholars (e.g., Benjamin, Bjork, & Schwartz, 1998; Schooler & Hertwig, 2005) used the term *retrieval fluency* to make clear that ease of processing is the mechanism underlying the availability heuristic.

Let us apply the term *availability* to our example: Both the husband and his wife overestimate their own contribution to the housework because information about their own contribution is more available than information about their spouse's contribution. As they are unable to come to an objective assessment of the proportion of housework that each of them has contributed, they use the availability of information as a heuristic for their estimate.

Overestimation of one's contribution to the joint products of a group has been only one of many applications of the availability heuristic. Although availability is often a valid cue to frequencies in the environment, it sometimes causes biased estimates. In this chapter, I first describe two of the experiments from the classical paper of Tversky and Kahneman (1973). I then turn to some early applications of the availability heuristic, such as overestimation of the frequency of sensational events and the effects of vividness of information, before I discuss some studies that explore mental mechanisms behind availability.

Two basic experiments

I now report two classical studies from Tversky and Kahneman's (1973) seminal paper. For each experiment, I first report methods and results of the original study and then turn to adaptations of the studies for classroom demonstrations.

Experiment 1: The famous-names experiment

The basic idea of Tversky and Kahneman's (1973, Exp. 8) famous-names experiment was to show that estimates of frequency of occurrence depend on availability (see Text box 10.2 for a classroom demonstration).

TEXT BOX 10.2 CLASSROOM DEMONSTRATION OF EXPERIMENT 1

This is an easy experiment that always worked when I used it as a classroom demonstration, even with small sample sizes of 20 to 30 students (though the result was numerically in the right direction but not necessarily significant by a sign test). The design of the experiment can be simplified. The experimenter compiles a list of nine famous women and ten less famous men, or vice versa. From time to time, he or she has to update the list because some less famous people have risen to stardom and some famous ones have sunk into oblivion. The important part of the experiment is the participants' judgment of the relative frequency of men compared to women, after they have listened to the list read by the experimenter. The independent variable is the fame of the name; the dependent variable is the estimated prevalence of female versus male names. The experimenter then can simply count the frequency with which the gender with the more famous names was chosen or vice versa, and the outcome can be tested by a sign test.

The idea behind the other part of Experiment 1 discussed below – the recall of names – is to check whether famous names are indeed more readily recalled than less famous names. As we will see later, however, number of recalled names is not necessarily a good indicator of ease of recall.

Method

Participants were presented with a tape-recorded list of 39 names, at a rate of 2 seconds per name. The independent variable, manipulated within participants, was fame of the names. Some names were famous (e.g., Richard Nixon, Elizabeth Taylor), others less famous (e.g., William Fulbright, Lana Turner). Some participants heard names of public figures (e.g., Richard Nixon, William Fulbright), others of entertainers (e.g., Elizabeth Taylor, Lana Turner). In one group, 19 of these names were

of famous women and the remaining 20 of less famous men. In the other group, 19 names were of famous men and the remaining 20 of less famous women. Note that non-famous names always outnumbered famous names. There were two dependent variables: (a) After listening to the recordings, about half of the participants had to recall as many names as possible; this measure indicated the availability with which an instance could be recalled. Participants were assumed to represent famous names more vividly than non-famous names and therefore to recall the former more readily than the latter. (b) The other participants had to judge whether the list contained more names of men or of women. If people use the availability heuristic, they are expected to judge that there are more instances with famous names despite the fact that there were only 19 famous names and 20 less famous names. If 19 famous women and 20 less famous men have been shown, participants were expected to judge that there were more women in the list. In contrast, if 19 famous men and 20 less famous women have been presented, participants were expected to judge that more men were presented.

Results

The results were clear-cut: Those participants who had to recall as many names as possible recalled 12.3 of the 19 famous names and 8.4 of the 20 less famous names. Of 86 participants in the recall condition, 57 recalled more famous than less famous names, and only 13 recalled fewer famous names than less famous names. A sign test revealed that this difference was highly significant. Among the 99 participants who compared the frequency of men and women in the lists, 80 erroneously believed that the class consisting of the more famous names was more frequent. Again, a sign test revealed that this difference was highly significant. The authors concluded that the participants used the availability heuristic because they recalled more famous names and they judged famous names as being more frequent on the list.

Experiment 2: The letter-frequency experiment

Another classical experiment instructs participants to judge letter frequency (Tversky and Kahneman, 1973, Exp. 3; see Sedlmeier, Hertwig, & Gigerenzer, 1998, for a critical examination). A classroom adaptation of this experiment can be found in Text box 10.3.

TEXT BOX 10.3 CLASSROOM DEMONSTRATION OF EXPERIMENT 2

For a classroom demonstration, the experimenter may choose an uneven number of letters that in his or her language is more frequent in the third than in the first position.[1] The respondents are then asked to indicate whether each

of these letters is more frequent in the first position or in the third position. The independent variable is the objective letter position (first versus third), the dependent variable is the judgment of the perceived letter position (first versus third). The experimenter then simply counts how many respondents chose the first and how many chose the third position for the majority of letters. This kind of data can be analyzed with a sign test. When the original experiment replicates, we expect that it is easier to recall letters at the first position and that participants therefore overestimate the number of letters in the first position, compared to the third position. This finding is likely to be significant with a sign test with a relatively large number of participants. However, as an instructor, you will probably get a higher number of participants who judge that there are more letters in the first position rather than the third position, even if this result is unlikely to be significant with small sample sizes (say, less than 30).

Method

Participants of this study were given the following instructions (Tversky & Kahneman, 1973, p. 211f):

> The frequency of appearance of letters in the English language was studied. A typical text was selected, and the relative frequency with which various letters of the alphabet appeared in the first and third positions in words was recorded. Words of less than three letters were excluded from the count.
>
> You will be given several letters of the alphabet, and you will be asked to judge whether these letters appear more often in the first or in the third position, and to estimate the ratio of the frequency with which they appear in these positions.

The authors assessed two dependent variables: First, participants were asked whether a certain letter, for example, R, is more likely to appear in the first or in the third position. The participants had to mark the correct answer. Second, they were asked to estimate the ratio of these two values, that is, Rs in the first position divided by Rs in the third position. In their original study, the authors used five letters, K, L, N, R, and V, all of them occurring more frequently in the third than in the first letter position in English words. There was no manipulation of an independent variable; the authors were interested in the question of whether participants judged these letters to appear more frequently in the first position despite the fact that all of them were more frequent in the third position in English language.

Results

As it is easier to retrieve letters in the first position than letters in the third position, a majority of participants judged the first position to be more likely for the majority

of letters: From 152 participants, 105 judged the first position to be more likely for a majority of the presented letters, and 47 judged the third position to be more likely for a majority of the letters. The authors employed a sign test and found a highly significant preference for the first letter position. Moreover, each of the five letters was judged to be more frequent in the first rather than in the third position, with a median ratio of about 2:1, despite the fact that each letter was more frequent in the third position.

I now turn to an overview on research on the availability heuristic that is partitioned into two sections: First, I review early research that applied the concept of availability before I then turn to more recent discussions of mental mechanisms underlying the availability heuristic.

Applications of the availability heuristic

The concept of availability has become very popular and has gone far beyond estimates of frequencies. Availability had a huge impact, especially on social psychology. In this section, I discuss first how biased encoding and retrieval influence availability of information. I then turn to vividness as a basis of availability of information. Finally, I highlight the role of perspective taking for availability. Perspective taking is an instructive example of how both retrieval processes and vividness jointly contribute to availability.

Biased encoding and retrieval of information

Many people are afraid of becoming a victim of a crime, maybe more than is justified by official crime statistics. One possibility is that more crimes are committed than revealed in official statistics. Alternatively, people may overestimate the prevalence of violent crime because these are exhaustively covered and sensationalized by the media. Due to high media coverage, violent crimes become more available in memory, and their frequency is thus overestimated. Lichtenstein, Slovic, Fischhoff, Layman, and Combs (1978) examined this assumption in a study about judging the frequency of lethal events.

They chose 41 causes of death that varied widely in frequency. It is very uncommon to die from botulism, whereas stroke is one of the more frequent causes of death. Some causes were natural, for example, stomach cancer, whereas others were unnatural, such as homicide. The authors predicted that unnatural causes with high media coverage were judged to be more frequent than quiet killers like stomach cancer. Their findings matched their predictions: Although stomach cancer is more than five times more frequent than homicide, participants estimated that homicide is about 1.6 times more frequent than stomach cancer. Moreover, media coverage was high for homicides, but zero for stomach cancer, and media coverage predicted the frequency estimates of causes of death. The authors concluded that estimates of frequency of lethal events are based on high availability of vivid or sensational events. Indeed, among the most overestimated causes were sensational events like

tornado, flood, homicide, and motor vehicle accidents. Most causes of death that were underestimated were those not much covered by the media, like asthma, tuberculosis, diabetes, stomach cancer, and heart disease.

This line of research showed that Information A may be more available than Information B because Information A is presented more frequently than Information B (see Chapter 14 on the mere exposure effect). Beyond estimating frequencies, like causes of death, stereotypes can be more available and influence the judgments on a group by repeatedly pairing an attribute (e.g., lazy) with the group (Rothbart, Fulero, Jensen, Howard, & Birrell, 1978). Another source of availability is the vividness of information.

Vividness of information

Estimations of frequency of lethal events are biased because of the disproportionate media coverage of some sensational, but relatively infrequent, events. Thus, two independent features of the information may cause the increase in availability of homicide compared to stomach cancer: (1) Homicides may be more available because instances of death from homicide are covered more frequently in the media than instances of death from stomach cancer, as discussed above. Or (2) homicides can be more available even if not encountered more frequently than stomach cancer because people can imagine violent crimes more vividly than quiet killers. Therefore, frequency of public coverage of an event and its vividness need to be manipulated independently.

Reyes, Thompson, and Bower (1980) were able to show that vividness of presented evidence from a trial affected both its retention and judgments of guilt after a 48-hour delay. The authors presented nine pieces of evidence from the prosecution and nine pieces of evidence from the defense, but for some participants, only the prosecution evidence was vivid, while for the other participants, only the defense evidence was vivid. The participants remembered more evidence of the prosecution and gave higher judgments of the defendant's apparent guilt when the prosecution presented the vivid evidence. In contrast, the participants remembered more evidence of the defense and gave lower judgments of the defendant's apparent guilt when the defense presented vivid evidence. This finding suggests that vividness and imaginability of an instance increase availability of the respective category that, in turn, increases judged frequency (or, in this experiment, judged probability) of occurrence of instances of the category.

There are many examples that vivid cases weigh more than pallid data summaries (see Nisbett & Ross, 1980). For example, in 20 years after the U.S. Surgeon General published a report that linked cigarette smoking to lung cancer, no decline in average cigarette consumption was observed. There was one exception: physicians, especially radiologists. The probability that a physician smokes is directly related to the distance of a physician's specialty from lung disease. It seems that those who diagnose and treat lung cancer daily have vivid illustrations of the dangers of cigarette smoking, while other people just see statistics that do not activate their imagination.

Adopting the perspective of others

If a husband thinks about how much housework his wife does, he has to adopt her perspective. As he does not see all the housework she does, he can try to think as if he were his wife and then estimate her contribution to joint outcomes. As we have already seen, adopting the other's perspective seems to be difficult, as suggested by the fact that spouses overestimate their own contribution to the housework. Both retrieval biases and vividness may contribute to the resulting bias: When estimating the share of the housework, people can retrieve more instances and probably have more vivid memories of their own housework than of the spouse's housework. Another well-known phenomenon that can at least partly be explained by the availability heuristic is *unrealistic optimism* about future life events (Weinstein, 1980). When people judge the chances that positive or negative life events happen to them, they believe that their chances, compared with the chances of their classmates, are higher to experience positive events and lower to experience negative events (cf. Chapter 7 on the illusion of control). Of course, the average of one's chances to experience positive or negative events should equal the chances of the whole group. Therefore, the optimism revealed in Weinstein's study is unrealistic. Among several mechanisms that contribute to this illusion, one is availability that may come into play in two ways: One factor that influences risk assessments is one's own experience. If one has experienced heart disease in his or her family, the risk of heart disease is more available than for someone who has no family history of heart disease. A second factor may be people's difficulties to adopt the perspective of others, comparable to the married partners that overestimated their share of the housework (Ross & Sicoly, 1979). People see their own actions that increase the probabilities for positive events and decrease their probabilities for negative events, but not what others do. If people assess their chances, they may see reasons for why they have better chances, but they may not understand that others also think about such reasons and may arrive at similar conclusions. Therefore, people perceive a difference in chances between themselves and others.

A similar lack of adopting the perspective of others may account for the *fundamental attribution error*, a bias in attributing behaviors of others to dispositional factors (see Nisbett & Ross, 1980). Research has shown that actors are more prone to attribute their behaviors to situational factors than observers of the same actions. For example, if a father yells at his child in the bus, he himself probably thinks that this is because he is tired and his child misbehaved for the third time, whereas observers of his behavior may conclude that he is a loony father. The actor – here the father – explains his behavior with factors that caused him to act as he did, whereas observers explain the father's behavior with his character or disposition. Actors see the situations and act in them, but they do not see themselves acting. For observers, however, the actor is in the focus of the observed action. Observers may see that the child behaves awkwardly, but they do not know that the child misbehaved repeatedly and do not feel how tired the father is.

Interim summary

So far, we have discussed how biased encoding and retrieval or vividness of information has an impact on the availability of information that, in turn, may influence frequency estimates or judgments of apparent guilt. The lack of ability to adopt another's perspective normally results in both more frequent encoding and more vivid memories of one's own actions, leading to overestimation of one's contribution to joint products or to attribution of one's own behavior to situational factors and of another's behavior to dispositional factors.

Availability has been a very popular theoretical framework to explain different phenomena. Part of this appeal, some critics have stated, has come from the vagueness of the term availability (e.g., Betsch & Pohl, 2002): It has been used in a very broad sense, and no process was specified that is unique to availability. It was unclear, for example, whether availability was tied to ease of recall or to amount of recall. Research by Norbert Schwarz and his colleagues addressed this issue (see Schwarz, 1998, and Wänke, 2012, for detailed discussions).

Availability: Ease or amount of recall?

Let us take a closer look at the first of the two basic experiments described above. Tversky and Kahneman (1973) found that people recalled more famous names and judged famous names to be presented more frequently. For example, if names of 19 famous women and 20 non-famous men were presented, participants responded that more names of women were presented. The authors concluded that people used availability – the ease with which they were able to bring instances to mind – as information to judge whether names of men or women were presented more frequently. But note that there is an ambiguity inherent to this finding: When famous names are more available, people can both retrieve them more easily and retrieve more of them. Ease of recall and amount of recall were confounded in this experiment. Thus, there are two alternative possibilities how people can arrive at the conclusion that names of (famous) women were more frequent than (non-famous) men: First, they may have recalled the famous women more easily than the non-famous men, concluding that if it is so easy to recall names of women, there must have been more of them in the list. Alternatively, they might simply have recalled more names of women than of men. From the fact that they have recalled more female names, they may conclude that there must have been more female names in the list. There is no way to resolve this ambiguity in the experiments as done by Tversky and Kahneman.

How can this ambiguity be resolved? Schwarz et al. (1991) used an experimental paradigm that separated ease of recall from amount of recall. They asked people to list six or twelve instances where they behaved self-assertively. In pilot studies, these authors had found that it is relatively easy to recall six instances of self-assertive behaviors, but it is quite difficult to recall twelve such instances. After the participants recalled these behaviors, they were asked how assertive they are. If people base their judgment of self-assertiveness on the experienced ease of recall,

rated assertiveness is expected to be higher after recalling six behaviors than after recalling twelve behaviors. In contrast, if people base their judgment on amount of recall, those who recalled twelve assertive behaviors should judge themselves as being more assertive than those who recalled six behaviors. The results supported the ease of recall view: Participants who listed six behaviors judged themselves to be more assertive than those who listed twelve behaviors. In other experimental conditions, the authors assessed the judgment of assertiveness after participants listed six or twelve instances of unassertive behaviors. The participants again based their judgments on ease of recall and judged themselves to be less assertive after recalling six rather than twelve behaviors. If it was easy to recall six unassertive behaviors, I cannot be assertive after all. The difficulty to recall twelve unassertive behaviors, in contrast, seems to indicate that I am rather assertive.

In the study by Schwarz et al. (1991), availability was related to ease of recall, not to amount of recall. However, do people always base their judgments on ease of recall, or are there instances where availability is better captured in terms of amount of recall? Several conditions have been found to moderate the role of ease of recall, among them the diagnosticity of the recall experience; its relevance for the judgment; the malleability of the judgment; the processing intensity, especially processing motivation; and the salience of the feeling. Finally, the interpretation of ease of recall is moderated by naïve theories of processing intensity, especially processing motivation (see Greifeneder, Bless, & Pham, 2011; Schwarz, 1998, 2004; Wänke, 2012, for reviews). I will discuss these variables in the next sections.

Is the recall experience diagnostic?

Imagine that a participant has recalled six examples of behavior where she behaved assertively. She now concludes from the ease with which she was able to recall these behaviors that she must be self-assertive. In a slightly different experiment, a participant under the same experimental condition listens to music, a new-age piece at half speed. He is told that this music *facilitates* the recall of the behaviors. After recalling six instances, he has to judge how assertive he is. What is the difference to the condition without music? The difference lies in the diagnosticity of the recall experience: The participant in the condition without music normally bases her judgment on ease of recall because she believes that ease of recall tells her something about her assertiveness. The participant who hears music experiences the same amount of ease of recall when he recalls instances of self-assertive behavior, but he believes that the experience of ease is caused by the music. Therefore, he has no reason to base his judgment of assertiveness on the experienced ease of recall. Ease of recall is considered as being undiagnostic as information for the judgment of self-assertiveness. Another participant has to recall six behaviors and hears music, but she is told that music *inhibits* the recall of examples. It is easy to recall six instances of self-assertive behavior, but the music is supposed to make recall difficult. This participant has reason to argue that if it is easy to recall instances of self-assertive behavior despite the inhibiting influence of music, she must be highly

assertive. In this case, ease of recall is considered as being diagnostic information for self-assertiveness. Schwarz et al. (1991) tested this assumption experimentally and indeed found that people used their recall experience only if it was diagnostic or even surprising (see Wänke & Hansen, 2015). If the informational value of the recall experience was undermined because participants could attribute these feelings to the music played to them, they no longer relied on their recall experiences. Further studies could replicate the finding that individuals no longer use ease of recall when the informative value of the experience is diminished (e.g., Raghubir & Menon, 2001; Winkielman, Schwarz, & Belli, 1998).

The relevance of the feeling for the judgment at hand

Several variables have been found to influence the use of feelings for the judgment at hand (see Greifeneder et al., 2011). For example, experts (car mechanics) and non-experts (holders of a driver's license) had to list few or many causes for car breakdowns. When they had to estimate the frequency of car breakdowns, laypeople but not experts were influenced by ease of retrieval (Ofir, 2000). Similarly, people relied on their recall experiences when they were in a positive mood (Ruder & Bless, 2003) or believed in the power of intuition (Keller & Bless, 2008). Finally, powerful individuals relied more on recall experiences than less powerful individuals, presumably because people in power can feel free to make judgments based on feelings while individuals who lack power need to pay attention to multiple social cues in order to get along with those in power (Weick & Guinote, 2008).

The malleability of the judgment

The third criterion is the malleability of the judgment. Let us assume that the recall experience is both informative for the target stimulus and relevant for the judgment at hand, but it might still be that people do not use the experience because they use a different criterion. One example is the direct access of a judgment in memory so that it is no longer malleable and experiences play a minor role.

When people can access a judgment directly, they do not need to rely on recall experiences. For example, people who have thought much about doctor-assisted suicide and are extremely in favor or against it do not need to inspect their feelings to determine how strong their attitude is; they retrieve this information directly. In line with this reasoning, Haddock, Rothman, Reber, and Schwarz (1999) found that experiences of ease of recall influenced judgments of the strength of students' attitude toward doctor-assisted suicide only when the pre-experimentally assessed attitude was not extreme. Those respondents who were strongly in favor or against doctor-assisted suicide did not rely on recall experiences when they judged attitude strength. Processing experiences influenced the participants' judgments about attitude strength only when attitudes were moderate and direct retrieval of information about attitude strength was not possible.

Effects of processing intensity: Processing motivation and processing opportunities

If you have to list behaviors that increase your risk for heart disease and are then asked to estimate your vulnerability for this disease, you can take it easy if there is no history of heart disease in your family, and you probably base your judgment of vulnerability on ease of recall. However, if heart disease has occurred in your family, you probably do not take it easy. You are much more motivated to process this information systematically, for example, by paying attention to the number of risk-increasing behaviors you are able to list. Rothman and Schwarz (1998) explored the consequences of processing motivation in an experiment. Participants had to list either three or eight behaviors that increased or decreased the risk of heart disease. About half of the participants had a family history of heart disease, the others had not. The results are shown in Table 10.1: Participants without a family history of heart disease based their judgments on ease of recall. They judged themselves to be more vulnerable and thought that they needed more urgently to change their behavior if they had to recall either *three* rather than *eight* examples of risk-*in*creasing behaviors or *eight* rather than *three* examples of risk-*de*creasing behaviors. This pattern reversed if there existed a family history of heart disease: These participants relied on the amount of information they listed and therefore judged themselves to be more vulnerable and thought that there was a higher need to change their

TABLE 10.1 Vulnerability to heart disease as a function of type and number of recalled behaviors and family history (adapted from Rothman & Schwarz, 1998)

	Family history of heart disease	
	Without family history	*With family history*
Vulnerability judgments		
Risk-increasing behavior		
3 examples	3.87	4.63
8 examples	3.18	5.37
Risk-decreasing behavior		
3 examples	3.09	5.75
8 examples	4.25	3.75
Need for behavior change		
Risk-increasing behavior		
3 examples	3.37	3.63
8 examples	2.87	6.25
Risk-decreasing behavior		
3 examples	3.00	5.20
8 examples	5.62	4.70

Note: Judgments of vulnerability and the need to change current behavior were made on 9-point scales, with higher values indicating greater vulnerability and need to change, respectively.

behavior if they had to recall either *eight* rather than *three* examples of risk-*in*creasing behaviors or *three* rather than *eight* examples of risk-*de*creasing behaviors. This study demonstrated the effect of processing motivation on the informational implications of processing experience: Participants without a family history of heart disease had a low motivation to examine the processed information and therefore based their judgments on ease of recall. Participants with a history of heart disease, on the other hand, were highly motivated to monitor how many risk-increasing or risk-decreasing behaviors they could list and based their judgments on amount of recall.

Processing intensity does not only depend on processing motivation but also on processing opportunities. The latter could be diminished by two variables. The first is lack of information, which makes it impossible to process relevant knowledge. The second is lack of cognitive resources. Greifeneder and Bless (2007) induced cognitive load in their study by instructing participants to keep an eight-digit number in mind while they formed their judgment. A control group did not have to keep the number in mind. As predicted, participants used ease of retrieval as information to make their judgments when they were under cognitive load and did not have sufficient processing opportunities.

The salience of the feeling

Another prerequisite for the impact of feeling on judgment is the salience of the experience. In a study by Kühnen (2010), participants had first to retrieve two versus eight examples of assertive behaviors and were then asked about their feeling before they had to provide a judgment of self-assertiveness. In order to restrict the processing opportunities, Kühnen applied a cognitive load manipulation, as discussed in the last section. With high cognitive load, the effects were stronger when participants had to gauge their feeling before rather than after the judgment. This experiment shows that the feeling has to be salient in order to influence judgments. If the feeling is in the background, it is less likely to influence the judgments at hand.

Salience of the feeling has not to be confused with salience of the source of the feeling. We have seen in the section on whether the experience is diagnostic that ease of recall is not used to judge a target when the person becomes aware of the source of the experience; as explained earlier, people discount the experience if an alternative source can explain it. In contrast to the salience of the source, the study by Kühnen (2010) showed that people are more likely to use the experience of ease when the feeling itself is in the foreground. Note that the use of the experience decreases when the source becomes salient, but it increases when the feeling itself is salient, without knowledge of its source.

The role of naïve theories

Some work has emphasized the importance of naïve beliefs in linking the experience of ease or difficulty of recall to subjective judgments (see Schwarz, 2004). For example, Winkielman and Schwarz (2001) showed that the same experience of ease

or difficulty in recalling childhood events can lead to opposite judgments, depending on participants' "theory" about the meaning of the subjective experience. Specifically, these researchers first manipulated the recall experience by asking participants to recall few or many childhood events. Then, they manipulated participants' naïve theories about the reason for their specific recall experiences. They told one group of participants that recall can be difficult because pleasant childhood events fade from memory; another group was told that recall can be difficult because unpleasant childhood events fade from memory. As expected, participants reached opposite conclusions about their childhood happiness when the same recall experience was suggested to have different causes: Those participants who experienced easy recall and believed that recall difficulty indicated an unpleasant childhood judged their childhood as more pleasant than those with easy recall and the belief that recall difficulty indicated a pleasant childhood. When recall was difficult, participants who believed that recall difficulty indicated a pleasant childhood judged their childhood to be more pleasant than those who believed that recall difficulty is caused by an unpleasant childhood. These findings show that people use their naïve beliefs to interpret their processing experiences.

From availability to retrieval fluency

The studies by Schwarz and colleagues showed that people rely on ease of retrieval – or retrieval fluency – when they are not motivated or are under time pressure. Else, people use amount of information. Tversky and Kahneman (1973) defined availability as the ease with which relevant instances come to mind, which corresponds to retrieval fluency, which has been the interest of later research in cognitive psychology.

Hertwig, Herzog, Schooler, and Reimer (2008) tested the use of retrieval fluency as a judgmental heuristic. Participants in their study had to choose which of two US-cities, such as San Antonio or Portland, is biggest. When participants recognized both cities, they indeed chose the city that was easier to recall, as measured by recognition latency. This measure of retrieval fluency indicates the relative familiarity of the two cities which is an automatic process.

Yet it would be wrong to think that the fluency heuristic relies on automatic processes alone. Not only familiarity but also controlled retrieval of information, or recollection, increases experienced retrieval fluency (Kurilla & Westerman, 2008). One could therefore think that the fluency heuristic depends both on automatic fluency effects due to having encountered an item – here, a city name – before and on recollection of relevant cues about the city. Based on a model by Hilbig, Erdfelder, and Pohl (2011) that considered both fluency and retrieval of knowledge, Schwikert and Curran (2014) used event-related potentials (ERPs) to distinguish these two contributors to the fluency heuristic. It is well-known that automatic processes related to familiarity lead to an increase in ERP waveforms after around 300 to 500 milliseconds while controlled processes related to recollection lead to an increase after around 500 to 800 milliseconds (see Schwikert & Curran, 2014). If

the fluency heuristic were exclusively based on automatic processes, there should be only one increase in activation after 300 to 500 milliseconds, but if both processes were involved, we should see an increase in activation at both times. Indeed, the authors observed the latter result, suggesting that the fluency heuristic depends on both automatic and controlled processes.

Benjamin et al. (1998) examined effects of retrieval fluency on predictions of learning. Participants in their study had to answer general knowledge questions and had to predict for each question whether they will later remember the answer. Participants predicted that they will remember answers that they could retrieve easily, but in fact they remembered answers best when they had difficulties retrieving them, as measured by response latencies. This yields the paradoxical phenomenon that difficulties experienced at retrieving the answer result in predictions of worse recall when recall in fact is better (see Reber, 2016, for a discussion of practical implications). Another series of experiments showed that learners use retrieval fluency not only in prospective performance judgments but also in retrospective performance estimates (Reber, Meier, Ruch-Monachon, & Tiberini, 2006). Similarly, participants had more confidence in solutions to a problem that were easy to retrieve (Ackerman & Zalmanov, 2012).

While the older studies examined how people draw inferences from frequency estimates based on experienced ease, more recent studies were interested in how fluency of retrieving an instance lead to inferences – frequency estimates were no longer involved. For example, in the Schwarz et al. (1991) study, participants inferred from the ease of retrieving instances of self-assertive behavior that there were many such behaviors in their lives, and they inferred from this frequency estimate that they were self-assertive. In the Ackerman and Zalmanov (2012) study, by contrast, participants did not estimate frequency of occurrence; they merely inferred from the ease with which they could retrieve a problem's solution that this solution must be true.

Conclusions

I have discussed in some detail whether availability as a judgmental basis is better described in terms of ease of recall or of amount of recall. In sum, participants relied on ease of recall, or retrieval fluency, when they thought that experienced ease was diagnostic for their judgmental task, when the information was relevant for the judgmental task, when the judgment was malleable (e.g., because they did not have direct access to relevant information), when individuals were either not motivated or not able to elaborate on a topic, and when the feeling was salient. Moreover, effects of ease of recall may depend on naïve theories people have about the meaning of experienced ease.

How do these findings relate to the frequency judgments assessed in the studies of Tversky and Kahneman (1973) discussed earlier? Can we conclude that their participants estimated the relative frequency of men and women or of word frequencies on the basis of retrieval fluency? The use of the availability heuristic is not the only way people can assess frequency. When people are confronted with

low frequencies, they simply try to count (Brown, 1995). If, for example, respondents in a survey are asked how many times they have eaten caviar in the last two years, most of them probably are able to count the frequency of this event. This means that availability – or ease of recall – may be used only when frequencies are sufficiently high. The findings that relate retrieval fluency to frequency judgments are mixed. While Schwarz et al. (1991) and later Wänke, Schwarz, and Bless (1995) found by applying the misattribution manipulation discussed above that retrieval fluency is related to judged frequency, Sedlmeier et al. (1998) found no evidence for effects of availability on judgments of letter frequencies in German language and concluded that people encode frequency automatically along with information about events. In another study, Reber and Zupanek (2002) manipulated ease of processing at encoding of frequently presented stimuli events and demonstrated an influence of this manipulation on frequency judgments. To sum up, the evidence for or against the use of the availability heuristic in estimating high frequencies is mixed.

The seminal paper by Tversky and Kahneman (1973) has opened a new way of thinking about how frequency judgments are performed, and subsequent research has shown the importance of the availability heuristic in different domains. As an important consequence, phenomena that formerly had been discussed in terms of motivational processes now were explained in terms of cognitive mechanisms. After some ambiguities about the mechanisms underlying the availability heuristic had been disentangled, research revealed different effects of retrieval fluency, and it is easy to think about new research directions that continue Tversky and Kahneman's work on the availability heuristic.

Summary

- Availability is the ease with which relevant instances of a class come to mind.
- Sources of biased availability are biased frequencies, different vividness of information, and the inability to adopt the perspective of another person.
- Availability affects frequency estimates and different kinds of judgments.
- Recent work disentangled the contributions of ease and amount of recall of instances to judgment formation.
- Whether people use ease or amount of recall as information depends on variables such as the perceived diagnosticity of experienced ease; the relevance of the experience for the judgment; malleability of the judgment; processing intensity; and salience of the feeling. Moreover, judgments depend on naïve theories about the meaning of experienced ease.

Further reading

The classical piece on this topic is the article by Tversky and Kahneman (1973) that has been cited over 6,000 times to date (source: Google Scholar). I recommend reading some elegant studies into availability, for example, by Lichtenstein et al. (1978),

Ross and Sicoly (1979), and Schwarz et al. (1991). Wänke (2012) provides a recent review of the literature. For early applications of the availability heuristic, see Nisbett and Ross (1980), for more recent studies, Hertwig et al. (2008) and Schwikert and Curran (2014).

Note

1 In English, the consonants D, K, L, N, R, V, X, and Z appear more frequently in the third than in the first position; the other 12 consonants appear more frequently in the first position (Tversky & Kahneman, 1973; see also Sedlmeier et al., 1998). In German, the letters A, C, E, I, L, N, O, R, and U appear more frequently in the second than in the first position, whereas the letters B, F, G, and S are more frequently in the first than in the second position (Sedlmeier et al., 1998).

References

Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency–confidence association in problem solving. *Psychonomic Bulletin & Review*, *19*, 1187–1192.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*, 55–68.

Betsch, T., & Pohl, D. (2002). Tversky and Kahneman's availability approach to frequency judgement: A critical analysis. In P. Sedlmeier & T. Betsch (Eds.), *Etc. frequency processing and cognition* (pp. 109–119). Oxford: Oxford University Press.

Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1539–1553.

Greifeneder, R., & Bless, H. (2007). Relying on accessible content versus accessibility experiences: The case of processing capacity. *Social Cognition*, *25*, 853–881.

Greifeneder, R., Bless, H., & Pham, M. T. (2011). When do people rely on affective and cognitive feelings in judgment? A review. *Personality and Social Psychology Review*, *15*, 107–141.

Haddock, G., Rothman, A. J., Reber, R., & Schwarz, N. (1999). Forming judgments of attitude certainty, intensity, and importance: The role of subjective experiences. *Personality and Social Psychology Bulletin*, *25*, 771–782.

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1191–1206.

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). New York: Guilford.

Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2011). Fluent, fast, and frugal? A formal model evaluation of the interplay between memory, fluency, and comparative judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 827–839.

Keller, J., & Bless, H. (2008). Predicting future affective states: How ease-of-retrieval and faith in intuition moderate the impact of activated content. *European Journal of Social Psychology*, *38*, 1–10.

Kühnen, U. (2010). Manipulation-checks as manipulation: Another look at the ease-of-retrieval heuristic. *Personality and Social Psychology Bulletin, 36*, 47–58.

Kurilla, B. P., & Westerman, D. L. (2008). Processing fluency affects subjective claims for recollection. *Memory & Cognition*, *36*, 82–92.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 551–578.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.

Ofir, C. (2000). Ease of recall vs recalled evidence in judgment: Experts vs laymen. *Organizational Behavior and Human Decision Processes, 81*(1), 28–42.

Raghubir, P., & Menon, G. (2001). Framing effects in risk perceptions of AIDS. *Marketing Letters*, *12*, 145–155.

Reber, R. (2016). *Critical feeling. How to use feelings strategically*. Cambridge: Cambridge University Press.

Reber, R., Meier, B., Ruch-Monachon, M.-A., & Tiberini, M. (2006). Effects of processing fluency on comparative performance judgments. *Acta Psychologica, 123*, 337–354.

Reber, R., & Zupanek, N. (2002). Effects of processing fluency on estimates of probability and frequency. In P. Sedlmeier & T. Betsch (Eds.), *Etc. frequency processing and cognition* (pp. 175–188). Oxford: Oxford University Press.

Reyes, R. M., Thompson, W. C., & Bower, G. H. (1980). Judgmental biases resulting from differing availabilities of arguments. *Journal of Personality and Social Psychology*, *39*, 2–12.

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology*, *37*, 322–336.

Rothbart, M., Fulero, S., Jensen, C., Howard, J., & Birrell, P. (1978). From individual to group impressions: Availability heuristics in stereotype formation. *Journal of Experimental Social Psychology*, *14*, 237–255.

Rothman, A. J., & Schwarz, N. (1998). Constructing perceptions of vulnerability: Personal relevance and the use of experiential information in health judgments. *Personality and Social Psychology Bulletin*, *24*, 1053–1064.

Ruder, M., & Bless, H. (2003). Mood and the reliance on the ease-of-retrieval heuristic. *Journal of Personality and Social Psychology*, *85*, 20–32.

Schooler, J. W., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112*, 610–611.

Schwarz, N. (1998). Accessible content and accessibility experiences: The interplay of declarative and experiential information in judgment. *Personality and Social Psychology Review*, *2*, 87–99.

Schwarz, N. (2004). Metacognitive 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.

Schwikert, S. R., & Curran, T. (2014). Familiarity and recollection in heuristic decision making. *Journal of Experimental Psychology: General*, *143*, 2341–2365.

Sedlmeier, P., Hertwig, R., & Gigerenzer, G. (1998). Are judgments of the positional frequencies of letters systematically biased due to availability? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 754–770.

Tulving, E. (1983). *Elements in episodic memory*. Oxford: Oxford University Press.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

Wänke, M. (2012). Almost everything you always wanted to know about ease-of-retrieval effects. In R. Greifeneder & C. Unkelbach (Eds.), *The experience of thinking: How the fluency of mental processes influences cognition and behavior* (pp. 151–169). London, England: Psychology Press.

Wänke, M., & Hansen, J. (2015). Relative processing fluency. *Current Directions in Psychological Science, 24*, 195–199.

Wänke, M., Schwarz, N., & Bless, H. (1995). The availability heuristic revisited: Experienced ease of retrieval in mundane frequency estimates. *Acta Psychologica*, *89*, 83–90.

Weick, M., & Guinote, A. (2008). Power increases reliance on experiential knowledge: Evidence from ease-of-retrieval. *Journal of Personality and Social Psychology*, *94*, 956–970.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39*, 806–820.

Winkielman, P., & Schwarz, N. (2001). How pleasant was your childhood? Beliefs about memory shape inferences from experienced difficulty of recall. *Psychological Science*, *12*, 176–179.

Winkielman, P., Schwarz, N., & Belli, R. F. (1998). The role of ease of retrieval and attribution in memory judgments: Judging your memory as worse despite recalling more events. *Psychological Science, 9,* 124–126.

11

JUDGMENTS BY REPRESENTATIVENESS

Karl H. Teigen

Imagine the following two situations:

- Example 1. You observe a person on the sidewalk who appears to be talking to himself. He is alone, but smiling and gesturing. You decide that he is probably crazy.
- Example 2. You take part in a raffle where tickets are numbered from 1 to 100. Someone offers you ticket No 1. You refuse. What about No. 50? You are still not satisfied. You are offered No. 63. You feel much better, and decide to keep the ticket.

These situations are, superficially, quite different from each other. In the first case, you try to explain a person's strange behavior by identifying a category where such behaviors appear to belong. You make a tentative diagnosis. In the second case, you want to pick a lottery ticket that maximizes your (subjective) chances of winning. You make a tentative prediction.

But the situations have also something in common. They are uncertain, and require you to make a guess. In both cases, you are searching for a probable solution. How do we make such guesses? You have no available statistics about the proportion of people talking to themselves who are actually crazy, so how can you conclude "he is probably crazy"? You may have some knowledge about the probabilities involved in a raffle, and even be able to calculate that all probabilities are equal, that is, p (ticket No. 1) = p (ticket No. 50) = p (ticket No. 63) = .01. And yet you do not find this knowledge very helpful because it does not tell you which ticket to accept, and worse still, it does not explain your uneasiness about the first two tickets.

In two famous articles Kahneman and Tversky (1972, 1973) suggested that people in such cases make use of a simple and reasonable mental shortcut to arrive at probability judgments. They simply ask themselves: How much does the target *look*

like a typical instance of the class, category, or parent population under consideration? How *similar* is this individual's behavior to that of a typical crazy person? How *representative* are tickets numbered 1, 50, or 63, as random instances of the ticket population?

Judgments by representativeness, commonly referred to as *the representativeness heuristic*, constitute a most useful way of making probability estimates.

- It is easy, requiring a minimum of cognitive resources.
- It can be used in a number of situations where objective probabilities cannot be calculated (e.g., in singular situations).
- It is often correct. In a unimodal, symmetrical distribution, the central outcome will also be the most frequent one. In many other distributions, including non-ordinal, categorical classifications, the modal outcome is both most probable and most typical. For instance, if I am going to meet a high-ranking military officer, I expect to see a man above 40, rather than a young woman. My stereotype of a "representative" officer corresponds in this case to the actual sex and age distribution of military commanders.

Philosophers from Aristotle to Hume have regarded representativeness, or similarity, as a perfectly legitimate way of estimating probabilities. In the *Rhetoric*, Aristotle lists similarity along with frequency as the basis of sound probability judgments. "If the thing in question *both* happens *oftener* as we represent it and more *as* we represent it, the probability is particularly great" (1941, p. 1433). Even Laplace, one of the founders of modern probability calculus, regarded judgment by analogy to be one of the "diverse means to approach certainty", claiming "the more perfect the similarity, the greater the probability" (1816, p. 263). But Laplace also wrote a chapter on "Illusions in probability estimation", being well aware that our intuitions about probabilities sometimes lead us astray.

In line with this, the representativeness heuristic is not infallible. The "crazy" person in Example 1 may not be crazy, even if he behaves like a typical madman. He could simply be talking in his hands-free mobile phone. The representativeness heuristic may in this case have enticed us to disregard alternative possibilities, and to forget the relative number of madmen compared to cellular-phone users. Similarly, the winning number in Example 2 could equally well be 1 or 50 as 63.

To show that people rely on the representativeness heuristic, rather than upon more rational calculations of frequencies, we need to construct situations in which probability judgments based on representativeness *differ* from judgments based on more normative considerations. In other words, the emphasis will be on errors of judgments, rather than successful judgments. These errors, or biases, often imply that some other, normative factors are neglected or given insufficient weight. The biases can accordingly be described as "base-rate neglect", "insensitivity to sample size" and similar labels, indicating the principles that are violated. It is, however, important to bear in mind that ignoring such principles is a phenomenon that is conceptually distinct from the representativeness heuristic (Bar-Hillel, 1984).

In addition, we need to show that subjective probability judgments and representativeness judgments are highly correlated. For instance, to check whether 63 is a more representative number than 1 or 50, we could ask people how typical these numbers are, as outcomes of a random draw.

Two demonstrations

The two most famous, and most intensely debated, demonstrations of representativeness in the research literature are the Linda problem and the engineers-and-lawyers problem. These are examples of the "conjunction fallacy" and "base-rate neglect", respectively, which are discussed in Chapters 1 and 2 in this volume. Instead, we will focus on two more simple problems, involving people's intuitions about randomness and their sensitivity (or insensitivity) to sample size.

Study 1: Intuitions about random sequences

Consider the problem described in Text box 11.1. When Kahneman and Tversky (1973) asked people to compare alternatives (a) and (b) in the example in Text box 11.1, the majority chose (a), because (b) did not "look" random. Sequence (a) was also preferred to (d), which appears biased and thus not representative of a "fair" coin. The truth is, however, that all series, including (c), are equally likely, with a probability of $.5^6 = .016$. This can be shown by writing all possible sequences of heads and tails; there are altogether 64 such series, each of them occurring once.

TEXT BOX 11.1 INTUITIONS ABOUT RANDOMNESS: PREDICTING CHANCE

Imagine a person tossing a fair coin six times in a row. In every toss, the outcome can be head (H) or tail (T). Which of the following series is most likely?

(a)	H	T	T	H	T	H
(b)	H	H	H	T	T	T
(c)	H	H	H	H	H	H
(d)	H	H	H	H	T	H

In a replication, Smith (1998) gave sets (a)–(c) to a sample of college students, including a fourth option, namely: All rows are equally likely. This (correct) answer was chosen by 40% of the students. However, a majority of 55% still opted for (a). Sequence (b) was chosen by 5%, and nobody chose (c). Smith found a preference for (a) among school children as young as 10 years. The popularity of options (b) and (c) decreased with age, whereas the correct answer, that all rows are equally likely, increased.

Significant differences between (a) and (b), or (a) and (c) can be checked using a sign test. For 20 participants, such differences will be significant ($p < .05$) if 15 (75%) or more prefer (a) to (b), or (a) to (c). Based on Smith's results we may expect at least 90% preference for (a) over (b), which will yield a significant result with 95% probability. If we include "equally likely" as an option, participants choosing this alternative can be excluded from the analysis. To test the difference between participants choosing (a) and "equally likely" makes no sense as no meaningful null hypothesis can be formed. (We don't need a significant majority of errors to identify a bias.)

Explanation

If we think of a fair coin as a device that can produce two equally likely outcomes, a representative series of outcomes should contain an approximate equal number of heads and tails. Three heads and three tails are viewed as more representative than a series containing more heads than tails. This makes (c) and (d) more unlikely than (a) or (b). But why should (a) be preferred to (b)? Two explanations, both invoking the concept of representativeness, are possible.

1 In a sequence of random tosses, the outcomes should not only be representative for the coin, but also for the process by which the outcomes are produced. Random outcomes should be *typically* random. They should *look* random. The sequence H-H-H-T-T-T looks too regular, whereas H-T-T-H-T-H has the proper random look.

   This explanation presupposes a lay theory about randomness. For most of us, this theory has two parts. First, random outcomes should balance each other out (there should be approximately an equal number of each). This requirement is violated by sequence (c) and (d). Second, random outcomes should not look orderly. This is the criterion that sequence (b) fails to meet.

2 A slightly different interpretation invokes the concept of *local* representativeness. A random sequence should not only be globally representative (i.e., for the series taken as a whole), we also expect each *part* of the series to reflect the balance between heads and tails, leading to frequent alternations and few runs. H-H-H-T-T-T consists of two biased parts, whereas H-T-T-H-T-H can be divided into three balanced pairs.

   We can extend the coin experiment by asking participants to predict the result of a *seventh* throw. Suppose that series (c) has actually occurred by chance. What will happen next? According to the so-called *gambler's fallacy* (Gold & Hester, 2008; Sundali & Croson, 2006) people will be more ready for a tail than for a seventh head. Also in series (b), we may feel (although less strongly) that head is due. These predictions are also in line with the notion of representativeness: A seventh head will make the sequence still less

representative, whereas a tail will contribute to a more balanced, and hence more "likely" pattern.

In a study by Holzworth and Doherty (1974) participants were shown random series of nine cards, colored black or white, and asked to predict the color of the next card. When the cards were drawn from a 90/10 distribution of black and white cards, the participants predicted (correctly) black cards. But when they believed that the cards came from a 70/30 distribution, they predicted *white* cards on almost 30% of the trials (see Chapter 3 in this volume). This is normatively nonsensical, because black cards have a better chance than white ones to be drawn in all trials, as long as the deck contains more black than white cards. It makes more sense from a representativeness point of view. Some of the displayed series may have included "too many" blacks. After drawing, say, four black cards in a row, one might think that a white card would make the sample more representative of a 70/30 distribution.

Study 2: Intuitions about sample sizes

Consider the problem described in Text box 11.2 (from Kahneman & Tversky, 1972, 1973). In a group of undergraduate college students, 56% answered alternative (c), the rest were equally divided between (a) and (b). Thus there was no general preference for either hospital, despite the fact that a large sample is much less likely to be biased than a small sample. That is precisely why investigators prefer large samples! In fact, the 'true' probability for random samples of size 15 to have 9 (60%) or more male babies is about .30, whereas the probability for samples of size 45 to have 27 (60%) or more male babies is about .12 (according to the normal-curve approximation of a binomial test). That is, normatively, (b) is the correct answer.

TEXT BOX 11.2 INTUITIONS ABOUT SAMPLE SIZE: FROM POPULATION TO SAMPLES

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower. For a period of one year, each hospital recorded the days on which more than 60 percent of the babies born were boys. Which hospital do you think recorded more such days?

a The larger hospital
b The smaller hospital
c About the same (that is, within 5 percent of each other)

This demonstration is in no need of a statistical test, because no null hypothesis can be meaningfully formed. The question is rather whether errors, that is, answers of type (a) and (c), are common, rare, or nonexistent. The number of such errors has shown to be highly variable, depending upon the way the question is asked (Sedlmeier & Gigerenzer, 1997).

Explanation

Instead of obeying the "law of large numbers" of probability theory, many students seem to think that equally biased samples are equally probable, regardless of *n*. The more biased, the less probable. This follows from a unique reliance on the representativeness heuristic.

Prediction versus diagnosis

In both demonstrations, people were asked to predict the occurrences of specific outcomes, which they seemed to do by asking themselves: How well do these outcomes match salient features of the populations from which they are drawn? The better the match, the higher the outcome probability.

However, a match can go both ways. In prediction tasks, we know some features of the population (or think we do), and ask for a matching sample. In other cases, we go from a given sample in search of a matching population. Do people use the representativeness heuristic in both cases? And should they?

Normatively, the probability of a sample, given the population, and the probability of a population, given a sample, are not the same. After a course in inferential statistics, most students know how to calculate $p(\text{Data}\,|\,H_o)$, that is, they can predict the probability of a particular outcome to occur by chance. They may also have learned that this probability is not identical to $p(H_o\,|\,\text{Data})$, or the probability that the null hypothesis is correct, given the results. Yet, these probabilities are often confused; even scientists who should know better sometimes refer to the probabilities involved in significance testing as probabilities for the null hypothesis, rather than probabilities for data, given H_0. A legitimate transition from $p(D\,|\,H_0)$ to $p(H_0\,|\,D)$ requires additional knowledge of the prior probability of H_0 and H_1, and of the compatibility of data to H_1, as dictated by Bayes' theorem. The tendency to confuse $p(D\,|\,H)$ and the inverse probability $p(H\,|\,D)$ is a very common error of probabilistic reasoning, which has been labeled the *inverse fallacy* (Villejoubert & Mandel, 2002).

Representativeness was originally described as a prediction rule, leading from a population (a distribution, or a category) to a sample (an instance, or an event). According to Kahneman and Tversky, the representativeness heuristic is used when "the probability of an uncertain event, or a sample, is evaluated by the degree to which it is: (i) similar in essential properties to its parent population; and (ii) reflects the salient features of the process by which it is generated" (1972, p. 431).

In practice, however, representativeness was soon applied to the inverse relationship (cf. Chapter 2). In the engineer-and-lawyer-problem, the cab problem, and

several other problems introduced by Kahneman and Tversky (1973), participants were asked to estimate the probability that a particular individual, or event, belonged to a certain category; in other words, they were invited to make an inference from event to category, or from sample to population, rather than the other way around. Such inferences should perhaps more accurately be termed problems of diagnosis than problems of prediction.

Are such inferences in everyday life also accomplished by a simple matching process, making representativeness a general heuristic for both kinds of probability judgments? To examine this question let us revisit our two demonstration cases, this time from the point of view of diagnosis rather than prediction.

Diagnosing randomness

A diagnosis version of the random sequence problem can be arranged by asking participants to estimate the probability that a particular sequence was actually produced by chance. Consider the example in Text box 11.3 (adapted from a study by Ayton & Wright, 1987). Following an "inverted" variant of the representativeness heuristic, we would believe that Alice in Text box 11.3 used a coin, because her sequence looks like a typical random sequence, whereas Susan probably cheated, because her sequence looks arranged.

TEXT BOX 11.3 INTUITIONS ABOUT RANDOMNESS: DIAGNOSING CHANCE

Alice and Susan are asked by their teacher to produce a sequence of heads and tails by tossing a coin six times. One of the girls does as she is told, whereas the other skips the coin and simply invents a sequence. Here are the results:

Alice:	H	T	T	H	T	H
Susan:	H	H	H	T	T	T

What is, in your opinion, more likely:

a Alice used a coin.
b Susan used a coin.

This problem has, however, no single normative answer. We cannot even say that both answers are equally likely, because the true probabilities do not only depend upon what kind of sequences can occur by chance, but also what kind of sequences will occur by cheating, which in turn depends upon how sophisticated a cheater you are! Susan's sequence looks arranged, but a cheater who tries to mimic chance would hardly produce such a sequence. So perhaps she is not the cheater after all.

Diagnoses based on sample size

A diagnosis version of the birth-clinic problem is more difficult to arrange, because most people will have strong a priori reasons to believe that there are around 50% male newborn babies (or slightly more) in the population. We could, however, make a slight change in content, as illustrated by the example in Text box 11.4. If we find in this case that answer (b) is preferred to answer (a), we can infer that people have more confidence in a greater than a smaller sample (as indeed they should). This should lead us to modify the previous conclusion that representativeness makes people "insensitive" to sample size. Prediction probabilities and diagnostic probabilities need not be the same.

TEXT BOX 11.4 INTUITIONS ABOUT SAMPLE SIZE: FROM SAMPLES TO POPULATION

Two teachers want to find out whether there are more male or female social science students at the university. One of them checks a small class of 15 students, finding 9 (60%) male and 6 female students. The other one studies a larger class of 45, finding 18 (40%) male and 27 female students. What is more probable: There are altogether (a) more male students; (b) more female students; (c) an equal number of male and female students.

The problems in Text boxes 11.3 and 11.4 differ in other ways as well. In the original birth-clinic problem, participants were asked to compare frequencies (number of days in the tail of a distribution); here, they are asked to compare probabilities. This has been shown to make a difference, in addition to other problem features as sample size and skew (for an overview, see Lem, Dooren, Gillard, & Verschaffel, 2011). What if the small class had only 5 students (80% males), and the large class contained 500 students (80% females)? In this case, sample size would no longer be neglected.

Representativeness as a general-purpose heuristic

The representativeness heuristic has been used to explain a number of judgmental biases. We have discussed the gambler's fallacy and insensitivity to sample size as examples of errors of prediction. In diagnosis problems, the representativeness heuristic has been held responsible for *base-rate neglect* (or rather insufficient weight of base rates). This theme is explored in detail in Chapter 2 in the present volume.

Tversky and Kahneman (1983) further argued that representativeness can lead to the fallacious belief that a combination of one likely and one unlikely event is more likely than the unlikely event taken by itself. This so-called *conjunction fallacy* is treated in this book in Chapter 1. For instance, people thought that

it was likely that Bjorn Borg would win the Wimbledon tennis final, because it looked like a typical thing for a champion like Borg to do. They thought it would be rather unlikely for him to lose the first set of the match. This would be less typical of Borg. The conjunction "losing the first set but winning the match" contains a combination of typical and less typical elements. This conjunction was believed by many participants to have an *intermediate* probability, rather than the even lower probability that follows logically from the combination of high and low *p* events.

Kahneman and Tversky (1973) also showed that use of the representativeness heuristic could lead to *non-regressive predictions.* It has been known since the time of Francis Galton that use of imperfect predictors should lead to less extreme predictions. Extremely tall parents will have tall offspring, but since the heights of parents and offspring are not perfectly correlated, we should expect these children to be, on the average, somewhat shorter than their parents; conversely, children of exceptionally short parents should be in general taller than their parents. Filial regression was, in fact, already observed by Homer, in a passage of the *Odyssey*: "Few are the sons that are like their father in breed; The most part are worse, scarce any their fathers excel" (Homer's *Odyssey*, 1953, Book 2, Verse 277–78, S. O. Andrew's translation). Being exclusively concerned with the superior part of the distribution, Homer failed to comment on the complementary fact that inferior fathers often have sons of a more hopeful breed. Evidently, Homer felt sons to be less representative of their illustrious origins than they ought to have been, documenting a very early instance of the representativeness heuristic failing to square with the facts.

It further follows from the concept of statistical regression, that when a measure is not perfectly reliable, we must expect the top scorers on one occasion to be distributed somewhat closer to the mean on the second occasion (and vice versa). From the point of view of representativeness, however, a typical top scorer should continue to excel, and an individual scoring in the 75th percentile should remain around the 75th percentile on the second occasion also. A drop in performance would accordingly be attributed to change or some other systematic process, rather than to chance. Kahneman and Tversky (1973) told the story about a flight instructor who used to praise students who had performed exceptionally well, only to find that, as a rule, they performed worse on the next occasion. Instead of realizing that performances are not completely reliable indicators of skill, and thus bound to regress for purely statistical reasons, he felt forced to conclude that praise has a negative rather than the intended positive effect.

A corollary of the problem of non-regressive predictions is that people tend to make the same predictions based on invalid measures as they would do on more reliable and valid ones. So, for instance, when two groups of participants were asked to predict the grades of hypothetical students based on their relative standing (percentile scores) on (a) a grade-point average scale, or (b) a mental concentration test, they produced in both cases almost identical, non-regressive predictions (Kahneman & Tversky, 1973). In other words, they appeared to use the less valid and reliable mental concentration test with the same confidence as a perfectly valid predictor.

Extreme predictions based on invalid predictors have been described as manifestations of an *illusion of validity* (Kahneman & Tversky, 1973).

The representativeness heuristic has over the years been applied to an increasing range of phenomena in the field of judgment and decision making. It has been proclaimed to be "perhaps our most basic heuristic" (Fiske & Taylor, 2013, p. 181). One of its attractions has been that it seems to be applicable also to expert judgments in a variety of fields. Another is its link to the area of causality judgments.

Expert judgments

In their very first paper of judgmental biases, Tversky and Kahneman (1971) showed that even scientists with a solid background in statistics place too much confidence in the results of small samples. They presented a questionnaire to a group of mathematical psychologists, asking for what kind of advice they would give a PhD student who just has performed two small-scale, inconclusive experiments (one barely significant and the other not). Many respondents thought it would be a good idea to speculate about the *difference* between the results (which could have been a statistical artifact). The majority thought that the experiment should be repeated a third time, again with a small sample (which could not be expected to reach significance). Despite their theoretical knowledge of sampling distributions and statistical hypothesis testing, these experts seemed to suppose that small samples are highly representative of their populations, apparently believing in a "law of small numbers" (as a proxy for the well-known "law of large numbers" in statistical theory).

Domain expertise can, however, sometimes counteract some of the more extreme biases due to representativeness thinking. For instance, experience with the ups and downs on the stock market could make the predictions of a professional investor more regressive than those of a novice. Yet even a real-world economic market may be biased by the power of representative predictions, manifested as overconfidence in stocks, firms, or football teams that have a recent history of good performance (Tassoni, 1996; see Chapter 16 in this volume). Moreover, risky stocks tend to be undervalued, and safe stocks overvalued, by representativeness reasoning: Safe stocks come from good companies, and investment in good companies should give good returns, that is, investors assume a match between the company and stock quality, making safe stocks attractive even when they are costly (Shefrin, 2001). Similarly, extrapolations of continued price growth (eventually creating price bubbles), and over- and under-reactions to good and bad news (prior to financial crises) have been attributed to representativeness by behavioral economists (Gennaioli, Shleifer, & Vishny, 2015).

Clinical judgments offer rich possibilities for studying diagnoses as well as predictions. Garb (1996) gave clinical psychologists a case description satisfying the DSM-IIIR criteria for antisocial personality disorder. They were then asked to rate (a) the likelihood for five possible diagnoses, as well as (b) the degree to which the case was similar to the "typical" person with these disorders. Only

27% of the clinicians made the "correct" diagnosis (according to the manual). The correlation between probability judgments (a) and representativeness judgments (b) was extremely high, $r = .97$, indicating that the clinicians used similarity to a prototype rather than a list of criteria to arrive at a diagnosis.

Causality judgments

John Stuart Mill (1856) observed that people, including philosophers, tend to assume a correspondence between cause and effects. Like begets like. Large effects prompt us to look for large causes. Good effects are attributed to good causes, whereas disasters and human suffering must be due to evil forces. While this is in general a sound heuristic – large objects make in general louder noises than smaller objects, and nice people often make us feel good – exceptions are not difficult to find (small whistles can be deafening, and nice people can be boring). The similarity between Mill's correspondence principle and the representativeness heuristic has made many investigators think that judgments by representativeness also apply to judgments of causation.

Again, these inferences may go both ways: from known causes to hypothetical effects and from known effects to hypothetical causes. We may for instance expect an acknowledged expert to be a source of valid and reliable information. Informed people (causes) should produce informative statements (effects) matching their level of expertise. Unfortunately, experts can be wrong, particularly outside their field of expertise. Even more risky, we may infer the expertise of the speaker from the confidence and specificity of his assertions. Unfortunately, a specific prediction will be more easily disconfirmed than the vague one, leading to a "preciseness paradox" (Teigen, 1990), where the speaker has to choose between being believed (by sounding like an expert) and being correct (by using more general and approximate terms).

If causes correspond to effects, we should expect people to prefer causes whose salient features match the salient features of the events-to-be-explained. Lupfer and Layman (1996) found that people favor religious explanations of uncontrollable events with life-altering outcomes, whereas they prefer naturalistic explanations for controllable events, and events with more mundane consequences. In each case the religious attributions were made in agreement with characteristics believed to be "representative" for supernatural versus natural sources of causality.

Gavanski and Wells (1989) suggested that representative causes also apply to hypothetical, counterfactual outcomes. For instance, when we think how an *exceptional* outcome could have been prevented, we focus on *exceptional* antecedents, whereas we change *normal* outcomes by changing a *normal* antecedent. Causes, or antecedents, are supposed to match outcomes also in magnitude (Sim & Morris, 1998). If an athlete makes a poor overall performance in a triathlon contest, we will blame the failure on her worst rather than on her average or best exercise, even if they all could, in principle, have been improved.

Representativeness, or similarity reasoning, may play a part in scientific theories as well.

- A stutterer behaves in some respects similar to a nervous person, and may indeed be anxious about not being able to communicate. This has suggested anxiety as an etiologic factor in some theories about stuttering (Attanasio, Onslow, & Packman, 1998).
- When children show few signs of empathy and social interest, a corresponding lack of empathy and interest on the part of their caregivers looks like a plausible cause. Thus childhood autism, with its remarkable impairment of reciprocal social interaction, was for many years believed to be due to inadequate mothering.

In both these cases, representativeness reasoning suggested a false lead. But there are probably many more cases where the same line of reasoning provides valuable hints. For instance, violent and abusive adults have themselves often been abused by their parents. Violence breeds violence. This looks like a similarity inference, but it is also a truth.

Representativeness broadly defined

If representativeness applies to all the cases we have listed in this chapter, the original definition (Kahneman & Tversky, 1972, see above) appears too narrow. A more general formulation was suggested by Tversky and Kahneman (1982, p. 85): "Representativeness is a relation between a process or a model, M, and some instance or event, X, associated with that model", as in the following four basic cases:

- M is a class and X is a value of a variable defined in this class (e.g., X could be the typical income of college professors)
- M is a class and X is an instance of that class (e.g., X is regarded to be a "representative" American writer)
- M is a class and X is a subset of M (e.g., X is a "representative" sample of the US population)
- M is a causal system and X is a possible consequence.

"In summary, a relation of representativeness can be defined for (1) a value and a distribution, (2) an instance and a category, (3) a sample and a population, (4) an effect and a cause. In all four cases, representativeness expresses the degree of correspondence between X and M" (Tversky & Kahneman, 1982, p. 87). This correspondence can be based on statistical beliefs (as in 1), causal beliefs (as in 4), and perceived similarity (as in 2 and 3). When this correspondence has been empirically established, for example by asking people to judge which of two events, X_1 or X_2, is more representative of M, we would expect probability judgments to be influenced by the representativeness relation. If X_1 is regarded as more representative than X_2, it will appear to be more likely.

Criticisms

The concept of a representativeness heuristic, as well as the biases it was supposed to explain, have often been challenged. Some of the main criticisms are summarized below.

Conceptual vagueness

Representativeness is a very broad concept, applicable to a number of situations. This generality makes it both imprecise and difficult to falsify. It covers concepts like typicality, similarity, correspondence, and match, which in themselves are open to a variety of interpretations. For instance, in the birth-clinic example, participants are assumed to think that samples are equally representative if they deviate equally from 50%. But people may also think that it is "typical" for small samples to produce variable results, and thus arrive at the correct answer. In contrast, some may even think that "large" samples can produce "large" variations, based on a misplaced assumption of correspondence in terms of magnitude.

Gigerenzer, the strongest critic of the heuristics-and-biases program, is not impressed by terms like representativeness, availability, and anchoring: "These one-word labels at once explain too little and too much: too little, because the underlying processes are left unspecified, and too much, because, with sufficient imagination, one of them can be fit to almost any empirical result post hoc" (Gigerenzer et al., 1999, p. 28). This is a serious criticism if we expect a full-fledged theory capable of modeling and predicting human judgments with a high degree of accuracy. However, representativeness was originally proposed as a more descriptive term, capable of elucidating some general characteristics of human reasoning under uncertainty. The concluding section of the present chapter presents some recent speculations about the nature of the "underlying processes".

Biases can be reduced

Not all studies show equally strong effects of representativeness. Moreover, in all studies there will be a substantial number of individual participants who appear less susceptible to representativeness reasoning. For instance, in the random-sequence experiment, many participants will say (correctly) that all sequences have the same probability of occurrence.

Such differences can be attributed to a variety of sources. One is situational transparency. A concrete situation, in which procedures and mechanisms are clearly visible, will increase the chances of a normative response. Within-subjects studies, in which participants are asked directly to compare the alternatives, will typically yield more normative answers than between-subjects designs, in which the focal variables are more disguised. A group in a between-subjects design, who is only shown the sequence H-T-T-H-T-H will probably characterize it as more likely than participants in another group who are asked to characterize the sequence

H-H-H-T-T-T, whereas individual participants who are asked to compare both sequences may "know" that they are equally likely.

People's use of heuristics is also influenced by their degree of statistical sophistication, ability differences (Stanovich & West, 2000), and more generally whether the task is conceived as a problem that should be solved by mathematical reasoning or simply by "gut feelings". One study showed that people relied less on the representativeness heuristic when frowning, presumably because the instructions to furrow their brows would serve as a cue that more cognitive processing was needed for answering the questions (Alter, Oppenheimer, Epley, & Eyre, 2007).

Probabilities versus frequencies

Problems can sometimes be made more concrete and transparent by translating probabilities into frequencies. Some evidence suggests, indeed, that people reason more normatively with natural-frequency formats (Gigerenzer, 1991). But despite claims to the contrary, the judgment "illusions" do not disappear. In several of the original demonstrations (including those presented in the first section of the present chapter) participants were in fact asked about frequencies.

Even so, it has been suggested that the representativeness heuristic is especially well suited for unique events, whereas the availability heuristic (Chapter 10) is more applicable to frequentistic probabilities (Jones, Jones, & Frisch, 1995). Frequency theorists, who believe that probabilities can only be meaningfully assigned to repeated events, have argued that probability judgments by representativeness cannot be given a mathematical interpretation, but invoke instead a credibility or plausibility concept (Hertwig & Gigerenzer, 1999).

Biases are not errors

Some critics have argued that when people appear biased, it is not because they commit errors of judgment, but because the norms do not apply. People may have been asked ambiguous questions, where a particular answer appears incorrect given a literal interpretation of the task, but might be justified given a more pragmatic interpretation. For instance, a question about the likelihood of the H-T-T-H-T-H sequence may be interpreted as a question about a sequence of "this type" (with alternating Hs and Ts) rather than about exactly this sequence. Conjunction tasks and base-rate tasks have similarly been given pragmatic interpretations that make "conjunction errors" and "base-rate neglect" less fallacious than they originally appeared (see Chapters 1 and 2).

A problem with this criticism is that it is typically raised post hoc (when the results are known) and often assumes that the participants are able to draw very fine distinctions in their interpretation of questions. Indeed, the participants are sometimes attributed a more sophisticated grasp of probability theory than the experimenters.

Alternative explanations

Not all the judgment "illusions" that have been attributed to the representativeness heuristic may, in fact, be due to it. The conjunction fallacy may in some cases be due to a misplaced averaging rule, or by judgments of surprise, as discussed in Chapter 1. Base-rate neglect, as discussed in Chapter 2, could sometimes be due to inversion errors (Villejoubert & Mandel, 2002). Similarly, the gambler's fallacy may be due to more magical "balancing beliefs", in addition to similarity judgments (Joram & Read, 1996). Finally, when middle numbers in a lottery are preferred to extreme numbers (Teigen, 1983), it could be due to representativeness, but it could also signify a preference for small errors over large ones (with ticket No. 1, one could be very wide off the mark).

Representativeness revisited

In a later article, Kahneman and Frederick (2002) offered a wider framework for heuristic judgments. In their view, representativeness illustrates a general feature of intuitive reasoning, where people solve a difficult task (estimation of probabilities) by transforming it into a simpler task (here: judgments of similarity). This can be described as a process of *attribute substitution*. A jury member who is asked to evaluate the probability that the defendant is telling the truth (the target attribute) may instead be performing the much easier evaluation: "How well did he answer the prosecutor's questions" (the heuristic attribute). There often is a valid link between these two attributes; convincing answers may be correlated with actual truth telling. But if the heuristic attribute is given too much credit, biased judgments ensue. A jury member who relies exclusively on his gut feelings may decide issues of guilt on the basis of credibility judgments rather than upon evidence.

Representativeness reasoning refers, by this account, to two processes:

- A judgment of what is the prototypical, or "representative" exemplar of a category, a population, or a distribution (judgment *of* representativeness).
- A probability judgment based on how similar a target outcome is to this prototype (judgments *by* representativeness).

In some problems, the first assessment is already implied by the instructions; for instance, in the birth-clinic problem, participants were told that typically around 50% of the babies are boys. In other problems, participants have to make their own typicality judgments, based on previous beliefs. For instance, a player with some coin-tossing experience may be less convinced that H-T-T-H-T-H is a prototypical random sequence; perhaps it contains too many alternations to be truly "representative". Thus, different individuals might arrive at different probability judgments simply by having different opinions about the prototypical chance outcome (or the prototypical engineer, or the prototypical bank teller, as the case may be). Experts could make better probability estimates than novices, not by relying less

on representativeness, but by having developed a more differentiated and accurate lexicon of prototypes.

Heuristic judgments are often described as quick, intuitive, effortless, and automatic. This means that they are hard to avoid, yet they do not have to be accepted. If I observe a fellow bus passenger in Oslo (Norway) with a striking similarity to the Russian president, the thought of Putin himself is unavoidable, but I will quickly convince myself that despite the similarity, the probability of Putin's twin brother riding the local bus is essentially zero. Similarly, statistical knowledge and logical arguments may convince me that large biased samples occur less frequently than small biased samples, that H-H-H-H-H-H is a perfectly acceptable random sequence, and that people apparently talking to themselves are not necessarily crazy, given the popularity of hands-free mobile phones. These "corrections" are usually due to more deliberate, reflective, and analytic afterthoughts that follow, and sometimes supersede, our initial, spontaneous intuitions.

Leaning on currently popular dual-process models, Kahneman and Frederick (2002), distinguished between the operations of two cognitive systems, System 1 and System 2. System 1 is exemplified by intuitive and spontaneous heuristic processing, whereas System 2 refers to our capacity for reflective, controlled, critical, and effortful thinking, where judgments can be evaluated according to rational rules. System 2 will monitor and control the output of System 1, with the implication that judgments by representativeness (as well as other heuristic judgments) are only expressed overtly if endorsed by System 2 processes. We may accordingly think of probability judgments as a compromise between simple and effortless spontaneous processes, on one hand, and more slow and careful review and revise procedures, on the other. Whether, in the end, the judgments will be biased or not depends upon the appropriateness of intuitive thinking, as well as the weight allotted to it by System 2.

The distinction between these two "systems", recently popularized in Kahneman's (2011) international best-seller "Thinking, fast and slow", strikes many as overly schematic, introducing two almost mystical entities with a life of their own (for a critique, see Keren & Schul, 2009). Kahneman himself hastens to "make absolutely clear that they are fictitious characters" (2011, p. 29), and yet gives them a central role as explanatory concepts. For those who prefer more mundane descriptions, we may think of cognitive heuristics, including representativeness, as ways of simplifying judgment and decision making, by relying on single cues that are easy to observe and require a minimum of effort to process (Shah & Oppenheimer, 2008). Sometimes we simplify too much, or in the wrong way.

Summary

- Representativeness is not in itself a bias (or an illusion), but a procedure for estimating probabilities by means of similarity or typicality judgments. Such judgments are often accurate, but will occasionally lead to biased estimates.
- Representativeness can be used to assess the probability of a particular outcome, based on its similarity with its source or its "parent population" (prediction tasks).

- It can also be used to assess the probability of a hypothesis, or a causal model, based on its match with a set of observations (diagnosis tasks).
- Representativeness was originally described as one of three basic heuristics by Kahneman and Tversky (1972, 1973). An over-reliance on representativeness has been used to explain a number of biases, including the conjunction fallacy, base-rate neglect, the gambler's fallacy, belief in "the law of small numbers", non-regressive predictions, and the illusion of validity.
- Representativeness has been studied both in lay and expert judgments and is related to beliefs in the similarity between causes and consequences.
- Many biases originally described as "due" to representativeness can also be given alternative explanations.

Further reading

Tversky and Kahneman's (1974) classic paper is still the best introduction to their early work on representativeness and other judgmental heuristics. A more thorough conceptual analysis of representativeness is provided by Tversky and Kahneman (1982), and a revised and updated version by Kahneman and Frederick (2002). A historical overview of the representativeness heuristic can be found in a handbook chapter by Griffin, Gonzalez, Koehler, and Gilovic (2012).

References

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.

Aristotle (1941). Rhetoric. In R. McKeon (Ed.), *The basic work of Aristotle* (pp. 1325–1454). New York: Random House.

Attanasio, J. S., Onslow, M., & Packman, A. (1998). Representativeness reasoning and the search for the origins of stuttering: A return to basic observations. *Journal of Fluency Disorders*, *23*, 265–277.

Ayton, P., & Wright, G. (1987). Tests for randomness? *Teaching Mathematics and its Applications, 6,* 83–87.

Bar-Hillel, M. (1984). Representativeness and fallacies of probability judgment. *Acta Psychologica*, *55*, 91–107.

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition – from brains to culture* (2nd ed.). London: SAGE.

Garb, H. N. (1996). The representativeness and past-behavior heuristics in clinical judgment. *Professional Psychology: Research and Practice*, *27*, 272–277.

Gavanski, I., & Wells, G. L. (1989). Counterfactual processing of normal and exceptional events. *Journal of Experimental Social Psychology*, *25*, 314–325.

Gennaioli, N., Shleifer, A., & Vishny, R. (2015). Neglected risks: The psychology of financial crises. *American Economic Review, 105*, 310–314.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology*, *2*, 83–115.

Gigerenzer, G., Todd, P. M., & The ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Gold, E., & Hester, G. (2008). The gambler's fallacy and the coin's memory. In J. I. Krueger (Ed.), *Rationality and social responsibility: Essays in honor of Robyn Mason Dawes* (pp. 31–46). New York: Psychology Press.

Griffin, D., Gonzalez, R., Koehler, D., & Gilovich, T. (2012). Judgmental heuristics: A historical overview. In K. Holyoak & R. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 322–345). Oxford: Oxford University Press.

Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*, 275–305.

Holzworth, R. J., & Doherty, M. E. (1974). Inferences and predictions: Normative vs. representative responding. *Bulletin of the Psychonomic Society*, *3,* 300–302.

Homer (1953). *Homer's Odyssey*. London: Dent.

Jones, S. K., Jones, K. T., & Frisch, D. (1995). Biases of probability assessment: A comparison of frequency and single-case judgments. *Organizational Behavior and Human Decision Processes*, *61*, 109–122.

Joram, E., & Read, D. (1996). Two faces of representativeness: The effects of response format on beliefs about random sampling. *Journal of Behavioral Decision Making*, *9,* 249–264.

Kahneman, D. (2011). *Thinking, fast and slow*. London: Penguin Books.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribution substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3,* 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science, 4*, 500–533.

Laplace, P. S. (1816). *Essai philosophique sur les probabilités*. Paris: Courcier.

Lem, S., Dooren, W. V., Gillard, E., & Verschaffel, L. (2011). Sample size neglect problems: A critical analysis. *Studia Psychologica, 53,* 123–135.

Lupfer, M. B., & Layman, E. (1996). Invoking naturalistic and religious attributions: A case of applying the availability heuristic? The representativeness heuristic? *Social Cognition*, *14*, 55–76.

Mill, J. S. (1856). *A system of logic*. London: Parker.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making*, *10*, 33–51.

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin, 134*, 207–222.

Shefrin, H. (2001). Do investors expect higher returns from safer stocks than from riskier stocks? *The Journal of Psychology and Financial Markets*, *2*, 176–181.

Sim, D. L. H., & Morris, M. W. (1998). Representativeness and counterfactual thinking: The principle that antecedent and outcome correspond in magnitude. *Personality and Social Psychology Bulletin*, *24*, 595–609.

Smith, H. D. (1998). Misconceptions of chance: Developmental differences and similarities in use of the representativeness heuristic. *Psychological Reports*, *83*, 703–707.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, *23,* 645–665.

Sundali, J., & Croson, R. (2006). Biases in casino betting: The hot hand and the gambler's fallacy. *Judgment and Decision Making, 1,* 1–12.

Tassoni, C. J. (1996). Representativeness in the market for bets on national football league games. *Journal of Behavioral Decision Making*, *9,* 115–124.

Teigen, K. H. (1983). Studies in subjective probability I: Predictions of random events. *Scandinavian Journal of Psychology*, *24*, 13–25.

Teigen, K. H. (1990). To be convincing or to be right: A question of preciseness. In K. Gilhooly, M. Keane, R. Logan, & G. Erdos (Eds.), *Lines of thinking: Reflections on the psychology of thought* (pp. 299–313). Chichester, UK: Wiley.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.

Tversky, A., & Kahneman, D. (1974). Judgments under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84–98). Cambridge: Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory & Cognition*, *30*, 171–178.

12

ANCHORING EFFECT

Štěpán Bahník, Birte Englich, and Fritz Strack

Imagine you are the judge in a legal case of rape. The prosecutor and the defense attorney have given their final speeches and the court hearing has just been interrupted for lunch. Thus you have roughly an hour to make up your mind about the sentence. All the necessary information is right in front of you. Once again, you go through the most important facts: The victim's account of what happened that night, the expert's assessment of how likely it is that the defendant will commit rape again, the prosecutor's and the attorney's pleas. Upon close inspection, the evidence seems mixed and you are uncertain about the sentence. In thinking about the core facts, the words of a journalist, asking you a question some days ago, echo in your mind: "Do you think that the sentence for the defendant in this case will be higher or lower than three years?" You start to think about the journalist's question: "Three years of prison confinement, is this an appropriate sentence? Or is it too severe, or too lenient?" Will the journalist's question influence your sentencing decision?

If so, your decision may be biased by one of the most pervasive influences on human judgment, namely the anchoring effect (Tversky & Kahneman, 1974). Although as a legal judge, you definitely do not want to be directly influenced by a journalist's question, if you were, you would be in good company. A study by Englich, Mussweiler, and Strack (2006) showed that accomplished trial judges with an average of more than 10 years of experience were influenced by a journalist's question containing a sentencing demand. In fact, the magnitude of this influence proved to be dramatic. Judges who considered a high demand of three years embedded in the journalist's question gave final sentences that were almost 8 months longer than judges who considered a low demand of one year. A difference of 8 months in prison for the identical crime.

Similar effects were shown for the prosecutor's sentencing demand, even if the demand was explicitly made by a layman – a computer science student in the role

of the prosecutor (Englich & Mussweiler, 2001)—or if the prosecutor's sentencing demand was determined at random by throwing dice (Englich et al., 2006). Furthermore, the prosecutor's sentencing demand seems not only to influence the final judgment, but also the defense's counter-demand (Englich, Mussweiler, & Strack, 2005).

The anchoring phenomenon

As is true in a legal setting, human judgment is often influenced by biased values (for a classroom demonstration, see Text box 12.1). Such judgmental "anchoring" is defined as the assimilation of a judgment to a previously considered standard. Anchoring is a remarkable influence on human judgment, for at least two reasons. First, anchoring effects are very pervasive and robust. Second, the mechanisms that produce anchoring are after many years of investigation still a matter of lively debate.

TEXT BOX 12.1 ANCHORING EXPERIMENT

Anchoring effects are among the most robust and easily replicated findings in psychology. The experimental design we outline as a basis for classroom demonstrations follows the standard anchoring paradigm (Tversky & Kahneman, 1974).

Method

Participants

A total of 20 participants should be sufficient to produce reliable effects.

Materials

Four pairs of difficult general knowledge questions pertaining to different content domains are used as materials (see Appendix for an example). The anchors are typically set one standard deviation above and below the mean estimates of a calibration group that answered only absolute questions. However, more extreme values should also produce the effect.

Each question pair consists of a comparative and an absolute judgment. In the *comparative* judgments, participants indicate whether the target quantity is higher or lower than the anchor value (e.g., "Is the mean winter temperature in Antarctica higher or lower than –17°C?"). In the subsequent *absolute* judgments, participants provide their best estimate of the target quantity (e.g., "How high is the mean winter temperature in Antarctica?"). Two questionnaires are constructed such that two comparison questions contain a high

anchor and the other two contain a low anchor in the first questionnaire, and complementary anchors are used in the second questionnaire. Each of the questionnaires is then given to half of the participants.

Procedure

The questionnaires can be administered in groups. However, participants should not communicate with each other during the experiment. The questionnaire is handed to a participant with an instruction to read it carefully. To reduce the perceived informativeness of the anchors and thus to discourage conversational inferences, participants may be informed that they are taking part in a pretest for the construction of a general-knowledge questionnaire. The purpose of the pretest is ostensibly to find the best wording for general-knowledge questions. Instructions should emphasize that the comparison values were randomly selected. It may be further pointed out that this random selection is necessary to minimize the impact the values have on the answers and to thus identify the impact of different question formats. Finally, participants are instructed to answer all of the questions in the given order and to do so as accurately as possible.

Analysis

To pool answers across different content domains, absolute estimates are transformed into z-scores across participants, separately for each question. These scores reflect participants' average deviation from the question mean in units of the pertinent standard deviation. A simple analysis can be conducted with a paired *t*-test using averaged z-scores for the two questions in the high-anchor condition and averaged z-scores for the two questions in the low-anchor condition. (More elaborate analyses of non-averaged data can be done by using analysis of variance or multilevel modeling.)

Results and discussion

Absolute estimates should be assimilated towards the provided anchor values, so that higher mean estimates result for those targets that were compared to high anchors than for those that were compared to low anchors.

Pervasiveness and robustness

Anchoring effects pervade a large variety of judgments, from the trivial (i.e., estimates of the mean temperature in the Antarctica; Mussweiler & Strack, 1999a) to the apocalyptic (i.e., estimates of the likelihood of nuclear war; Plous, 1989).

They have been also observed in a broad array of different domains, such as general knowledge questions (Jacowitz & Kahneman, 1995; Tversky & Kahneman, 1974), price estimates (Mussweiler, Strack, & Pfeiffer, 2000; Northcraft & Neale, 1987), estimates of self-efficacy (Cervone & Peake, 1986), probability assessments (Plous, 1989), valuation of products (Ariely, Loewenstein, & Prelec, 2003), legal judgments (Englich & Mussweiler, 2001; Englich et al., 2005, 2006), and negotiations (Galinsky & Mussweiler, 2001; Loschelder, Stuppi, & Trötschel, 2014).

Not only is the anchoring effect influential in a plethora of settings, this influence is also remarkably robust. For one, anchoring occurs even if the anchor values are clearly uninformative for the critical estimate, for example, because they were randomly selected (e.g., Tversky & Kahneman, 1974). Moreover, even implausibly extreme values can yield the effect (e.g., Chapman & Johnson, 1994; Strack & Mussweiler, 1997). For example, in one study (Strack & Mussweiler, 1997) estimates for Mahatma Gandhi's age were assimilated to an unreasonably high anchor value of 140 years. Furthermore, anchoring may be in some cases uninfluenced by manipulations of accuracy motivation (e.g., Wilson, Houston, Etling, & Brekke, 1996; but see Simmons, LeBoeuf, & Nelson, 2010). In addition, it has been demonstrated that anchoring is often not affected by knowledge and expertise (Cheek, Coe-Odess, & Schwartz, 2015; Englich & Mussweiler, 2001; Englich et al., 2006; Northcraft & Neale, 1987). In the above-mentioned study in the legal domain (Englich & Mussweiler, 2001), for example, experienced judges and inexperienced law students were similarly influenced by the anchor sentencing demand if it was given by a computer science student. Furthermore, anchoring effects can persist even over fairly long periods of time. For example, anchoring effects were still apparent one week after the anchor value had been considered (Mussweiler, 2001). Moreover, anchor values do not need much attention to influence judgments. Even incidental anchors like a number on a football player's jersey (Critcher & Gilovich, 2008) or subliminally presented anchors (Mussweiler & Englich, 2005; Reitsma-van Rooijen & Daamen, 2006) can result in anchoring effects (but see Newell & Shanks, 2014). Probably the most striking demonstration of the robustness of the phenomenon, however, stems from a study showing that explicit instructions to correct for a potential influence of an anchor may not mitigate the effect (Wilson et al., 1996). That is, even explicitly forewarning judges about the potential distortion and informing them about its direction did not diminish the effect in the study. While the examples above show that anchoring is a robust phenomenon, its robustness depends on its underlying mechanism in a given judgment.

Relevance

Anchoring has not only been shown to be a robust judgmental effect that has been demonstrated in a variety of domains, it has also been suggested to play a role in a wide array of judgmental phenomena. For example, anchoring has been used to explain another eminent cognitive illusion, namely, the assimilation of a recollected estimate towards an outcome, which is also known as the hindsight bias (Hawkins &

Hastie, 1990; see also Chapter 23). The egocentricity of social judgment has also been attributed to an anchoring mechanism (Gilovich, Medvec, & Savitsky, 2000). Specifically, people may overestimate the extent to which they are noted by others, because they anchor the estimate on their own rich experiences. Similarly, the illusion of transparency – the tendency to underestimate ambiguity in communication when intentions are known – also shares some similarities with anchoring (Keysar & Barr, 2002).

In the psychology of judgment and decision making, anchoring has been primarily applied to probabilistic inferences. Thus, preference-reversal effects (Lichtenstein & Slovic, 1971), the distortion of estimates for the probability of disjunctive and conjunctive events, and the assessment of subjective probability distributions (Tversky & Kahneman, 1974) have been all attributed to judgmental anchoring.

Finally, applications of the anchoring concept are also found in applied contexts, such as negotiations, consumer behavior, and sentencing decisions (see the example at the beginning). For example, first offers may serve as anchors and thus influence the final negotiation outcome (Galinsky & Mussweiler, 2001). In consumer research, it has been suggested that price claims in advertisements influence behavior because they function as anchors in product evaluation (Biswas & Burton, 1993). Similar to anchoring effects on criminal sentencing decisions, research in the civil context of damage awards has shown that the higher the plaintiff's request in court, the higher the damage award that is given (Hastie, Schkade, & Payne, 1999; Malouff & Schutte, 1989; Marti & Wissler, 2000). In personal injury verdicts, the requested compensation systematically influences the compensation awarded by the jury as well as the judged probability that the defendant caused the plaintiff's injuries (Chapman & Bornstein, 1996). Ironically, even limits on damage awards serve as anchors and therefore increase damage awards (Hinsz & Indahl, 1995). While some research clearly points to anchoring effects in the field, the effect seems to be much more fragile than in the laboratory (Jung, Perfecto, & Nelson, 2015).

These accounts stand witness to the great diversity of phenomena that have been connected to judgmental anchoring. It is important to note however that these phenomena are not sufficiently explained by evoking an unspecific notion of anchoring. As such, the anchoring notion does not by itself illuminate the underlying mechanisms, but only describes the direction of the observed influence (assimilation). In this respect, the term "anchoring" constitutes a descriptive rather than an explanatory concept that does not go beyond the terms "assimilation" and "contrast" (Strack, 1992). In order to be used as an explanatory concept, the psychological mechanisms that underlie anchoring first have to be sufficiently understood.

Paradigms

Anchoring effects are typically examined in the standard paradigm introduced by Tversky and Kahneman (1974). Anchors are explicitly provided there by having judges compare the target to the anchor value. This is usually achieved by posing a comparative anchoring question and asking participants to indicate whether

a characteristic of the target is larger or smaller on the judgmental dimension than the anchor value. In order to avoid inferences about the intention that led to the selection of a particular anchor value (see Grice, 1975), it is typically presented as randomly selected. This may be achieved by spinning a wheel of fortune (Tversky & Kahneman, 1974), by emphasizing the random selection in the instructions (Strack & Mussweiler, 1997), by throwing dice (Mussweiler & Strack, 2000b), or by generating the value as the outcome of a clearly irrelevant process (Ariely et al., 2003).

In what is probably the best known demonstration of anchoring in this paradigm, Tversky and Kahneman (1974) asked their research participants two consecutive questions about the percentage of African nations in the UN. In a first *comparative* anchoring question, participants indicated whether the percentage of African nations in the UN was higher or lower than an arbitrary number (the anchor) that had ostensibly been determined by spinning a wheel of fortune (65% or 10%). In the subsequent *absolute* anchoring question, participants gave their best estimate of the correct percentage. Absolute judgments were assimilated to the provided anchor value, so that the mean estimate of participants who received the high anchor was 45%, compared to 25% for participants who received the low anchor.

Alternatively, the anchor may be provided to the participants in cases in which it is clearly informative for the judgment at hand. For example, Northcraft and Neale (1987) demonstrated that real-estate pricing decisions depended on the listing price for the property. Real-estate agents and lay subjects were given a 10-page booklet including all the information that is important for real-estate pricing. This booklet also contained the listing price of the house, which constituted the central independent variable. The price provided was either above or below the actual appraisal value of the property (e.g., $83,900 vs. $65,900). Replicating the typical anchoring finding, both expert and lay participants' estimates for the value of the property were assimilated toward the provided anchors.

In a third paradigm, anchors are self-generated rather than provided by the experimenter (Tversky & Kahneman, 1974). In one such study, participants were given 5 seconds to estimate the result of a product that was either presented in an ascending sequence ($1 \times 2 \times \ldots \times 8$) or in a descending sequence ($8 \times 7 \times \ldots \times 1$). Participants' estimates for the ascending sequence proved to be lower than for the descending sequence, presumably because participants based their estimates on the product of the first few numbers (which is lower for the ascending than for the descending sequence), which served as a self-generated anchor to which their final estimate was assimilated. Likewise, numerical estimates may be assimilated to self-generated anchors that are closely associated with the target quantity. For example, participants who are asked to give their best estimate for the freezing point of vodka may use the freezing point of water as an anchor and then adjust downwards because they know that the freezing point of alcohol is lower than that of water (Epley & Gilovich, 2001).

The "sequential judgment paradigm" also relies on self-generated anchors. However, an anchor is self-generated by answering an unrelated question. For example,

answering a question about the weight of a raccoon influences subsequent judgments of the weight of a giraffe. The estimated weight of a giraffe is assimilated to the anchor self-generated by estimating the weight of a raccoon. Importantly, the first answer influences the subsequent judgment only if it is made using the same response scale (Frederick & Mochon, 2012; Mochon & Frederick, 2013).

Finally, anchoring effects may be obtained by surreptitiously increasing the accessibility of the anchor value (Critcher & Gilovich, 2008; Wilson et al., 1996). In one experiment (Wilson et al., 1996) demonstrating such a "basic" anchoring effect, participants were first asked to copy either 5 pages of numbers ranging from 4,421 to 4,579 or 5 pages of words and subsequently estimated the number of students at their university who will contract cancer within the next 40 years. The participants who had copied 5 pages of high numbers estimated this number to be higher than those who had copied 5 pages of words. Thus, the irrelevant high anchor presented in the preceding task influenced the judgment. In another study (Critcher & Gilovich, 2008), the anchor value was made more accessible by its mere presence in materials. For example, participants judged that an American football linebacker, who was presented on a photo, had a higher probability of registering a sack (i.e., stopping the offensive team in a specific way) in a game if he had the number 94 displayed on a jersey than if the number was 54.

In sum, anchoring effects have been demonstrated using five different experimental paradigms, in which the anchor values are either explicitly or implicitly provided by the experimenter, self-generated, or made more accessible by a mere presence. Most of the anchoring research, however, first asks participants a comparative and then an absolute anchoring question, employing the standard paradigm introduced by Tversky and Kahneman (1974).

Theoretical accounts

To date five theoretical accounts of anchoring effects have been proposed. In particular, it has been suggested that anchoring effects result from (1) insufficient adjustment from an anchor, (2) conversational inferences, (3) numerical priming, (4) mechanisms of selective accessibility, and (5) distortion of a response scale.

1. Insufficient adjustment

In their initial description of the phenomenon, Tversky and Kahneman (1974) described anchoring in terms of insufficient adjustment from a starting point. They argued that "people make estimates by starting from an initial value that is adjusted to yield the final answer [. . .]. Adjustments are typically insufficient. That is, different starting points yield different estimates, which are biased toward the initial value" (p. 1129). Adjustment may be insufficient because it terminates once it reaches a region of acceptable values for the estimate (Epley & Gilovich, 2006; Quattrone et al., 1984). For example, participants who are asked whether the percentage of African nations in the UN is higher or lower than 65% may use this anchor value

as a starting point, determine whether it is too high or too low, and then adjust in the appropriate direction until the first acceptable value is found. Such insufficient adjustment is only possible if the anchor value falls outside of the distribution of acceptable values – that is, it constitutes an unacceptable value itself. This may be the case because the anchor value is extreme, or because it is known to be wrong. For example, participants self-generate the duration of Earth's orbit as an anchor in order to estimate the number of days it takes Mars to orbit the Sun. They are likely to know that 365 days constitutes an unacceptable value because Mars's orbit takes longer than Earth's (Epley & Gilovich, 2001). As a consequence, they may adjust from this unacceptable value until an acceptable value is reached.

However, anchoring effects are not obtained only for clearly implausible and unacceptable anchor values (e.g., Strack & Mussweiler, 1997). It seems difficult to explain effects of plausible and acceptable anchors by "insufficient adjustment" because for such anchors, there is no reason to adjust in the first place. The anchoring-and-adjustment account also cannot explain why an anchor influences the proportion of people making judgments higher and lower than the anchor. That is, people should know the direction of adjustment from the anchor; however, the direction is influenced by the comparative question (Jacowitz & Kahneman, 1995).

The insufficient adjustment thus appears to contribute to anchoring effects mainly if the critical anchors are unacceptable and self-generated rather than acceptable and externally provided values. Consistent with this assumption, participants' answers are more likely to be closer to the anchor within the range of acceptable values if the anchor is self-generated than if it is provided externally (Epley & Gilovich, 2001). Furthermore, adjustment is an effortful process and the availability of cognitive resources should therefore influence the size of the anchoring effect. Consistently, anchoring is reduced under cognitive load as well as after previous alcohol consumption in case of self-generated anchors (Epley & Gilovich, 2006). Similarly, forewarning and incentives reduce anchoring in case of self-generated anchors, but the results are mixed in case of externally provided anchors (Epley & Gilovich, 2005; Simmons et al., 2010).

2. Conversational inferences

A second account attributes anchoring to conversational inferences. According to this reasoning, applying implicit rules of natural conversations (Grice, 1975) to standardized situations (e.g., Schwarz, 1994) allows participants to use the anchor value to infer the actual range of possible answers. Participants who expect the experimenter to be maximally informative (Grice, 1975) in asking his or her questions may assume that the provided anchor value is close to the actual value and consequently position their estimate in its vicinity. Such conversational inferences may well underlie the effects of considering anchor values that are of clear relevance for the estimate to be made (e.g., Northcraft & Neale, 1987). Conversational inferences may also explain some other effects found in the anchoring literature. For example, more precise anchors (e.g., 4.85 rather than 5) lead to larger anchoring

effects (Janiszewski & Uy, 2008). Importantly, precision influences the size of the anchoring effect only if the anchor may be perceived as informative. In one study (Zhang & Schwarz, 2013), precision of an anchor influenced anchoring only if the anchor was allegedly created by a person and not if it was generated by a computer program. The conversational account also explains why people show weaker anchoring effects on their estimation of the population of Chicago if a high anchor is presented as a part of a question ("Do more or less than 5 million people live in Chicago?") than if it is a part of a statement ("The population of Chicago is less than 5,000,000."; Klein et al., 2014). Presumably, it is easier to infer that the anchor is informative in case of the statement than in case of the question. A recent study also shows that conversational inferences may play some role even in anchoring paradigms where experimenters try to make the anchor uninformative. In particular, Frederick, Mochon, and Savary (2014) found that the anchoring effect was smaller if participants took part in the random generation of the anchor than if the randomness of the anchor was conveyed by the experimenter. Additionally, while anchoring is present even in cases where the anchor is clearly irrelevant, relevance can increase the size of anchoring at least in an applied context (Glöckner & Englich, 2015).

It is important to note, that this account presupposes that the anchor value is indeed seen as informative for the judgment. Anchoring effects, however, also occur if the anchor values are clearly uninformative because they were randomly selected (Frederick et al., 2014; Tversky & Kahneman, 1974), are implausibly extreme (Strack & Mussweiler, 1997), or are not related to the question at all (Critcher & Gilovich, 2008; Frederick & Mochon, 2012). Thus, although conversational inferences are potential determinants of anchoring in natural situations, they are not a necessary precondition.

3. Numeric priming

A third theoretical account assumes that anchoring can be rather superficial and purely numeric in nature (Critcher & Gilovich, 2008; Wilson et al., 1996; Wong & Kwong, 2000). In particular, an anchor may simply render the anchor value itself more accessible, which influences the subsequent absolute judgment. From this numeric-priming perspective, the sole determinant of anchoring effects is the anchor value itself, regardless of its context, the target with which it is compared, and the judgmental operations in which it is involved. One account even goes so far as to claim that anchoring effects may be so superficial that not the anchor itself, but only its absolute value (e.g., "50" for an anchor of "–50°C") is represented in memory and exerts the primary anchoring influence (Wong & Kwong, 2000). A study by Oppenheimer, LeBoeuf, and Brewer (2008) further suggests that numeric priming may be only a specific example of a more general magnitude priming. According to the study, magnitude may be primed cross-modally – for example, drawing a long line increased subsequent numeric judgment.

However compelling such a simple numeric account may appear, a careful analysis of anchoring research reveals that focusing exclusively on the numeric value of an anchor is insufficient to explain most of anchoring effects. In particular, abundant evidence demonstrates that semantic content associated with the anchor has to be taken into account to understand the complete pattern of findings in the standard paradigm. For example, a purely numeric account cannot explain why anchoring effects depend on changes of the response scale (Frederick & Mochon, 2012) or the target of the comparative judgment (Bahník & Strack, 2016; Mussweiler & Strack, 2001). If anchoring effects were evoked by the anchor value itself, then identical effects should result irrespective of the semantic content with which the anchor is associated. For example, comparing the average *annual* and *summer* temperature in New York City to a given anchor value should both have identical effects on subsequent judgments of the average *summer* temperature in New York City because the numeric properties of the anchor are left unchanged by changing the target of the comparative judgment. This, however, is not the case. Rather, the anchoring effect disappears if the comparative anchoring question pertains to the average *annual* temperature (Bahník & Strack, 2016).

The temporal robustness of anchoring effects is also at odds with a purely numeric account which implies that anchoring effects are fairly transient and short-lived. Because we are constantly exposed to arbitrary numbers, our daily routines (e.g., calling a friend, paying a bill) should immediately wipe out the effects of solving a comparative anchoring task. The fact that anchoring effects can prevail for a week (Mussweiler, 2001) is clearly in conflict with this implication and further renders a purely numeric conceptualization of the standard anchoring paradigm unconvincing.

While numeric priming seems to be able to parsimoniously explain some effects (Critcher & Gilovich, 2008; Wilson et al., 1996; Wong & Kwong, 2000), its influence might be limited to judgment under cognitive load (Blankenship, Petty, Detweiler-Bedell, & Macy, 2008; see also Wegener, Petty, Blankenship, & Detweiler-Bedell, 2010), and it plays only a marginal role in most anchoring effects (Brewer & Chapman, 2002; Newell & Shanks, 2014).

4. Selective accessibility

The fourth theoretical account, the selective accessibility model of anchoring, argues that the anchoring effect is the result of increased accessibility of information consistent with an anchor (Mussweiler & Strack, 1999a, 1999b; Strack & Mussweiler, 1997; for a related account, see Chapman & Johnson, 1994, 1999). The model attempts to explain anchoring by linking it to two principles that are fundamental to social cognition research: (1) *hypothesis-consistent testing* and (2) *semantic priming*. More specifically, the model postulates that comparing the judgmental target to the anchor value changes the accessibility of knowledge about the target. In particular, the accessibility of knowledge that is consistent with the anchor is selectively increased because judges compare the target with the anchor by testing the

possibility that the target's value is equal to the anchor value. For example, judges who are asked whether the percentage of African nations in the UN is higher or lower than a high anchor of 65% test the possibility that this value actually is 65%. To do so, they selectively retrieve knowledge from memory that is consistent with this assumption (e.g., "Many African nations that are probably members of the UN come easily to mind", etc.). Such hypothesis-consistent testing is a general tendency that contributes to a variety of judgmental processes (Klayman & Ha, 1987; see Chapter 5 in this volume). In order to generate the final numeric estimate, judges then rely primarily on easily accessible knowledge (Higgins, 1996), so that their estimate is heavily influenced by the anchor-consistent knowledge generated before. In our example, absolute estimates of the percentage of African nations in the UN would be based on the subset of target knowledge that was retrieved specifically because it was consistent with the assumption that this percentage is fairly high. Conceivably, using this knowledge leads to high estimates, and the final estimate is thus assimilated to the anchor value.

The most direct support for this notion stems from a series of studies which directly assessed the accessibility of target knowledge subsequent to the critical comparative judgment (Englich et al., 2006; Mussweiler & Strack, 2000a, 2000b). In one of these studies (Mussweiler & Strack, 2000a), participants were asked to compare the average price for a German car to either a high or a low anchor value (40,000 vs. 20,000 German Marks). Subsequent to this comparative judgment, the accessibility of target knowledge was assessed with a lexical decision task. In particular, participants judged whether presented items are existing words. The items included target words that are closely associated with expensive cars (e.g., Mercedes, BMW) and words associated with inexpensive cars (e.g., VW). Response latencies for these two types of target words clearly depended on the anchoring condition. In particular, judges were faster in recognizing words associated with expensive cars after a comparison with the high anchor than after a comparison with the low anchor. In contrast, words associated with inexpensive cars were recognized faster after a comparison with the low anchor. These findings demonstrate that the accessibility of anchor-consistent semantic knowledge about the target (e.g., knowledge associated with high prices after a comparison with a high anchor) is increased as a consequence of the comparative judgment.

Further evidence suggests that this accessibility increase is specific to the judgmental target itself. That is, the knowledge that is rendered accessible specifically pertains to the judgmental target. In one study demonstrating this specificity, for example, comparing the self as a judgmental target to a high anchor of general knowledge only increased the accessibility of knowledge indicating that the self is knowledgeable, whereas the accessibility of knowledge about a close other remained unchanged (Mussweiler & Strack, 2000a). These findings provide support for the core assumption of the selective accessibility model. Comparing the target to the anchor value does indeed appear to increase the accessibility of anchor-consistent semantic knowledge about the target. Using this knowledge as a basis for the absolute estimate produces the assimilation effect that is known as the typical consequence of anchoring.

Additional support for the selective accessibility model comes from a recent study that tested a prediction derived from the model (Bahník & Strack, 2016). In particular, an anchor should not influence the absolute judgment if it activates information that would have been used for the judgment even without the anchor. For example, if the comparative judgment makes information about summer in New York City more accessible, it should not influence the judgment of the average summer temperature in New York City because the information overlaps with information that is used for making that judgment in any case. Bahník and Strack (2016) achieved the informational overlap by using a different target for the comparative and absolute judgment. As already mentioned, the study showed that the judgment of the average summer temperature in New York City was assimilated to a high anchor if the comparative question asked about the average *summer* temperature, making especially hot periods of summer more accessible, but not if it asked about the average *annual* temperature, which presumably activates information about summer. Importantly, a low anchor compared with the average annual temperature led to assimilation of the anchor in judgment of the average summer temperature. The low anchor presumably activates information about winter that does not overlap with information normally used for making the absolute judgment of the average summer temperature and that therefore exerts influence on the judgment.

The selective accessibility model is also consistent with other findings. For example, the time that is needed to generate a given judgment depends on the degree of accessibility of judgment-relevant knowledge (Neely, 1977). Accordingly, response latencies for the absolute anchoring judgment have been shown to depend on the extent to which the accessibility of judgment-relevant knowledge had been increased during the comparative judgment (Mussweiler & Strack, 1999a, 2000a, 2000b; Strack & Mussweiler, 1997). For example, judges were faster in giving absolute judgments if they had ample time to generate knowledge during the preceding comparison than when they had made the comparison under time pressure – a condition that is likely to limit the accessibility increase (Mussweiler & Strack, 1999a). Different levels of accessibility do not influence only response latencies for absolute judgments, but also the content of these judgments. In particular, larger anchoring effects occur under conditions that promote the extensive generation of anchor-consistent target knowledge and thus lead to a more substantial accessibility increase. For example, judges who have more target information available during the comparative task show more anchoring than those who have little information available (Chapman & Johnson, 1999). Consistently, anchoring effects do not diminish but get even stronger under sad mood compared to happy mood (Bodenhausen, Gabriel, & Lineberger, 2000; Englich & Soder, 2009). These pronounced anchoring effects under sad mood might be explained by the deeper and more thorough information processing under sad mood (e.g., Schwarz & Clore, 2007).

Recent studies (Frederick & Mochon, 2012; Mochon & Frederick, 2013) showed that anchoring can occur even when the target of the comparative and absolute judgment are largely dissimilar. For example, the absolute judgment of annual

rainfall at the driest place in the US (Death Valley) is influenced to the same degree by a comparative question related to the wettest place in the US (Mount Waialeale) as to Death Valley (Frederick & Mochon, 2012). Similarly, the estimate of the price of a camera is influenced to the same degree by a comparison question related to a GPS device as to the camera (Mochon & Frederick, 2013). According to the selective accessibility model (Mussweiler & Strack, 2000a), the size of the anchoring effect should be influenced by the applicability of the information made accessible by the comparative question to the absolute judgment. Presumably, the applicability is lower when the target of a judgment changes between the comparative and absolute question. The studies therefore suggest that the selective accessibility mechanism may not be necessary for producing anchoring in the standard paradigm, even though the evidence indicates that it is a potent theoretical account.

5. Scale distortion

A more recent theoretical account argues that anchoring may result from distortion of a response scale. The scale distortion theory argues that contrast effects that can be seen in perception are also present in the use of response scales (Frederick & Mochon, 2012). As water appears to be warmer if judges have previously put their hand in cold water, they also feel that 100 kilograms is heavier if they previously thought about 5 kilograms. This contrast effect can easily lead to the anchoring effect. For example, considering 10% of African states results in the feeling that 45% percent of states is relatively a large number. People would then choose a lower number as the answer for the absolute question because the lower number would be perceived as subjectively larger due to the comparison with 10%. The anchor may thus be assimilated to a subsequent judgment by distorting the response scale for that judgment.

The scale distortion theory of anchoring was studied mainly using the sequential judgment paradigm. The results of experiments using the paradigm support several predictions of the scale distortion theory. Anchoring was demonstrated, for example, by having participants estimate the weight of a raccoon and subsequently estimate the weight of a giraffe (Frederick & Mochon, 2012). While the participants who estimated the weight of a raccoon subsequently answered that the weight of a giraffe is on average 709 pounds, those who did not estimate the weight of a raccoon estimated the weight of a giraffe to be on average 1254 pounds. The anchor in a form of the estimated weight of a raccoon was therefore embedded in the subsequent judgment. Presumably, by answering a question about the weight of a raccoon, large numbers on the response scale felt even larger in comparison. The participants thus mapped the same representation of a giraffe to a smaller number if they previously estimated the weight of a raccoon. Since scale distortion operates only on a given response scale, anchoring should not occur if the response scale is changed between the two judgments. Consistently, participants who judged the weight of a raccoon on a 7-point heaviness scale were not influenced by this judgment and estimated the weight of a giraffe to be on average 1265 pounds (Frederick & Mochon, 2012).

Apart from the assimilation of an anchor value in the sequential judgment paradigm, the scale distortion theory predicts a contrast effect when objects are mapped to a certain value on the response scale. For example, participants that were first asked to estimate the weight of a wolf chose a heavier animal when asked afterwards which animal has its weight closest to 1000 pounds than participants that did not estimate the weight of a wolf. Apparently, the same 1000 pounds felt heavier in comparison after the judgment of the weight of a wolf (Frederick & Mochon, 2012). Scale distortion should not be influenced by what is the target of the comparative judgment. For example, the scale should be similarly distorted if the judgment of a price of a camera is preceded by the comparison of an anchor with a price of the camera or of a GPS device. The prediction was supported by some experimental data (Mochon & Frederick, 2013), but other research is inconsistent with it (Bahník & Strack, 2016).

Additionally, even in the sequential judgment paradigm the second judgment can be influenced differently by a similar anchor value depending on the target of the first judgment. For example, Chernev (2011) found that asking about a calorie estimate of a *salad* leads to a contrast effect in a subsequent estimate of calories in a cheesesteak while an estimate of calories in a *cake* leads to assimilation. Both calorie estimates of a salad and cake are lower than that of a cheesesteak, so they should both lead to a lower judgment of calories in a cheesesteak, according to the scale distortion theory, which assumes that anchoring should be largely independent of the target of judgment. In summary, while scale distortion theory explains parsimoniously some findings in the sequential judgment and standard anchoring paradigms, other findings are inconsistent with its predictions. The specific conditions under which scale distortion operates are yet to be explored.

Conclusions

Anchoring effects are among the most robust and ubiquitous psychological phenomena in judgment and decision making. Different underlying mechanisms were traditionally used to explain anchoring effects in different anchoring paradigms. Anchoring and insufficient adjustment is used to explain assimilation of judgment to an anchor in cases of unacceptable and self-generated anchors. Conversational inferences may particularly play a role when the anchor itself or the context of its presentation are perceived as informative. Numeric priming might contribute to anchoring effects, especially in cases of incidental anchors and under cognitive load. Selective accessibility seems to lie behind anchoring in cases of externally provided anchors. Scale distortion explains most of the effects in the sequential judgment paradigm. Its role in the standard anchoring paradigm is not yet known, but it may be relatively more prominent in cases of little knowledge about the judgmental domain.

While the evidence suggests operation of different processes under different circumstances, there is little evidence that other processes cannot operate simultaneously. Indeed, the theoretical accounts invoked to explain anchoring effects are not

mutually exclusive, even though, they are often described as such in the literature. However, the interaction of the various processes is not thoroughly explored and may stimulate future research. Given the broad definition of anchoring, it is not surprising that there is not a single mechanism that can explain it under all circumstances. Focusing on the judgmental processes rather than on judgmental effects, we may discover that the assimilation of a numeric judgment toward a previously considered value (also known as "anchoring") may be the result of different psychological mechanisms. Identifying the concomitant determinants may transform the "anchoring heuristic" into a psychologically rooted phenomenon with rich conceptual and applied implications.

Summary

- An assimilation of an estimate towards a previously considered standard is defined as judgmental anchoring.
- Anchoring constitutes a ubiquitous phenomenon that occurs in a variety of laboratory and real-world settings.
- Anchoring effects are remarkably robust. They may occur even if the anchor values are clearly uninformative or implausibly extreme, are sometimes independent of participants' motivation and expertise, may persist over long periods of time, and are sometimes not reduced by explicit instructions to correct.
- There are various underlying mechanisms that contribute to anchoring effects. Anchoring may under various conditions result from insufficient adjustment from a previously considered standard, use of conversational inferences, numeric priming, selective accessibility of information consistent with an anchor, or distortion of a response scale.

Further readings

A recent review of anchoring research is given by Furnham and Boo (2011). The main accounts of anchoring are explored in more detail in Mussweiler and Strack (1999a), Epley and Gilovich (2001), and Frederick and Mochon (2012).

References

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent arbitrariness": Stable demand curves without stable preferences. *Quarterly Journal of Economics*, *118*, 73–105.

Bahník, Š, & Strack, F. (2016). Overlap of accessible information undermines the anchoring effect. *Judgment and Decision Making, 11*, 92–98.

Biswas, A., & Burton, S. (1993). Consumer perceptions of tensile price claims in advertisements: An assessment of claim types across different discount levels. *Journal of the Academy of Marketing Science*, *21*, 217–229.

Blankenship, K. L., Wegener, D. T., Petty, R. E., Detweiler-Bedell, B., & Macy, C. L. (2008). Elaboration and consequences of anchored estimates: An attitudinal perspective on numerical anchoring. *Journal of Experimental Social Psychology*, *44*, 1465–1476.

Bodenhausen, G. V., Gabriel, S., & Lineberger, M. (2000). Sadness and the susceptibility to judgmental bias: The case of anchoring. *Psychological Science, 11*, 320–323.

Brewer, N. T., & Chapman, G. B. (2002). The fragile basic anchoring effect. *Journal of Behavioral Decision Making*, *15*, 65–77.

Cervone, D., & Peake, P. K. (1986). Anchoring, efficacy, and action: The influence of judgmental heuristics on self-efficacy judgments and behavior. *Journal of Personality and Social Psychology, 50*, 492–501.

Chapman, G. B., & Bornstein, B. H. (1996). The more you ask for, the more you get: Anchoring in personal injury verdicts. *Applied Cognitive Psychology, 10*, 519–540.

Chapman, G. B., & Johnson, E. J. (1994). The limits of anchoring. *Journal of Behavioral Decision Making, 7*, 223–242.

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes, 79*, 1–39.

Cheek, N. N., Coe-Odess, S., & Schwartz, B. (2015). What have I just done? Anchoring, self-knowledge, and judgments of recent behavior. *Judgment and Decision Making*, *10*, 76–85.

Chernev, A. (2011). Semantic anchoring in sequential evaluations of vices and virtues. *Journal of Consumer Research, 37*, 761–774.

Critcher, C. R., & Gilovich, T. (2008). Incidental environmental anchors. *Journal of Behavioral Decision Making*, *21*, 241–251.

Englich, B., & Mussweiler, T. (2001). Sentencing under uncertainty: Anchoring effects in the courtroom. *Journal of Applied Social Psychology, 31*, 1535–1551.

Englich, B., Mussweiler, T., & Strack, F. (2005). The last word in court – A hidden disadvantage for the defense. *Law and Human Behavior*, *29*, 705–722.

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, *32*, 188–200.

Englich, B., & Soder, K. (2009). Moody experts – How mood and expertise influence judgmental anchoring. *Judgment and Decision Making, 4*, 41–50.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, *12*, 391–396.

Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: Differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making*, *18*, 199–212.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, *17*, 311–318.

Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General*, *141*, 124–133.

Frederick, S., Mochon, D., & Savary, J. (2014). The role of inference in anchoring effects (Working paper). New Haven, CT: Yale University.

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, *40*, 35–42.

Galinsky, A. D., & Mussweiler, T. (2001). First offers as anchors: The role of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology, 81*, 657–669.

Gilovich, T., Medvec, V. H., & Savitsky, K. (2000). The spotlight effect in social judgment: An egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of Personality and Social Psychology*, *78*, 211–222.

Glöckner, A., & Englich, B. (2015). When relevance matters: Anchoring effects can be larger for relevant than for irrelevant anchors. *Social Psychology, 46*, 4–12.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics. Vol. 3: Speech acts* (pp. 41–58). New York: Academic Press.

Hastie, R., Schkade, D. A., & Payne, J. W. (1999). Juror judgments in civil cases: Effects of plaintiff's requests and plaintiff's identity on punitive damage awards. *Law and Human Behavior*, *23,* 445–470.

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin*, *107*, 311–327.

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). New York: The Guilford Press.

Hinsz, V. B., & Indahl, K. E. (1995). Assimilation to anchors for damage awards in a mock civil trial. *Journal of Applied Social Psychology*, *25,* 991–1026.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*, 1161–1166.

Janiszewski, C., & Uy, D. (2008). Precision of the anchor influences the amount of adjustment. *Psychological Science*, *19*, 121–127.

Jung, M. H., Perfecto, H., & Nelson, L. D. (2015). Anchoring in payment: Evaluating a judgmental heuristic in field experimental settings. Available at SSRN: http://ssrn.com/abstract=2442153 or http://dx.doi.org/10.2139/ssrn.2442153

Keysar, B., & Barr, D. J. (2002). Self-anchoring in conversation: Why language users don't do what they "should". In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 150–166). Cambridge, England: Cambridge University Press.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211–228.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š, Bernstein, M. J., . . . Nosek, B. A. (2014). Theory building through replication: Response to commentaries on the "Many Labs" replication project. *Social Psychology, 45*, 307–310.

Lichtenstein, S., & Slovic, P. (1971). Reversal of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, *89*, 46–55.

Loschelder, D. D., Stuppi, J., & Trötschel, R. (2014). "14,875?!": Precision boosts the anchoring potency of first offers. *Social Psychological and Personality Science, 5*, 491–499.

Malouff, J., & Schutte, N. S. (1989). Shaping juror attitudes: Effects of requesting different damage amounts in personal injury trials. *The Journal of Social Psychology*, *129,* 491–497.

Marti, M. W., & Wissler, R. L. (2000). Be careful what you ask for: The effect of anchors on personal-injury damages awards. *Journal of Experimental Psychology: Applied*, *6,* 91–103.

Mochon, D., & Frederick, S. (2013). Anchoring in sequential judgments. *Organizational Behavior and Human Decision Processes*, *122*, 69–79.

Mussweiler, T. (2001). The durability of anchoring effects. *European Journal of Social Psychology, 31*, 431–442.

Mussweiler, T., & Englich, B. (2005). Subliminal anchoring: Judgmental consequences and underlying mechanisms. *Organizational Behavior and Human Decision Processes*, *98*, 133–143.

Mussweiler, T., & Strack, F. (1999a). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, *35*, 136–164.

Mussweiler, T., & Strack, F. (1999b). Comparing is believing: A selective accessibility model of judgmental anchoring. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 10, pp. 135–167). Chichester, England: Wiley.

Mussweiler, T., & Strack, F. (2000a). The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of Personality and Social Psychology, 78*, 1038–1052.

Mussweiler, T., & Strack, F. (2000b). Numeric judgment under uncertainty: The role of knowledge in anchoring. *Journal of Experimental Social Psychology*, *36*, 495–518.

Mussweiler, T., & Strack, F. (2001). The semantics of anchoring. *Organizational Behavior and Human Decision Processes, 86*, 234–255.

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin, 26*, 1142–1150.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading of activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *3*, 226–254.

Newell, B. R., & Shanks, D. R. (2014). Prime numbers: Anchoring and its implications for theories of behavior priming. *Social Cognition*, *32*, 88–108.

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes, 39*, 84–97.

Oppenheimer, D. M., LeBoeuf, R. A., & Brewer, N. T. (2008). Anchors aweigh: A demonstration of cross-modality anchoring and magnitude priming. *Cognition*, *106*, 13–26.

Plous, S. (1989). Thinking the unthinkable: The effects of anchoring on likelihood estimates of nuclear war. *Journal of Applied Social Psychology, 19*, 67–91.

Quattrone, G. A., Lawrence, C. P., Warren, D. L., Souza-Silva, K., Finkel, S. E., & Andrus, D. E. (1984). Explorations in anchoring: The effects of prior range, anchor extremity, and suggestive hints. *Unpublished manuscript*.

Reitsma-van Rooijen, M., & Daamen, D. D. L. (2006). Subliminal anchoring: The effects of subliminally presented numbers on probability estimates. *Journal of Experimental Social Psychology*, *42*, 380–387.

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 123–162). San Diego, CA: Academic Press.

Schwarz, N., & Clore, G. L. (2007). Feelings and phenomenal experiences. In T. Higgins & A. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (2nd ed., pp. 385–407). New York: Guilford.

Simmons, J. P., LeBoeuf, R. A., & Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors? *Journal of Personality and Social Psychology*, *99*, 917–932.

Strack, F. (1992). The different routes to social judgments: Experiential versus informational strategies. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgment* (pp. 249–275). Hillsdale, NJ: Erlbaum.

Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, *73*, 437–446.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1130.

Wegener, D. T., Petty, R. E., Blankenship, K. L., & Detweiler-Bedell, B. (2010). Elaboration and numerical anchoring: Implications of attitude theories for consumer judgment and decision making. *Journal of Consumer Psychology*, *20*, 5–16.

Wilson, T. D., Houston, C., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, *4*, 387–402.

Wong, K. F. E., & Kwong, J. Y. Y. (2000). Is 7300 m equal to 7.3 km? Same semantics but different anchoring effects. *Organizational Behavior and Human Decision Processes*, *82*, 314–333.

Zhang, Y. C., & Schwarz, N. (2013). The power of precise numbers: A conversational logic analysis. *Journal of Experimental Social Psychology*, *49*, 944–946.

APPENDIX

Comparative anchoring questions and high (and low) anchor values:

1 Is the mean winter temperature in Antarctica higher or lower than −17 (−43) °C?
2 Was Leonardo da Vinci born before or after 1698 (1391) A.D.?
3 Was Albert Einstein's first visit to the US before or after 1939 (1905)?
4 Was Mahatma Gandhi older or younger than 79 (64) years when he died?

13

VALIDITY EFFECT

Catherine Hackett Renner

The philosophical roots of the validity effect go back to Descartes and Spinoza, both of whom first developed models to explain how the truth of a statement is represented in human memory. Descartes postulated that information is "tagged" as true or false, respectively. Spinoza argued that we automatically assume information we hear is true and revise if/when we learn it is false. Spinoza's model infers a truth bias, whereas Descartes' model suggests we comprehend information separate from evaluating whether it is true or false (Nadarevic & Erdfelder, 2013; Street & Richardson, 2015).

Research can be found supporting both models. The distinction between the models is found in the context within which the information is presented as well as the amount of relevant background knowledge one has on the topic. When the actual truth status of the information is known, people tend to be more Cartesian in nature (Dechêne, Stahl, Hansen, & Wänke, 2010). Here, effects of repetition affecting the assessments of truth for information known to be true are larger in magnitude than for statements known to be false. However, when the truth status of the information is unknown, people tend to be more Spinozan in nature, with the processing fluency of information acting as a proxy for the validity of the statement.

The effect of repetition on enhancing the perceived validity of information has long been independently established by several researchers (e.g., Arkes, Hackett, & Boehm, 1989; Bacon, 1979; Begg, Armour, & Kerr, 1979; Gigerenzer, 1984; Gude & Zechmeister, 1975; Hasher, Goldstein, & Toppino, 1977; Parks & Toth, 2006; Schwartz, 1982). Collectively, these studies have found that if information has been heard previously, people are likely to ascribe more truth or validity to it than if they are hearing it for the first time. This phenomenon, referred to as the "validity effect" or illusion of truth, occurs regardless of the type of information (factual, nonfactual, or political), regardless of whether the information was originally believed to be true or false, and regardless of whether the exact wording was repeated.

Overview: Demonstrating and explaining the occurrence of the validity effect

Establishing the phenomenon of the validity effect

Text box 13.1 includes the detailed description of a typical validity-effect experiment that can be used as a classroom demonstration. In 1977, Hasher et al. demonstrated that the mere repetition of a statement to research participants prompted the participants to rate the statement as more true or valid. On three successive occasions, separated by two-week intervals, Hasher et al. presented participants with 60 factual statements that were either true or false. These statements had been pre-tested in order to ensure that on a 1–7 scale of truth assessment

TEXT BOX 13.1 THE TYPICAL VALIDITY EFFECT EXPERIMENT

Generating statements

The typical validity effect experiment requires that numerous "trivia" statements be generated and pilot-tested first in order to ascertain how true or false participants perceived them. The most recent almanac is an excellent source of information for creating statements of fact. During the pilot test, the participants should rate the statements on the same rating scale that will be used in the experiment. The majority of the experiments have used a Likert scale ranging from 1 (definitely false) to 7 (definitely true). Once the statements have been generated and pilot-tested, the process of choosing the statements that will be used in the research or classroom demonstration begins. In most experiments, statements that have a neutral mean rating (near 4.0) are chosen. For the experiment described below, 100 of such statements are needed, 50 of which are actually true and 50 of which are actually false.

Implementing the experiment

Participants

Since any experiment in this area will necessarily be a repeated-measures study, having at least 20 participants is important (in order to have enough power to detect differences between the sessions), and the participants must be able to attend two sessions. This type of research lends itself nicely to a classroom setting as most of the participants will return for the necessary subsequent sessions. (One note of caution: The participants used for the pilot test should not also participate in the research or classroom demonstration.)

Materials

After the list of factual statements has been pilot-tested, the statements that will be repeated in the second session need to be randomly selected. The list should contain 30 true and 30 false statements in the first session. In the second session, 20 new true and 20 new false statements will be needed to mix with only 10 true and only 10 false statements from the first session. Each session will have a total of 60 statements, 20 of which will appear in both sessions (see Table 13.1 for examples of statements used in the Arkes et al., 1989, study). The statements should be in random order. It would be a good idea to be sure that the first 5–8 statements are always non-repeated statements in order to serve as primacy and recency buffers.

Procedure

For ease of explanation, this procedure assumes that the statements will be given in a written list to the participants. In the first session, the participants are instructed to read each statement and to indicate on a 1 (definitely false) to 7 (definitely true) scale how true or false they believe the statement to be. After an interval of time has passed (across studies this interval has varied from three days to one month), the participants are asked to return for a second session to engage in the same task of rating statements. In this second session, the participants are told that some of the statements may be repeated

TABLE 13.1 Examples of true and false statements*

True statements	*False statements*
More presidents of the United States were born in Virginia than any other state.	New Delhi, India, is the world's most populous city.
The thigh bone is the largest bone in the human body.	Willie Mays had the most home runs in one season of any National League player.
Canada is the world's largest producer of silver.	New Jersey was the first state to ratify the eighteenth amendment.
Bolivia borders the Pacific Ocean.	The Indian Ocean is the smallest ocean on earth.
The largest dam in the world is in Pakistan.	The planet Venus is larger than the earth.

*At the time of the Arkes et al. (1989) experiment, these statements were factually true and factually false. Since then the status of some of these statements may have changed. These statements were used in a sample of participants from the United States and therefore reflect some information that is specific to the United States.

TABLE 13.2 Design and results table of the classroom experiment

	Session 1		*Session 2*	
	True	*False*	*True*	*False*
Repeated				
Non-repeated				

Note: Recall that "repeated" means that the same statements are presented in Sessions 1 and 2, whereas "non-repeated" means that different statements are used in Sessions 1 and 2.

from the first session. The second list will be comprised of the 20 statements repeated from the first session and the 40 new statements.

Statistical analysis

To analyze the data, the mean of the ratings given to the true and false repeated and non-repeated statements across the two sessions will need to be computed for each participant. Therefore, each participant will have eight mean ratings as shown in Table 13.2.

From here a 2 (Session: 1 vs. 2) × 2 (Veracity: true vs. false statement) × 2 (Repetition: repeated vs. non-repeated statements) repeated-measures analysis of variance can be performed. The hallmark of the validity effect is a statistically significant Session X Repetition interaction. Veracity ratings should increase from Session 1 to Session 2 for repeated statements, but not for non-repeated statements. There should be no effect of the actual veracity of the statements.

(higher values indicating higher levels of truth) all statements averaged a rating of close to 4.0. The statements were statements of trivia of the type "the population in Greenland is about 50,000". Twenty of the statements from the first session were repeated in Sessions 2 and 3; the remaining 40 items in each session were new. The results revealed that the rated validity of the repeated statements increased across sessions regardless of whether the statements were originally true or false. Non-repeated statements (those presented only once during one of the three sessions) were given approximately the same validity rating during each session. From these results, Hasher et al. concluded that the frequency of occurrence is a criterion we use to judge the referential validity of information.

The validity effect as an automatic process

Prior to the work of Hasher and her colleagues, Gude and Zechmeister (1975) had demonstrated the occurrence of the validity effect regardless of whether the

exact wording of the sentences was used. In their research, the effect of repetition was found similarly in sentences that were both exact in wording of the original sentences as well as sentences that were not exact in wording but had the same underlying meaning or gist. To further explain this phenomenon, Hasher and Chromiak (1977) performed two studies to ascertain whether detecting frequency may be an automatic process. In this research, participants from grades 2, 4, 6, and college were asked to judge how frequently a word had been presented to them during a testing session. In some of the conditions, the participants were explicitly told that some items would be repeated, whereas in others, they were not told. The results of this research revealed that the ability to count frequency did not show a developmental trend, nor was it affected by instructions indicating that repetition would occur. This type of finding implies that the validity effect is a cognitive skill that is based on recognition memory, which has historically shown limited development with age (Brown, 1975). However, as we age, this cognitive skill may become practiced enough that it becomes an automatic process (i.e., one that is not under the conscious control of the individual). If the ability to detect frequency is an automatic process, it would then make sense that it would be insensitive to explicit instructions regarding whether the information has been heard, as these instructions now become redundant to the individual.

In sum, the collective works by Hasher suggest that detecting the frequency with which information has been previously heard is an automatic process. The detection of frequency is used to determine whether information is new or old and may also be used to ascertain how likely the information is true. The next question to answer pertains to whether the validity effect occurs for information believed to be repeated regardless of its actual repetition status. Here we turn to Bacon (1979). (*Note*: "Frederick T. Bacon" is a pseudonym for a research group that included Ian Begg, Grant Harris, John Mitterer, and Douglas Upfold.) In their research, statements judged by the participants to be repeated were rated as more true regardless of whether the statement was actually repeated or not. Bacon concluded that it is the judgment that the information was repeated that leads one to believe that the information has been repeated rather than the actual repetition status. In addition, this research also supports the contention that the repetition sparks recognition memory. Bacon further believed that the validity effect is a recognition effect rather than a repetition effect.

Exploring the types of information that produce the validity effect

In all of the studies mentioned thus far, the statements used in the experiments were trivia statements that had an overall rating of 4 on a 1 (definitely false) to 7 (definitely true) scale. In essence, these statements were of neutral truth value to the participants. Also, all statements were statements of fact. In the first of two experiments, Arkes et al. (1989) were interested in whether the validity effect would occur for statements that were opinion statements (e.g., "Competition in schools is

not good for young children") as well as statements of fact (e.g., "The thigh bone is the largest bone in the human body"). In addition, my colleagues and I were interested in whether the validity effect would occur similarly for statements that were obviously true or obviously false.

In a separate sample, Arkes et al. pilot-tested numerous statements to ascertain the perceived truth status. Statements that obtained a mean rating of 1.8–2.8 were used as the "perceived false" statements. Statements that obtained a mean rating of 3.95–4.95 were used as the "perceived neutral" statements and statements that obtained a mean rating of 5.3–6.5 were used as the "perceived true" statements. We labeled these statements as "perceived" as half of the statements in each category chosen were factually false while others were factually true. Therefore, we were also able to look at the occurrence of the validity effect within those statements that were erroneously believed to be true and false as well as those that were correctly believed to be true and false. In two sessions, participants were asked to rate the validity of these statements with some of the statements in Session 2 being statements repeated from Session 1.

The results of this study revealed that the validity effect occurred similarly for statements of fact as well as statements of opinion in that repeated statements were given higher ratings at Session 2 than Session 1. In addition, the validity effect occurred similarly for all statements regardless of their perceived truth rating. This means that those statements initially perceived to be false rose in rated validity at Session 2 as did statements initially perceived to be true and those initially perceived to be neutral. This finding separates the validity effect from an attitude-polarization effect. If the effect of repetition was simply an extension of attitude-polarization, then the effect of repetition on the statements believed to be true should be that these statements are rated as more true, and by extension, the effect of repetition of the false statements should be that these statements are rated as more false. As stated, this was not the case, repeated statements that were initially perceived to be false were rated as more true (i.e., less false) in the second session. In conclusion, the validity effect seems to occur for statements of fact as well as statements of opinion. It also occurs similarly for factual and opinion statements that are originally believed to be true as well as those that are originally believed to be false.

The role of the source of information in the validity effect

It is possible that the source of the information could be contributing to the occurrence of the validity effect. In an experimental setting in which participants are hearing statements they believe they have heard before, it may be the case that they also think they heard the statement elsewhere first (Source 1) and then again in the experiment (Source 2). In Experiment 1 of Arkes et al. (1989) participants were asked to indicate whether they had heard the statement before, and if so where (only within the experiment or from a source outside of the experiment). The results indicated that information attributed to sources outside of the experimental setting was given the highest validity ratings.

This prompted further study by Arkes, Boehm, and Xu (1991). In a series of three studies, these researchers sought to further explain the impact of source dissociation, need for cognition, and the outside boundaries of the validity effect (i.e., how many repetitions continue to produce the effect and could the effect be found in statements that are related to assertions but not exact replications). A structural equation model of the relationship between source dissociation, validity, and familiarity found that source dissociation does not affect validity but it does affect familiarity (which then affects validity). Statements that were attributed to a source outside of the experiment were rated as more familiar. In this analysis, need for cognition was not found to be associated with any of the factors studied. With respect to the boundaries of the validity effect, Arkes et al. (1991) replicated and extended the earlier work of Begg et al. (1979). In a second experiment, participants read a passage concerning information about China and in subsequent weeks rated the validity of sentences that were (a) related to the passage, (b) not related to the passage but related to China, and (c) not related to the passage or China. Those statements that were related to the passage were given higher validity ratings than statements that were related to China but not in the passage or statements that were not related to either the passage or China. This result demonstrates that the validity effect extends beyond the exact presentation of the information and includes information that is similar to what was first heard. Finally, in a third experiment, Arkes et al. (1991) found that across six sessions, the bulk of the validity effect occurred in the second session (i.e., as a result of the first repetition). While rated validity further increased in Session 3–6, the increase was not statistically significant. In conclusion, the results of this research demonstrated that it only takes one repetition that is similar (rather than identical) to the first presentation of the information to increase the rated and perceived validity of information.

The role of expertise in the validity effect

In a second study, Arkes et al. (1989) attempted to explore the impact of levels of knowledge or expertise with the topic on the validity effect. Past research in the area of experts and novices (e.g., Chiesi, Spilich, & Voss, 1979) has consistently found that within their area of expertise, experts have better memory for information than novices. Given this, Arkes et al. questioned whether or not the validity effect would occur for information in the area of one's expertise compared to information not in one's area of expertise. In an attempt to determine if expertise would have an impact, statements that came from 7 different categories of information (food, literature, science, art, history, entertainment, and sport) were generated. In the first session, participants were asked to rank-order their knowledge of information in these seven categories before they rated how true the statements were. Participants returned for a second session and rated another list of statements in which some of the statements had been repeated from the first session and some were new. The participants were not asked to rank their knowledge of the seven categories at this session. Analysis of the data from this study revealed that the validity effect

occurred only in those topics for which the participants rated themselves as being knowledgeable. Given this, familiarity with information mediates the impact of repetition such that information from topic areas in which one has little knowledge will be rated less valid than information from topic areas in which one has more knowledge.

The validity effect and the hindsight bias

The validity effect has been used as an explanation for the asymmetry found in the hindsight bias (Hertwig, Gigerenzer, & Hoffrage, 1997; see Chapter 23 in this volume). In the hindsight-bias research design, participants are presented with a list of statements and asked to assess their confidence that each statement is true. Participants are then informed which statements are in fact true and which are false. For some (control) statements, participants do not receive any feedback. After the feedback phase, participants are asked to recall their original confidence estimates. When the hindsight bias occurs, the resulting outcome is that the recalled confidence estimates are higher than the original estimates when told that the statement was true, but lower than the original estimates when told that the statement was false. However, as Hertwig et al. reported, these changes in recalled confidences are not symmetric for true and false statements: The change in confidence was larger for positive statements. In addition, when no feedback was provided, recalled confidence estimates were higher than the original confidence estimates regardless of whether the statements were true or false.

The observed asymmetry for statements known to be true or false could be explained in the following way: The no-feedback condition of the hindsight-bias study design clearly demonstrates the occurrence of the validity effect (labeled "reiteration effect" by Hertwig et al., 1997). The two feedback conditions can be viewed as a combination of this reiteration effect and the typical hindsight bias. However, for true statements both point towards the "true" end of the confidence scale, thus leading to a relatively large change in confidence, whereas for false statements, both point into opposing directions (reiteration to the "true" end and feedback to the "false" end), thus leading to a relatively small net change.

The validity effect outside of the laboratory: Is there ecological validity?

Up to this point, the information presented to participants has been trivia statements presented in a university laboratory. The participants are students enrolled in the university and the climate is such that they are encouraged and expected to believe what they are told by professors. It is possible that the validity effect is an epiphenomenon of the setting, and has little ecological validity. It now becomes important to generalize the validity effect to field settings.

That the validity effect occurs in "real life" is not at question here. There are numerous accounts that confirm that the validity effect does in fact occur outside of

the laboratory. For example, it is well known in the world of advertising that repetition works (Krugman, 1977). Krugman clearly supported that a "frequency of three" is critical for advertising effectiveness. Krugman linked each of the three repetitions to awareness in the population with the first exposure interpreted by the consumer as "What is it", the second exposure interpreted as "What is the personal relevance of it?", and the third exposure taking the role of a reminder. Krugman went on to state that with enough mere repetition of product brand names, we have a higher tendency to buy the product solely because we recognize it (see Chapter 14). While researchers in advertising have disagreed with Krugman's "frequency of three", it has only been to provide evidence that increased repetition brings more success. For example, Meyers (1993) demonstrated the need to increase the minimum number of repetitions for effective advertising. Meyers argued that due to media fragmentation, repetition needs to become the most important tool of the advertiser.

But where else might there be a demonstration of the validity effect? One suggestion is in a college classroom. Many introductory psychology instructors can attest to the fact that students often come to them and say they have heard all of this information before and believe they know it already. The source of the information may have been their high-school psychology class, the media in general, or other classes within the college setting that have some degree of overlap with the introductory psychology course (e.g., the information on the brain may well be covered in an introductory biology class). According to the validity effect, each time students have heard the information they are then more likely to believe it is true. I was interested in how this truth assessment manifested itself in a learning environment. Given the past research, the repetition is based on recognition; therefore the students would recognize the information as having been heard before. How would this familiarity be interpreted in a learning environment?

In a series of experiments looking at the impact of repetition and familiarity on learning, my colleagues and I explored a number of potential ways in which course content information may have been recognized as a repetition by the students (Renner, Renner, & Shaffer, 2004). In our first experiment, we defined familiarity with course material in two ways. In the first way we assumed that if students had taken the course previously they would be familiar with the material in the course due to the previous exposure. In the second way we compared terms that are more common in the psychological literature to those that are not, under the assumption that common terms would have been heard or learned before in other courses. The participants were college students enrolled in a general psychology class, 46% (66/142) of whom had taken high-school psychology.

We needed to develop a measure of knowledge of psychology that was separate from course exams and that would equally represent the various areas in an introductory psychology course. To accomplish this, a 54-item Psychology Information Test (PIT) was constructed. The PIT contained three items from each of the 18 chapters represented in the text book used by the instructor, which was organized similarly to most contemporary texts. The questions were selected based

on the commonness of the term being queried in an introductory psychology text. The determination of commonness came from research by Boneau (1990) who created a list of terms used in general psychology from surveying glossaries and indexes of general psychology text books. He then asked textbook authors to rate the list of terms from most to least common. From this data, Boneau created a list of the 100 most common terms in each of the general areas of psychology. For each chapter, a highly common term, moderately common term, and less common term was chosen from the Boneau list. A question representing each term was then chosen from the test-item file that accompanied the text book (all questions were of moderate difficulty level as assessed by the authors of the test-item file). This procedure resulted in a 54-item PIT.

On the first day of class, the students were told that in order to have a general idea of how much knowledge students have of psychology before taking a psychology course, they would be asked to answer the PIT. On the second to last day of class, the students were asked to fill out the PIT again so that the instructor could assess how much information they gained as a result of the course. The results of this study revealed that terms that are high in commonness resulted in worse performance on the PIT than terms that were low in commonness. Students who had a high-school psychology course before taking a college-level psychology course performed significantly worse than students who had not taken a high-school psychology course. Having prior exposure to psychology does not enhance performance in the classroom. We suggest that the prior exposure to the material creates the perception that the material is known. We suspect that the increased familiarity with the information creates a false sense of knowing the information. We designed a second study to attempt to ascertain whether this suspicion is true. In our second study, we were interested in determining whether increased familiarity would lead to a false assumption that the material was known. If this is true, students should have a high level of confidence in their knowledge of the material that they indicate is familiar. However, if the material is familiar only because of a previous exposure, this confidence would be misleading and result in poorer class performance.

The participants were students enrolled in an undergraduate psychology course. On the first day of class, the instructor explained to the students that she was trying to determine ways to enhance learning of course material. One way to do this was to find out what students knew about the course content before the course started. The students were then given a list of 100 terms that would be discussed throughout the course and asked to indicate (Yes or No) if they had ever heard each of the 100 terms before. Approximately 10 terms were randomly chosen from each chapter for inclusion in the list.

Throughout the course, the items in the terms list appeared on tests. For each term, the percentage of students who had heard of the term, and the percent correct on the course examination for that term, was recorded. We then looked at the accuracy of each term according to the number of students who said they had heard the term before. The highest accuracy was obtained when the term had never been

heard before. For other terms, as the number of students who had heard of the term increased, the percent correct on the terms increased, but never reached that of the group who had never heard of the term before.

The conclusion of these two studies is that there is ecological validity to the validity effect in an educational setting. We suspect familiarity is used as a proxy measure indicating how true it is that we know information. The more we hear information, the more familiar it is. Increased familiarity leads to an increase in processing fluency. The increase in processing fluency leads to an increase in the veracity of our knowledge such that "I've heard it before, therefore I know it" results. Unfortunately, based on our research, this assumption is not true. Familiarity is not a valid predictor for one's knowledge in that simply having heard something before does not mean there is also a depth of accurate knowledge about that item or topic.

Debiasing the validity effect

Unkelbach (2007) was able to reverse the effects of processing fluency by providing feedback to participants. In a series of three experiments, Unkelbach was able to demonstrate that the pairing of the assessment of truth to processing fluency is learned. Consequently, providing feedback when processing fluency is high, yet information is not true, can reverse this effect.

Relation of the validity effect to other phenomena

It is important to distinguish the validity effect from other phenomena that have similar characteristics. Tesser (1978) performed a study in which he measured the impact of repetition on affect and found increased affect with increased repetition among positive items but decreased affect with increased repetition among negative items. Tesser's study better represents the impact of repetition on attitude-polarization than the validity effect. Remember that in the Arkes et al. (1989) study, false statements (the most comparable to Tesser's negative items) increased in perceived validity with repetition. If the validity effect were nothing more than attitude-polarization, the effect of repetition on the false statements should have prompted the participants to rate the statements as more false rather than more true.

Harrison (1977) and Zajonc (1968) have summarized a large number of studies which suggest that repeated exposure increases the liking for a wide variety of stimuli. This effect has been called the "mere exposure" effect (see Chapter 14). Research in this area used research materials that are not factual in nature (e.g., photographs, melodies, and tastes). The difference between the mere-exposure effect and the validity effect lies in the distinction between assessments of liking and validity. It is entirely possible that liking and validity are very different concepts that operate separately and distinctly from each other. Given this, we may come to believe a statement is increasingly true regardless of whether or not we like the statement.

The truth effect (or truth bias) has also been linked to the validity effect (Dechêne, Stahl, Hansen, & Wänke, 2009; Dechêne et al., 2010; Levine, 2014; Unkelbach, Bayer, Alves, Koch, & Stahl, 2010). In the truth effect, the repetition of information

increases one's subjective estimation of the information's truth. This effect is stronger when the information is ambiguous as well as when the information is presented within a context of mixed or heterogeneous (some repeated and some new statements) lists. The processing fluency that occurs with repeated information is the likely explanation for this finding. While the truth effect and the validity effect are similar when statements are ambiguous (i.e., political statements and opinion statements), they differ when statements are not ambiguous. Therefore, it is more likely that the truth bias may be an explanation for one facet of the validity effect but not for the entirety of the phenomenon.

Finally, the term "validity effect" has been used within the area of visio-spatial attention (Vossel, Thiel, & Gereon, 2006) to describe the difference in reaction time in location-cueing paradigms when a cue provides either correct or incorrect information about a target. In this scenario, the validity effect is used as a measure of the cost of inattention. There is no overlap between this definition of the construct and the definition used in this chapter.

Conclusion

For more than 25 years, the phenomenon of the validity effect has been established by several different researchers, from numerous research laboratories, using different methodologies. Within these studies the validity effect has occurred regardless of age (grade school vs. college), types of information (factual vs. opinion), length of delay between presentations (same day vs. several weeks), type of presentation of the statements (written vs. auditory), and precision of the repeated information (exact duplication vs. gist). The repeated replication of this phenomenon across time and methods demonstrates that the validity effect is a robust phenomenon.

Attention should now turn to enhancing our understanding of the impact of the validity effect in numerous areas of life. For example, the research by Renner et al. (2004) suggests that students equate repetition of information with familiarity of information that further prompts them to think they are knowledgeable about the information. If this is true, would students who retake courses study less for the course the second time around, because they are under the assumption that they "know" the information since they had it before? Is it only necessary to continually repeat to constituents the message a particular political candidate wants them to believe rather than provide evidence? Finally, does the validity effect deter critical thinking about the information presented? These are all important questions, raising important issues. Finding the role of the validity effect in these situations may assist in moving people from blindly accepting pre-thought thoughts (Gatto, 2001) to employing critical evaluation of information. It is time to move beyond demonstrations that the validity effect occurs to demonstrations of its extended impact.

Summary

- The validity effect occurs when the mere repetition of information increases the perceived truthfulness of that information.

- The validity effect appears to be based on recognition memory and may be an automatic process.
- The validity effect occurs similarly for statements of fact that are true and false in origin as well as political or opinion statements.
- Factors such as the belief that the information was heard previously or having some expertise in the content of the statement prompts information to be rated as more valid even if the information has not been repeated.
- The validity effect is not similar to the mere-exposure effect in that liking and validity are separate and distinct phenomena.
- The validity effect is not similar to attitude-polarization as evidenced by the fact that false statements demonstrate an increase in truth rating with repetition (the validity effect) rather than a stronger rating of being false (attitude-polarization).

Further reading

A good place to start in developing an understanding of the beginning research on the validity effect is by reading the Hasher et al. (1977) article. The cognitive processes that do and do not mediate the validity effect are examined by Gude and Zechmeister (1975), Hasher and Chromiak (1977), Parkes and Toth (2006), Unkelbach (2007), and Unkelbach et al. (2010). Familiarity with information and the boundaries of the validity effect are explored in a series of studies by Arkes et al. (1989, 1991). Finally, whether or not we believe that the information has been repeated and the subsequent impact on the perceived truth of information was first demonstrated in studies by Bacon (1979), Begg et al. (1979), Schwartz (1982), and more recently in Dechêne et al. (2009, 2010).

References

Arkes, H. R., Boehm, L., & Xu, G. (1991). Determinants of judged validity. *Journal of Experimental Social Psychology*, *27*, 576–605.

Arkes, H. R., Hackett, C., & Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making*, *2*, 81–94.

Bacon, F. T. (1979). Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 241–252.

Begg, I., Armour, V., & Kerr, T. (1979). On believing what we remember. *Canadian Journal of Behavioral Science*, *17*, 199–202.

Boneau, A. C. (1990). Psychological literacy: A first approximation. *American Psychologist*, *45*, 891–900.

Brown, A. L. (1975). The development of memory: Knowing, knowing about knowing, and knowing how to know. In H. W. Reese (Ed.), *Advances in child development and behavior* (Vol. 10, pp. 327–431). New York: Academic Press.

Chiesi, H. L., Spilich, G. J., & Voss, J. F. (1979). Acquisition of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, *18*, 275–290.

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2009). Mix me a list: Context moderates the truth effect and the mere-exposure effect. *Journal of Experimental Social Psychology*, *45*, 1117–1122.

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, *14*(2), 238–257.

Gatto, J. T. (2001). *A different kind of teacher: Solving the crisis in American schooling*. Berkeley, CA: Berkeley Hills Books.

Gigerenzer, G. (1984). External validity of laboratory experiments: The frequency-validity relationship. *American Journal of Psychology*, *97*, 285–295.

Gude, C., & Zechmeister, E. B. (1975). Frequency judgments for the "gist" of sentences. *American Journal of Psychology*, *88*, 385–396.

Harrison, A. A. (1977). Mere exposure. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 39–83). New York: Academic Press.

Hasher, L., & Chromiak, W. (1977). The processing of frequency information: An automatic mechanism? *Journal of Verbal Learning and Verbal Behavior*, *16*, 173–184.

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*, 107–112.

Hertwig, R., Gigerenzer, G., & Hoffrage, U. (1997). The reiteration effect in hindsight bias. *Psychological Review*, *104*, 194–202.

Krugman, H. E. (1977). Memory without recall, exposure without perception. *Journal of Advertising Research*, *17*, 7–12.

Levine, T. R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, *33*(4), 378–392.

Meyers, J. (1993). More is indeed better. *Media Week*, 3, 14.

Nadarevic, L., & Erdfelder, E. (2013). Spinoza's error: Memory for truth and falsity. *Memory & Cognition*, *41*, 176–186.

Parks, C. M., & Toth, J. P. (2006). Fluency, familiarity, aging, and the illusion of truth. *Aging, Neuropsychology, and Cognition, 13*, 225–253.

Renner, C. H., Renner, M. J., & Shaffer, V. A. (2004). "I've heard that before, therefore I know it": The validity effect in the classroom. *Unpublished manuscript*.

Schwartz, M. (1982). Repetition and rated truth value of statements. *American Journal of Psychology*, *95*, 393–407.

Street, C. N. H., & Richardson, D. C. (2015). Descartes versus Spinoza: Truth, uncertainty, and bias. *Social Cognition*, *33*, 1–12.

Tesser, A. (1978). Self-generated attitude change. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 289–338). New York: Academic Press.

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*(1), 219–230.

Unkelbach, C., Bayer, M., Alves, H., Koch, A., & Stahl, C. (2010). Fluency and positivity as possible causes of the truth effect. *Consciousness and Cognition, 20*(2), 594–602.

Vossel, S., Thiel, C. M., & Gereon, R. F. (2006). Cue validity modulates the neural correlates of covert endogenous orienting of attention in parietal and frontal cortex. *NeuroImage*, *32*(3), 1257–1264.

Zajonc, R. B. (1968). The attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*, 1–27.

14

MERE EXPOSURE EFFECT

Robert F. Bornstein and Catherine Craver-Lemley

Folk wisdom tells us that "familiarity breeds contempt", but studies suggest otherwise. Beginning with the work of Titchener (1910), psychologists have been intrigued by the possibility that repeated, unreinforced exposure to a stimulus would result in increased liking for that stimulus. Zajonc (1968) coined the term *mere exposure effect* (MEE) to describe this phenomenon, and since the publication of Zajonc's seminal (1968) paper, there have been more than 400 published studies of the MEE. The MEE occurs for a broad array of stimuli (e.g., drawings, photographs, musical selections, real words, nonsense words, ideographs) under a variety of laboratory and real-world conditions. Bornstein's (1989) meta-analysis of research on the MEE indicated that the overall magnitude of the effect (expressed in terms of the correlation coefficient *r*) was 0.26, a moderate effect size. Subsequent investigations have confirmed this result (e.g., Gillebaart, Förster, & Rotteveel, 2012; Monahan, Murphy, & Zajonc, 2000; Seamon, McKenna, & Binder, 1998).

Without question, repeated exposure to a stimulus biases our attitude regarding that stimulus: Even though the stimulus itself remains the same, the way we think and feel about the stimulus changes as we become familiar with it (see Chapter 13 for a related discussion). In this respect, researchers agree that the MEE represents a form of cognitive *bias*. But is it a genuine cognitive *illusion*? Is our attitude regarding a repeatedly exposed stimulus changed so profoundly that we can no longer perceive and judge the stimulus accurately, no matter how much effort we devote to the task? Several decades of research can help us resolve this question.

Examples

There are numerous everyday instances of increased liking following repeated exposure to a stimulus. As these examples illustrate, not only does repeated exposure affect our attitude regarding a stimulus, but the process is so subtle that in most

cases we are unaware that mere exposure played a role in altering our judgments and feelings.

Repetition and liking for music

Several MEE experiments have shown that repeated exposure to unfamiliar music leads to more positive ratings of this music (Harrison, 1977; Moreland & Topolinski, 2010). Similar patterns emerge in real-world settings. In fact, the impact of repeated exposure on music sales is so strong that it is often illegal (and always unethical) for radio and Internet show hosts to accept any sort of compensation from music companies, for fear that this will bias song selection and produce an exposure-induced spike in sales.

Exposure and preference for novel types of art

When Impressionist paintings were first displayed publicly, they received scathing reviews. The same thing occurred when Cubist and Expressionist works first appeared. An initial negative reaction occurs almost any time a new art form emerges, but over time – and with repeated viewings – aesthetic judgments shift, and attitudes regarding the now-familiar style become more positive. What was once despised is now embraced.

Unfamiliar people

To a surprising degree, we affiliate with people we encounter most frequently. This is why first-year college students' friendship patterns are determined in part by housing proximity, and why our attitudes regarding other morning commuters become more positive over time (even if we never exchange a word with our fellow traveler). Mere exposure to an unfamiliar person enhances our attitude toward that person.

Relevance

The most obvious applications of MEE principles are in product sales, and marketing researchers have incorporated findings from mere exposure research into a number of contemporary advertising programs (Janiszewski, 1993; Ruggieri & Boca, 2013). Along similar lines, studies suggest that frequency of exposure is a significant determinant of the number of votes garnered by a candidate for elected office, even when other factors (e.g., popularity of the candidate's policy positions) are controlled for statistically (Bornstein, 1989). The impact of repeated exposure on election outcome is not just statistically significant, but ecologically significant as well: The 5–10% shift in voting behavior attributable to candidate familiarity is enough to alter the outcome of many real-world elections.

Another potentially important application of MEE principles and methods concerns intergroup behavior. Beginning with the work of Amir (1969), psychologists have investigated the degree to which repeated, unreinforced exposure could enhance the attitudes of different groups toward each other. Findings in this area have been mixed: Although mere exposure can enhance the attitudes of unfamiliar groups, it does not produce a parallel effect – and sometimes even leads to increased tension and conflict – in groups who have initial negative attitudes (Bornstein, 1993). History is replete with examples of neighboring groups for whom decades of exposure have only heightened hostility (e.g., Israelis and Palestinians).

Research methods

Designs

MEE studies use two types of designs: naturalistic and experimental. Each has certain advantages and certain disadvantages as well.

Naturalistic designs

Naturalistic MEE studies examine the relationship between the naturally occurring frequency of a stimulus and people's attitudes regarding that stimulus. Thus, common names receive more positive liking ratings than do uncommon names, and familiar foods are rated more positively than unfamiliar ones (Bornstein, 1989; Harrison, 1977). The primary advantage of a naturalistic design is that it provides a good approximation of naturally occurring MEEs. The primary disadvantage of a naturalistic design is that it does not allow firm conclusions to be drawn regarding causal relationships between exposure and affect: It may be that common names become better liked because people are exposed to them more frequently, but it is also possible that people are inclined to give their children names that are popular to begin with.

Experimental designs

In experimental MEE studies, participants are exposed to varying numbers of exposures of unfamiliar stimuli (e.g., novel photographs, nonsense words), after which they report how much they like each stimulus. Most experimental studies of the MEE use within-participants designs, so each person is exposed to an array of stimuli at different frequencies. For example, a participant might rate six different stimuli, with each stimulus having been exposed 0, 1, 2, 5, 10, or 20 times during the familiarization phase of the study. In these investigations affect ratings typically become more positive with increasing exposure frequency, leveling off after about 10 exposures then diminishing somewhat at higher exposure frequencies (Bornstein, 1989; Gillebaart et al., 2012).

The primary advantage of an experimental design is that it allows strong conclusions to be drawn regarding the causal relationship between stimulus exposures and

subsequent affect ratings. The primary disadvantage of an experimental design is its artificiality: Because novel stimuli are presented under highly controlled laboratory conditions, the degree to which these findings generalize to real-world situations is open to question.

Measures

A key aspect of MEE research is assessing participants' attitudes regarding stimuli that vary in familiarity. Three types of measures have been used.

Likert ratings

The most common outcome measure in MEE research is a Likert-type rating of each stimulus. Many different rating dimensions have been used (e.g., liking, pleasantness, attractiveness, interestingness), with the specific rating dimension based on the type of stimulus being investigated. Thus, liking ratings are commonly employed when people (or photographs of people) are used as stimuli; pleasantness or interestingness ratings are often employed when paintings or music selections are used.

Likert ratings are not only the most common MEE outcome measure; they are also the most sensitive. Participants' liking ratings of a merely exposed stimulus typically shift by 1 or 2 points on a 9-point scale (e.g., Bornstein, Kale, & Cornell, 1990; Caruso, Vohs, Baxter, & Waytz, 2013). Though this degree of attitude change may seem trivial, it is not: If unfamiliar stimuli receive neutral (midpoint) ratings, a one-point positive shift represents a 20% increase in liking for a familiarized stimulus.

Forced-choice preference judgments

Some MEE studies use forced-choice preference judgments in lieu of Likert-type ratings (e.g., Mandler, Nakamura, & Van Zandt, 1987). In these studies, participants are asked to choose which of two stimuli they like better during the rating phase of the study, with one member of each stimulus pair being previously exposed, and the other being novel. Although forced-choice judgments are less sensitive than Likert-type ratings, they are a better approximation of preference judgments in vivo (e.g., wherein a person must choose between two similar products that vary in familiarity).

Behavioral measures

A small number of MEE studies have used behavioral outcome measures in lieu of self-reports (e.g., Bornstein, Leone, & Galley, 1987; Jones, Young, & Claypool, 2011). Behavioral outcome measures include agreement with familiarized and unfamiliarized confederates in a laboratory negotiation task, voting behavior in a campus election, electrodermal responses to familiar versus novel stimuli, and

willingness to sample different types of food. Most behavioral outcome measures in MEE studies take the form of dichotomous decisions (e.g., choosing between two foods), but on occasion, behavioral outcome measures are analogous to Likert-type ratings (e.g., when percentages of agreement with familiar and unfamiliar people are used; see Bornstein et al., 1987).

Moderating variables

Researchers have examined the impact of numerous moderating variables on the MEE. These can be grouped into three categories: (1) stimulus variables; (2) exposure variables; and (3) participant variables. Assessment of moderating variables is not only useful in understanding the parameters of the MEE, but also in testing competing theoretical models. Different frameworks make contrasting predictions regarding the impact of various moderating variables, and the most influential models are those that have shown good predictive power in this domain.

Two general procedures have been used to assess the impact of moderating variables on the MEE: individual experiments (e.g., Murphy & Zajonc, 1993) and meta-analytic reviews of the mere exposure literature (Bornstein, 1989, 1992). Individual experiments allow for *direct* assessment of the impact of a particular variable by comparing the magnitude of the exposure effect under different conditions (e.g., for complex versus simple stimuli). Meta-analyses allow for *indirect* assessment of the impact of a moderating variable by comparing the magnitude of the MEE across different studies (e.g., those that used a brief delay between exposures and ratings versus those that used a longer delay). As is true of research in many areas of psychology, some moderating variables have been assessed within MEE studies, others have been assessed by contrasting outcomes across studies, and still others have been evaluated using both procedures (see Moreland & Topolinski, 2010, for an extensive review).

Stimulus variables

Two stimulus variables have been assessed by MEE researchers: type of stimulus (e.g., photograph versus drawing) and stimulus complexity.

Type of stimulus

Ten different types of stimuli have been used in MEE studies: nonsense words, meaningful words, ideographs, photographs, drawings, auditory stimuli, olfactory stimuli, gustatory (i.e., food) stimuli, actual people, and objects (e.g., toys). Studies contrasting the magnitude of the MEE as a function of stimulus type have generally found no consistent differences across stimulus classes (e.g., Stang, 1974, 1975; Suzuki & Gyoba, 2008). Meta-analytic data support this result, confirming that different types of stimuli produce comparable exposure effects (Bornstein, 1989).

Stimulus complexity

The majority of experiments that compare the magnitude of the MEE produced by simple versus complex stimuli find that complex stimuli yield stronger exposure effects (Berlyne, 1970; Bornstein et al., 1990). Two processes are involved. First, complex stimuli typically produce a more rapid increase in liking at lower exposure frequencies (i.e., 1, 2, and 5 exposures). Second, complex stimuli produce a less pronounced downturn in liking at higher exposure frequencies (i.e., 10 or more exposures). It appears that simple stimuli are less interesting to begin with (hence, the less rapid increase in liking at lower frequencies), and become boring more quickly at higher exposure frequencies (leading to an "overexposure effect").

Exposure variables

The most widely studied exposure variables in MEE studies are number of presentations, stimulus exposure sequence, stimulus exposure duration, and delay between exposures and ratings.

Number of presentations

MEE studies typically present stimuli a maximum of 50 times, although there is considerable variability in this area (Bornstein, 1989). In most studies MEE researchers obtain an increase in liking ratings through about 10 stimulus exposures, after which ratings plateau, and gradually decline to baseline (Kail & Freeman, 1973; Stang, 1974). These frequency-liking patterns characteristic of individual MEE experiments were confirmed in Bornstein's (1989) meta-analysis, which found that across different stimuli and rating dimensions, the strongest MEEs occurred following a maximum of 5–9 stimulus exposures.

Exposure sequence

Significantly stronger MEEs are obtained when stimuli are presented in a heterogeneous (i.e., randomized) sequence than a homogenous (i.e., massed) sequence during the familiarization phase of the study (Bornstein, 1989; Gillebaart et al., 2012). Consistent with the results of individual experiments, meta-analytic comparisons indicated that while heterogeneous exposures produce a robust MEE ($r = 0.30$), homogenous exposures do not ($r = -0.02$).

Exposure duration

There is an inverse relationship between stimulus exposure duration and magnitude of the exposure effect (Bornstein, 1989). Studies that use stimulus exposures less than 1 sec produce an overall MEE (r) of 0.41, whereas studies that use stimulus exposures between 1 and 5 sec produce an MEE of 0.16, and those that use longer exposures produce an MEE of 0.09. Individual studies comparing MEEs in

identical stimuli presented at different exposure durations support this meta-analytic result (e.g., Hamid, 1973).

Delay between exposure and rating

Seamon, Brody, and Kauff (1983), and Stang (1975) found stronger exposure effects with increasing delay between stimulus exposures and ratings. These results not only indicate that delay enhances the MEE, but confirm that MEEs can persist for up to one week (Seamon et al., 1983) or two weeks (Stang, 1975) following stimulus exposures.

Meta-analytic data confirm these experimental results (Bornstein, 1989) and further indicate that naturalistic MEE studies (which examine affect ratings of stimuli whose frequency varies naturally in vivo) produce a stronger exposure effect (r = 0.57) than do laboratory studies (r = 0.21). The particularly strong MEEs produced by real-world stimuli (e.g., common names) are in part a consequence of the comparatively long delays between stimulus exposures and affect ratings in naturalistic settings.

Participant variables

Participant variables have been studied less frequently than other moderating variables in MEE investigations, but in certain respects these variables have yielded the most intriguing results. Researchers have examined the effects of stimulus awareness, imagery, and individual difference (i.e., personality) variables on the magnitude of the MEE.

Stimulus awareness

More than a dozen published studies have obtained robust exposure effects for stimuli that are not recognized at better-than-chance levels (e.g., Huang & Hsieh, 2013; Kunst-Wilson & Zajonc, 1980; Murphy & Zajonc, 1993; Seamon et al., 1983). Not only do subliminal stimuli produce robust MEEs, but meta-analysis of the MEE literature indicates that stimulus awareness actually inhibits the MEE. Experiments using stimuli that were not recognized at better-than-chance accuracy produce an overall MEE of 0.53, whereas experiments using briefly presented, recognized stimuli produce an overall MEE of 0.34. The magnitude of the MEE produced by stimuli that were recognized at 100% (or close to 100%) accuracy is 0.12 (Bornstein, 1989, 1992).

The inverse relationship between stimulus recognition accuracy and magnitude of the MEE has been replicated in individual experiments as well (Bornstein & D'Agostino, 1992, 1994). For example, Bornstein and D'Agostino (1992) found that photographs and Welsh figures (i.e., simple line drawings) presented for 5 ms during the exposure phase of a typical MEE experiment produced a significantly greater increase in liking than did identical stimuli presented for 500 ms during the exposure

phase. (Follow-up data confirmed that 5 ms stimuli were not recognized at better-than-chance level, whereas 500 ms stimuli were recognized at close to 100% accuracy.)

Additional support for the existence of robust MEEs in the absence of stimulus awareness comes from studies of neurologically impaired participants (e.g., Alzheimer's patients, patients with Korsakoff's syndrome). These experiments confirm that even when neurological deficits obviate explicit memory for previously seen stimuli, robust exposure effects are obtained (Halpern & O'Connor, 2000). In fact, these results are so consistent and compelling that researchers now view MEE-type affect ratings as one of the most reliable indicators of implicit memory for previously encountered stimuli (Marin-Garcia, Ruiz-Vargas, & Kapur, 2013; Whittlesea & Price, 2001).

Imagery effects

Given that MEEs persist for up to two weeks in laboratory studies, and almost indefinitely in vivo, repeated exposure to a stimulus must lead to the construction of a mental representation of that stimulus – a representation that is encoded deeply enough to be maintained from exposures through affect ratings (Mandler et al., 1987). In a compelling demonstration of the impact of mental imagery on the MEE, Craver-Lemley and Bornstein (2006) found that when participants were exposed to the ambiguous duck-rabbit figure and instructed to visualize the image as a duck or as a rabbit consistently throughout exposures, they showed the typical MEE at test only for a disambiguated version of the figure that matched the mental image they generated during encoding (see also Compton, Williamson, Murphy, & Heller, 2002; Rhodes, Halberstadt, & Bracjkovich, 2001, for additional evidence that merely exposed stimuli are conceptually categorized during familiarization).

Along somewhat similar lines, Bornstein, Craver-Lemley, and Alexander (2013) explored the possibility that self-generated mental images would produce exposure effects comparable to those produced by exposure-based mental images. This hypothesis was confirmed: Repeatedly exposed and repeatedly imagined stimuli yielded comparable MEEs. These results dovetail with reports that self-generated images may share properties in common with images that are actually perceived (Craver-Lemley, Arterberry, & Reeves, 1999). In a related experiment, Bornstein et al. (2013) found that self-generated imagery can moderate – or even obviate – the MEE: When participants were instructed to generate positive or negative images during repeated exposures of photographs, subsequent affect ratings of the individuals pictured in the photographs were biased in the direction of these self-generated images (this occurred despite the fact that participants were not asked to generate images during the rating phase of the experiment).

Individual differences

Several individual difference variables have been examined in MEE studies, including need for approval, alexithymia, manifest anxiety, tolerance of ambiguity,

evaluation apprehension, boredom-proneness, and sensation-seeking. For the most part, these variables had modest moderating effects, with two exceptions. Bornstein et al. (1990) found that boredom-prone participants produced significantly weaker MEEs than did non-boredom-prone participants. Kruglanski, Freund, and Bar-Tal (1996) found that high levels of evaluation apprehension undermined the MEE.

Example of a mere exposure experiment

This section describes Bornstein et al.'s (1990) Experiment 2, illustrating two important principles relevant to a broad array of laboratory and real-world exposure effects: (1) the moderating impact of stimulus complexity; and (2) the downturn in the frequency–affect curve that often occurs after many stimulus exposures. A simplified version of this experiment may be used as a classroom demonstration (see Text box 14.1).

TEXT BOX 14.1 MERE EXPOSURE CLASSROOM DEMONSTRATION

This is a simplified version of Bornstein et al.'s (1990) Experiment 2. It focuses on the mere exposure effect for relatively small frequencies and its possible downturn for larger frequencies.

Method

Participants

Because MEE effect sizes are typically moderate, an ideal sample size for this experiment is about 80 participants when alpha is set at .05. Gender does not moderate the MEE, so the distribution of women and men is unimportant (55 women and 45 men participated in the original experiment).

Materials

Deviating from the original experiment, only one set of stimuli is used. These stimuli consist of six line-drawn visual illusions taken from Gregory (1968): the Hering illusion, Wundt's converse of the Hering illusion, the Nekker illusion, the Zollner illusion, the Poggendorf illusion, and a reversible figure–ground drawing. (These are all available on the Internet.)

The stimulus set contains 43 slides. Stimuli are presented at the following frequencies: 0, 1, 2, 5, 10, or 25 (the original study also included a 50-exposure condition). Order of stimuli within the stimulus set is random, and counterbalancing is used to ensure that different stimuli are presented at different

frequencies in different participants. Across participants, each stimulus should appear in each frequency condition approximately the same number of times. In Bornstein et al.'s (1990) experiment, booklets were used during the rating phase of the experiment; these consisted of one copy of each visual illusion per page, along with two 9-point rating scales for each stimulus: like–dislike, and simple–complex. Each rating scale was anchored with the terms *Not at all*_____ (1) and *Very*_____ (9). Rating scales appeared directly below the relevant stimulus. Within each booklet, stimuli were presented in random order. You can follow this procedure as well, or obtain ratings by presenting the stimuli on PowerPoint, one at a time, after the stimulus exposure portion of the experiment is complete.

Design

This demonstration uses a one-factor within-participants design: Each participant provides ratings of stimuli at all six exposure frequencies. The primary dependent measure is participants' like–dislike ratings.

Procedure

Participants can be tested in class. The experimenter provides standardized instructions:

> This is a study of people's responses to visual stimuli. You will be presented a series of images one at a time, and you should examine each image as it's presented. After all the images have been presented, I'll ask you some questions about your reactions to the stimuli. There are about 40 stimuli in all, and this part of the experiment will take about 4 minutes.

After answering any final questions, the experimenter presents the stimuli on PowerPoint slides, one stimulus per slide. Exposure times for each stimulus may be manually controlled by the experimenter, or set automatically so that PowerPoint advances each slide after 5 seconds.

Immediately following stimulus presentations, participants are given the rating booklet, and asked to provide ratings of each stimulus. Participants circle the number on each rating scale corresponding to their rating of the stimulus pictured on that page.

Analysis

The analysis consists of a one-factor within-participants analysis of variance (ANOVA), with stimulus exposure frequency (0, 1, 2, 5, 10, and 25) as the independent variable and participants' like–dislike ratings as the dependent variable.

Results

Results of this experiment should parallel those of Bornstein et al. (1990, Exp. 2) as summarized in Figure 14.1. Liking ratings of visual illusions should increase through five exposures, and then gradually decline to baseline (i.e., 0-frequency level). Statistically, there should be a significant main effect of exposure frequency (with liking ratings increasing through five exposures, then declining).

FIGURE 14.1 Effects of stimulus type and exposure frequency on liking ratings of merely exposed stimuli (from "Boredom as a limiting condition on the mere exposure effect" by Robert F. Bornstein, Amy R. Kale, and Karen R. Cornell, 1990, *Journal of Personality and Social Psychology*, Volume 58, p. 795. Copyright © 1990 by the American Psychological Association. Adapted with permission of the publisher).

Method

The experiment tested 100 participants with two sets of stimuli. *Simple* stimuli consisted of seven line drawings (Figures 8, 10, 20, 33, 42, 55, and 66) from the Barron-Welsh Art Scale (Barron & Welsh, 1949). *Complex* stimuli consisted of seven line-drawn visual illusions taken from Gregory (1968). Within each stimulus category, stimuli were presented at the following frequencies: 0, 1, 2, 5, 10, 25, or 50. Order of stimuli within the stimulus set was random, and counterbalancing was used to ensure that different stimuli are presented at different frequencies in different participants. Across participants, each stimulus appeared in each frequency condition approximately the same number of times.

The stimuli were presented with a slide projector exposing each stimulus for 5 seconds. Subsequent to the presentation phase, participants rated the seven stimuli from each set on two 9-point rating scales: like–dislike and simple–complex, both rating scales anchored with the terms *Not at all* (1) and *Very* (9).

Results

The results are summarized in Figures 14.1 and 14.2. As Figure 14.1 shows, liking ratings of visual illusions increased through five exposures, then gradually declined to baseline (i.e., 0-frequency levels). Liking ratings of Welsh figures increased slightly through 5 exposures then declined below baseline levels at higher exposure frequencies. Statistically, a 2 × 7 within-participants ANOVA showed (1) a significant main effect for stimulus type, $F(1, 99) = 98.88$, $p < .0001$ (with visual illusions receiving more positive ratings than Welsh figures); (2) a significant main effect of exposure frequency, $F(6, 594) = 17.79$, $p < .0001$ (with liking ratings of both types of stimuli increasing through five exposures, then declining); and (3) a significant Stimulus Type × Exposure Frequency interaction, $F(6, 594) = 2.44$, $p < .05$ (with visual illusions showing a more rapid increase in liking than Welsh figures through five stimulus exposures).

Two follow-up ANOVAs assessed the effect of stimulus type and exposure frequency on participants' liking ratings. The first ANOVA assessed the effect of

FIGURE 14.2 Effects of stimulus type and exposure frequency on complexity ratings of merely exposed stimuli (from "Boredom as a limiting condition on the mere exposure effect" by Robert F. Bornstein, Amy R. Kale, and Karen R. Cornell, 1990, *Journal of Personality and Social Psychology*, Volume 58, p. 796. Copyright © 1990 by the American Psychological Association. Adapted with permission of the publisher).

stimulus type and exposure frequency on liking ratings at 0, 1, 2, and 5 exposures; the second assessed the effect of these variables on liking ratings at 5, 10, 25, and 50 exposures. The first ANOVA yielded a significant interaction between stimulus type and exposure frequency, with liking ratings of visual illusions increasing more rapidly than liking ratings of Welsh figures through five exposures. The second ANOVA yielded significant main effects for stimulus type and exposure frequency, but no interaction: Liking ratings of visual illusions and Welsh figures both declined at higher exposure frequencies, with visual illusions continuing to receive more positive ratings than Welsh figures through 50 exposures (see Figure 14.1).

Figure 14.2 summarizes the effects of stimulus type and exposure frequency on simple–complex ratings. As Figure 14.2 shows, there was a significant main effect of stimulus type on complexity ratings, with visual illusions receiving higher complexity ratings than Welsh figures at all exposure frequencies.

Discussion

The results of this experiment illustrated three aspects of the MEE: (1) Liking increased with increasing stimulus exposures. This is the classic MEE, and it is reflected in the significant increase in liking for both types of stimuli at lower exposure frequencies. (2) The stimulus type moderated the MEE. As noted earlier, complex stimuli tend to yield stronger MEEs than do simple stimuli. This is reflected in the significant Stimulus Type × Exposure Frequency interaction at lower exposure frequencies. (3) The downturn in liking ratings for both types of stimuli illustrates the "overexposure effect": At higher exposure frequencies, stimuli become predictable and boring, and as a result, liking ratings decline.

Neurological correlates

Paralleling findings obtained with Alzheimer's and Korsakoff's patients, studies have shown that robust MEEs are obtained in patients who suffer from transient global amnesia (Marin-Garcia et al., 2013). Along somewhat similar lines, Greve and Bauer (1990) found that patients suffering from prosopagnosia prefer familiarized faces over novel ones, though they do not recognize the familiarized faces as having been seen before. Both sets of results are consistent with findings indicating that robust MEEs occur even in the absence of conscious awareness of stimulus exposures (e.g., Monahan et al., 2000).

Two studies have found enhanced MEEs for faces presented to the left visual field (and therefore processed primarily in the right cerebral hemisphere; Compton et al., 2002; Zarate, Sanders, & Garza, 2000). In the only study to date that used functional Magnetic Resonance Imaging (fMRI) to examine the neurological correlates of the exposure effect, Kongthong, Minami, and Nakauchi (2014) found increased gamma activity in the parietal-occipital region in participants who showed strong MEEs. These studies, taken together, suggest some promising avenues for continued

research, but as yet do not offer definitive conclusions regarding the neurological underpinnings of the MEE.

Theoretical accounts

Though the neurological underpinnings of the MEE remain open to question, considerable effort has been devoted to delineating the psychological processes that underlie the effect. Since publication of Zajonc's seminal (1968) paper, more than a dozen theoretical frameworks have been developed to explain the psychological processes that help account for the MEE (see Bornstein, 1989, 1992; Moreland & Topolinski, 2010; Seamon et al., 1998; Whittlesea & Price, 2001; Zajonc, 2001). Five of these models have been particularly influential.

1. The arousal model

Berlyne's (1970) arousal model contends that unfamiliar stimuli are unpleasant because they produce high levels of physiological arousal. Over time, familiarity-induced habituation leads to less and less arousal following each additional stimulus exposure, while at the same time an opponent-process "rebound effect" causes the participant to experience increased liking for the now-familiar stimulus.

Studies support Berlyne's (1970) prediction that unfamiliar stimuli lead to high levels of arousal, in part because of their unpredictability (Kruglanski et al., 1996). Studies also confirm that arousal in response to stimulus presentations diminishes at higher levels of exposure. Despite these supportive findings, Berlyne's (1970) model is incomplete. It cannot account for the moderating effects of stimulus complexity or homogenous-heterogenous exposure sequence on the MEE, nor can it explain the downturn in affect ratings that occurs at higher exposure frequencies.

2. The nonspecific activation model

Mandler et al.'s (1987) nonspecific activation model contends that MEEs result from activation of previously encoded stimulus representations. The basic premise of this perspective is that repeated exposures lead to increasingly elaborated mental images of a stimulus (see also Craver-Lemley, Bornstein, Forys, Lake, & Thomas, 2002). When participants are asked to provide liking ratings during the test phase of the study, the elaborated stimulus representations are easily primed, and participants interpret the resulting ease of processing as evidence that they like the stimulus (cf. Chapter 10 on availability).

A key prediction of the nonspecific activation model is that MEEs should occur for a variety of stimulus judgments, including (but not limited to) affect ratings. In support of this prediction, Mandler et al. (1987) demonstrated that repeated exposure to polygon stimuli led to increases in judgments of stimulus brightness – and stimulus darkness – in addition to the usual increases in liking ratings.

3. The two-factor model

Stang's (1974) two-factor model contends that MEEs reflect two interacting processes: learning and boredom. Learning leads to increased liking for a stimulus at lower exposure frequencies, as the participant becomes familiar with the properties of the stimulus. Boredom leads to a downturn in the frequency-liking curve at higher exposure frequencies, as the stimulus becomes predictable and uninteresting. The first portion of the two-factor model parallels closely the earlier framework of Berlyne (1970), but emphasizes cognitive processing of the stimulus rather than physiological arousal per se. The other important contribution of Stang's (1974) framework is the prediction that boredom will undermine the exposure effect.

Myriad experiments demonstrating that stronger exposure effects are obtained for complex than simple stimuli support this latter prediction of Stang's (1974) two-factor model (Bornstein, 1989): Simple stimuli do indeed become boring more easily than complex stimuli with repeated exposure. Bornstein et al.'s (1990) results provide further evidence for Stang's (1974) framework: Not only do complex stimuli produce stronger MEEs than simple stimuli, but participants who score high on a measure of boredom-proneness show weaker exposure effects than participants who are not boredom-prone (Bornstein et al., 1990, Exp. 1).

4. The perceptual fluency/attributional model

Bornstein and D'Agostino's (1992, 1994) perceptual fluency/attributional (PF/A) model interprets the MEE in terms of increased perceptual fluency (i.e., ease of perceptual processing) for repeatedly exposed stimuli. Consistent with the perspectives of Seamon et al. (1983) and Mandler et al. (1987), the PF/A model contends that participants in typical MEE studies misattribute perceptual fluency to increased liking for a stimulus. The PF/A model extends earlier thinking in this area by positing that to the degree that participants attribute increased fluency to the stimulus familiarization procedure (rather than to properties of the stimulus itself), they will adjust their liking ratings downward, inferring that their reactions to the stimulus are the result of repeated exposure.

The initial portion of the PF/A model is a variation of Mandler et al.'s (1987) hypothesis that repeated exposures lead to the construction of increasingly elaborated mental representations of a stimulus (see also Craver-Lemley et al., 2002). The latter ("attributional") portion of the PF/A model is supported by findings that indicate that: (1) subliminal stimuli produce significantly stronger MEEs than do clearly recognized stimuli; (2) delay between stimulus exposures and ratings enhances the effect; and (3) naturalistic MEE studies yield stronger exposure effects than do laboratory MEE studies (Bornstein, 1989, 1992). All three variables – subliminality, experimentally determined delay, and in vivo delay – interfere with participants' ability to attribute familiarity to stimulus exposures and prevent them from adjusting downward their liking ratings of the stimuli.

5. The affective primacy model

Zajonc (1980) argued that MEEs represent a "pure" affective response that occurs with minimal intervening cognitive activity beyond rudimentary encoding of stimulus properties. The existence of MEEs in primates and other mammals supports the affective primacy hypothesis, and – as noted earlier – in recent years some progress has been made in identifying the neurological underpinnings of exposure-based affective responding in humans, even in the absence of higher-level cognitive processing of stimulus elements (Marin-Garcia et al., 2013; Zarate et al., 2000).

Zajonc's (1980) affective primacy hypothesis is consistent with findings demonstrating robust MEEs for subliminal stimuli (Murphy & Zajonc, 1993), and with results showing affective "spillover" effects to related – and even unrelated – stimuli following repeated, unreinforced exposures (Caruso et al., 2013; Monahan et al., 2000). Bornstein et al.'s (2013) finding that repeated association of merely exposed stimuli with positive or negative images altered participants' affective reactions is also consistent with the affective primacy hypothesis. The affective primacy hypothesis does not account for the downturn in liking ratings at higher exposure frequencies.

Conclusions

Few psychologists question the robustness of the MEE, but researchers continue to debate the processes that underlie the effect. Over time, researchers have divided into two camps with respect to this issue: those who favor an affect-based model of the MEE, and those who focus on the cognitive processes that mediate and moderate the effect. Compelling evidence has been obtained in support of both positions, and in certain respects these two viewpoints are actually quite compatible. It may be that MEEs occur in stages, the first of which is a "pure" affective response that requires minimal cognitive processing beyond rudimentary encoding of stimulus properties. This initial affective response is then moderated by more extensive cognitive processing of the mental representation of the merely exposed stimulus. A key challenge for researchers during the coming years will be to specify the links between affective and cognitive components of the MEE.

Whatever psychological and neurological processes underlie the MEE, there is no doubt that this phenomenon has important implications for a broad array of psychological phenomena. In the cognitive arena, the MEE paradigm has been increasingly applied to the investigation of implicit memory and schema-priming effects (e.g., Whittlesea & Price, 2001). Social researchers have used MEE procedures to examine the impact of familiarity on intergroup attitudes and behaviors (Kruglanski et al., 1996). Developmental psychologists have become interested in a very different aspect of the MEE: Because infants actually show a reverse MEE (i.e., preference for novel over familiar stimuli), while toddlers and older children show typical exposure effects, developmentalists have begun to explore the processes that delay the onset of the MEE beyond the first two years of life (Berg & Sternberg, 1985).

In this context it is worth noting that the MEE overlaps to some degree with two other phenomena that may also qualify as cognitive illusions. The validity effect (Chapter 13), which refers to the impact of repetition on the perceived validity (that is, truthfulness) of information, represents an example of the way that mere repeated exposure alters people's perceptions of messages (e.g., political statements). Moreover, the validity effect is consistent with the PF/A model of the MEE, because when repeated messages are processed more easily than unfamiliar ones, ease of processing is misattributed to the veracity/truthfulness of the message (see, e.g., Renner & Renner, 2001). Along somewhat similar lines, the recognition heuristic may represent yet another example of the misattribution of perceptual fluency to other types of judgments (see Hilbig, Erdfelder, & Pohl, 2010). Although the recognition heuristic typically refers to the impact of familiarity on inferences (i.e., the conclusion that a true or correct solution exists) rather than preferences, Oeusoonthornwattana and Shanks (2010) obtained preliminary results extending the recognition heuristic from inferences to preferences.

One of the most stunning findings in MEE research is the inverse relationship between stimulus awareness and magnitude of the exposure effect. This result – which has now been obtained in more than a dozen independent laboratories – played a major role in establishing the existence of perception without awareness in mainstream psychology (Bornstein, 1992; Zajonc, 2001). The robustness of subliminal MEEs continues to intrigue researchers interested in automaticity, implicit perception, and the dynamics of unconscious mental processing (Huang & Hsieh, 2013; Moreland & Topolinski, 2010).

The question remains: Given what we know about the MEE, can this phenomenon be described as a genuine cognitive illusion? The answer to this question is yes. Robust MEEs are produced with a complete absence of stimulus recognition on the part of participants (Zajonc, 2001). Even in situations where participants are aware of having been exposed to stimuli, they rarely attribute their liking for a stimulus to repeated exposure, instead believing that some property of the stimulus itself is particularly attractive or interesting (Bornstein & D'Agostino, 1994). It is here that the crux of the illusion lies: Although repeated exposure does not alter a stimulus at all, it alters attitudes regarding that stimulus. Insofar as people attribute their positive attitude to properties of the stimulus – not familiarity with the stimulus – the true source of this positive attitude remains unknown, and the illusion remains strong.

Summary

- The mere exposure effect (MEE) refers to increased liking for a stimulus that follows repeated, unreinforced exposure to that stimulus.
- MEEs are obtained for a wide variety of stimuli (e.g., visual, auditory), in a broad array of contexts (e.g., laboratory, field).
- MEEs have numerous real-world implications, helping explain voting behavior, advertising effects, and attitudes toward people and objects encountered in everyday life.

- Boredom is a limiting condition on the MEE: Simple stimuli and a homogeneous exposure sequence weaken the effect, and liking ratings tend to decrease after a large number of stimulus exposures.
- Stimulus awareness inhibits the MEE: Stronger effects are produced by stimuli perceived without awareness than those that are consciously recognized.
- Myriad theoretical models have attempted to explain the MEE, and it appears that the effect is a product of two processes: A rapid, reflexive affective response followed by more controlled, deliberate cognitive processing of stimulus content.

Further reading

Zajonc's (1968) classic monograph summarizes the history of the mere exposure effect (MEE) and the relationship of the effect to other psychological phenomena; Moreland and Topolinski's (2010) extensive review provides an excellent summary of contemporary MEE research. Kunst-Wilson and Zajonc's (1980) experiment has served as a model for most subliminal MEE studies during the past 35 years, while Bornstein's (1989) meta-analysis remains one of the most useful references on the parameters and limiting conditions of the MEE; it not only identifies variables that moderate the effect, but also illustrates how MEEs differ in different situations and settings. More recently, Whittlesea and Price's (2001) experiments have demonstrated how participants' information processing strategies can enhance or undermine the effect, Zajonc's (2001) review discusses the implications of the MEE for contemporary models of unconscious mental processing, and Kongthong et al.'s (2014) study provides preliminary evidence regarding cortical activity patterns that may mediate the MEE.

References

Amir, Y. (1969). Contact hypothesis in ethnic relations. *Psychological Bulletin*, *71*, 319–342.

Barron, F., & Welsh, G. S. (1949). *Barron-Welsh art scale*. Palo Alto, CA: Consulting Psychologists Press.

Berg, C. A., & Sternberg, R. J. (1985). Response to novelty: Continuity and discontinuity in the developmental course of intelligence. In C. Schooler & K. W. Schaie (Eds.), *Advances in child development and behavior* (Vol. 19, pp. 1–47). New York: Academic Press.

Berlyne, D. E. (1970). Novelty, complexity, and hedonic value. *Perception and Psychophysics*, *8*, 279–286.

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin*, *106*, 265–289.

Bornstein, R. F. (1992). Subliminal mere exposure effects. In R. F. Bornstein & T. S. Pittman (Eds.), *Perception without awareness: Cognitive, clinical, and social perspectives* (pp. 191–210). New York: Guilford Press.

Bornstein, R. F. (1993). Mere exposure effects with outgroup stimuli. In D. M. Mackie & D. L. Hamilton (Eds.), *Affect, cognition, and stereotyping* (pp. 195–211). New York: Academic Press.

Bornstein, R. F., Craver-Lemley, C., & Alexander, D. N. (2013). Mental imagery moderates the mere exposure effect: Impact of self-generated images on affect ratings of faces. *Journal of Mental Imagery*, *3/4*, 1–12.

Bornstein, R. F., & D'Agostino, P. R. (1992). Stimulus recognition and the mere exposure effect. *Journal of Personality and Social Psychology*, *63*, 545–552.

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.

Bornstein, R. F., Leone, D. R., & Galley, D. J. (1987). The generalizability of subliminal mere exposure effects: Influence of stimuli perceived without awareness on social behavior. *Journal of Personality and Social Psychology*, *53*, 1070–1079.

Caruso, E. G., Vohs, K. D., Baxter, B., & Waytz, A. (2013). Mere exposure to money increases endorsement of free-market systems and social inequality. *Journal of Experimental Psychology: General*, *142*, 301–306.

Compton, R. J., Williamson, S., Murphy, S. G., & Heller, W. (2002). Hemispheric differences in affective response: Effects of mere exposure. *Social Cognition, 20*, 1–17.

Craver-Lemley, C., Arterberry, M. E., & Reeves, A. (1999). "Illusory" illusory conjunctions: The conjoining of features of visual and imagined stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1036–1049.

Craver-Lemley, C., & Bornstein, R. F. (2006). Self-generated visual imagery alters the mere exposure effect. *Psychonomic Bulletin & Review*, *13*, 1056–1060.

Craver-Lemley, C., Bornstein, R. F., Forys, K., Lake, C., & Thomas, N. (2002). *Repeatedly exposed and imagined stimuli yield comparable exposure effects*. Paper presented at the annual meeting of the Eastern Psychological Association, Boston, MA.

Gillebaart, M., Förster, J., & Rotteveel, M. (2012). Mere exposure revisited: The influence of growth versus security cues on evaluations of novel and familiar stimuli. *Journal of Experimental Psychology: General*, *141*, 691–714.

Gregory, R. L. (1968). Visual illusions. *Scientific American*, *219*, 66–76.

Greve, K. W., & Bauer, R. M. (1990). Implicit learning of new faces in prosopagnosia: An application of the mere exposure paradigm. *Neuropsychologia*, *28*, 1035–1041.

Halpern, A. R., & O'Connor, M. G. (2000). Implicit memory for music in Alzheimer's disease. *Neuropsychology*, *14*, 391–397.

Hamid, P. N. (1973). Exposure frequency and stimulus preference. *British Journal of Psychology*, *64*, 569–577.

Harrison, A. A. (1977). Mere exposure. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 39–83). New York: Academic Press.

Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2010). One-reason decision making unveiled: A measurement model of the recognition heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 123–134.

Huang, Y. F., & Hsieh, P. J. (2013). The mere exposure effect is modulated by selective attention but not visual awareness. *Vision Research*, *91*, 56–61.

Janiszewski, C. (1993). Preattentive mere exposure effects. *Journal of Consumer Research*, *20*, 376–392.

Jones, I. F., Young, S. G., & Claypool, H. M. (2011). Approaching the familiar: On the ability of mere exposure to direct approach and avoidance behavior. *Motivation and Emotion*, *35*, 383–392.

Kail, R. V., & Freeman, H. R. (1973). Sequence redundancy, rating dimensions, and the exposure effect. *Memory & Cognition*, *1*, 454–458.

Kongthong, N., Minami, T., & Nakauchi, S. (2014). Gamma oscillations distinguish mere exposure from other likeability effects. *Neuropsychologia*, *54*, 129–138.

Kruglanski, A. W., Freund, T., & Bar-Tal, D. (1996). Motivational effects in the mere exposure paradigm. *European Journal of Social Psychology*, *26*, 479–499.

Kunst-Wilson, W. R., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, *207*, 557–558.

Mandler, G., Nakamura, Y., & Van Zandt, B. J. (1987). Nonspecific effects of exposure to stimuli that cannot be recognized. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 646–648.

Marin-Garcia, E., Ruiz-Vargas, J. M., & Kapur, N. (2013). Mere exposure effect can be elicited in transient global amnesia. *Journal of Clinical and Experimental Neuropsychology*, *35*, 1007–1014.

Monahan, J. L., Murphy, S. T., & Zajonc, R. B. (2000). Subliminal mere exposure: Specific, general, and diffuse effects. *Psychological Science*, *11*, 462–466.

Moreland, R. L., & Topolinski, S. (2010). The mere exposure phenomenon: A lingering melody by Robert Zajonc. *Emotion Review*, *2*, 329–339.

Murphy, S. T., & Zajonc, R. B. (1993). Affect, cognition, and awareness: Priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology*, *64*, 723–739.

Oeusoonthornwattana, O., & Shanks, D. R. (2010). I like what I know: Is recognition a non-compensatory determiner of consumer choice? *Judgment and Decision Making*, *5*, 310–325.

Renner, C. H., & Renner, M. J. (2001). But I thought I knew that: Using confidence estimation as a debiasing technique to improve classroom performance. *Applied Cognitive Psychology*, *15*, 23–32.

Rhodes, G., Halberstadt, J., & Brajkovich, G. (2001). Generalization of mere exposure effects to averaged composite faces. *Social Cognition*, *19*, 57–70.

Ruggieri, S., & Boca, S. (2013). At the roots of product placement: The mere exposure effect. *Europe's Journal of Psychology*, *9*, 246–258.

Seamon, J. G., Brody, N., & Kauff, D. M. (1983). Affective discrimination of stimuli that are not recognized: Effect of delay between study and test. *Bulletin of the Psychonomic Society*, *21*, 187–189.

Seamon, J. G., McKenna, P. A., & Binder, N. (1998). The mere exposure effect is differentially sensitive to different judgment tasks. *Consciousness and Cognition*, 7, 85–102.

Stang, D. J. (1974). Methodological factors in mere exposure research. *Psychological Bulletin*, *81*, 1014–1025.

Stang, D. J. (1975). Effects of mere exposure on learning and affect. *Journal of Personality and Social Psychology*, *31*, 7–12.

Suzuki, M., & Gyoba, J. (2008). Visual and tactile cross-modal mere exposure effects. *Cognition and Emotion*, *22*, 147–154.

Titchener, E. B. (1910). *A textbook of psychology.* New York: MacMillan.

Whittlesea, B. W. A., & Price, J. R. (2001). Implicit/explicit processing versus analytic/nonanalytic processing: Rethinking the mere exposure effect. *Memory & Cognition*, *29*, 234–246.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology Monographs, 9*(*2*, Part 2), 1–27.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35*, 151–175.

Zajonc, R. B. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science*, *10*, 224–228.

Zarate, M. A., Sanders, J. D., & Garza, A. A. (2000). Neurological dissociations of social perception processes. *Social Cognition*, *18*, 223–251.

15

HALO EFFECTS

Joseph P. Forgas and Simon M. Laham

The halo effect is a general cognitive bias in impression formation. It can be defined as a tendency of judges to assume that once a person possesses some known good (or bad) characteristics, their other, unrelated and unknown characteristics are also likely to be consistent, that is, good or bad. For example, one of the most common instances of the halo effect is when people tend to judge physically attractive people as having better personal and intellectual qualities than unattractive persons, even though they have no relevant direct evidence about these inferred qualities. It seems that physical attractiveness casts a 'halo' over general impressions about the target, biasing judgments on other attributes in a positive direction. The halo effect is conceptually similar to other kinds of constructive cognitive illusions characterized by a general confirmation bias, including, for example, stereotype effects, prototype effects and the like (see also Chapter 5 in this volume).

The term 'halo effect' was first suggested by Thorndike (1920) to describe this phenomenon. The concept is essentially a metaphor derived from medieval, renaissance and byzantine religious imagery, where a glowing circle or 'halo' is depicted as floating above the heads of saints indicating their general goodness and chosen status.

A 'negative halo effect' also exists, sometimes known as the devil effect or the horns effect. In the case of negative halo effects, a single negative attribute can bias subsequent impression formation judgments on unrelated dimensions in a negative direction. In general, then, halo effects can be identified in situations whenever a judge learns that a target possesses attribute A, and this knowledge subsequently produces a diffuse positive (halo) or negative (devil) influence on subsequent judgments on unknown attributes B, C, D, etc.

A related phenomenon, known as halo error, has also been identified in the literature (Thorndike, 1920). This occurs when unjustifiably inflated positive correlations are observed among multiple attribute ratings about a target (or inflated inter-category correlations; see Cooper, 1981, for a review; cf. Chapter 6 in this volume on illusory correlations). In contrast to the halo effect, a halo error may also occur in situations when judges do not know anything specific about how a target is positioned on any single attribute.

Halo effects can influence a wide variety of social judgments and decisions, but most typically, halo effects have been studied by social psychologists interested in impression formation and social judgments (Asch, 1946). In contrast, the halo error has been of greater interest to psychometricians, concerned with the psychometric properties of multi-attribute rating instruments and inflated inter-attribute correlations.

In practice, the halo effect is typically measured by presenting participants with information about a target's position on one attribute, followed by an examination of the influence of this information on subsequent ratings on one or more unrelated attributes. The halo error, on the other hand, is typically studied by asking judges to simultaneously rate a target on multiple dimensions and then assessing inter-correlations among these ratings. The focus of this chapter is on the halo effect, although we refer to the halo error on occasion, to clarify certain distinctions.

Typical experiment and class illustration

A typical recent experiment illustrating the halo effect was published by Forgas (2011). In that study, participants were asked to form impressions on multiple dimensions about the writer of a brief philosophical essay, as well as the quality of the essay. The halo effect was manipulated by also providing judges with an alleged 'photo' of the writer, showing either a middle-aged, bespectacled man (positive halo, 'typical' academic philosopher), or the image of a young, casually dressed woman (negative halo, an 'atypical' philosopher). It was expected and found that the physical appearance of the writer in the photo would produce a general halo effect, resulting in more positive judgments when the writer appeared to be a middle-aged male, rather than a young female (see Figure 15.1).

This study also contained another manipulation, inducing judges into either a positive or a negative mood before forming impressions. It was expected and found that positive mood should recruit a more heuristic, superficial and simplified information processing strategy that is conducive to constructive cognitive illusions such as the halo effect. In contrast, negative mood triggered a more detailed, systematic and attentive information processing style that reduced the observed halo effects. A further interesting real-life illustration of the halo effect can also be viewed on YouTube at the following URL: www.psychologyconcepts.com/halo-effect.

FIGURE 15.1 Halo effects produced by the appearance of the writer on evaluating the competence of the author of a short philosophical essay: A middle-aged male was rated as more competent than a young female. The mood of the judges also influenced the size of the halo effect: Positive mood, producing more heuristic and assimilative processing, increased, and negative mood, producing more systematic and accommodative processing, reduced the size of the halo effect. (Adapted from "She just doesn't look like a philosopher . . .? Affective influences on the halo effect in impression formation", by J. P. Forgas, 2011, *European Journal of Social Psychology*, Volume 41, p. 815. Copyright © 2011 by Wiley Publishers. Adapted with permission of the publisher.)

Overview of research on halo effects

History

Halo biases were first discussed in any detail in the early twentieth century by Thorndike (1920). He was among the first to notice that when supervisors were asked to judge their subordinates on multiple attributes, surprisingly, the intercorrelations between their judgments were "all higher than reality" (p. 25), indeed, "too high and too even" (p. 27). It was also Thorndike who first coined the terms halo error and halo effect to describe the emergence of this unexpected pattern by judges to overestimate the strength of the expected relationship between different attitudes, and the tendency to infer the valence of unknown personal qualities from known characteristics when forming impressions about people.

The next major contribution to the study of halo effects is associated with Solomon Asch's (1946) classic studies on impression formation. In a series of simple but ingenious experiments, Asch found that known personality traits exert a significant influence on other traits when forming impressions. For example, Asch found that important central traits of a person, such as his or her warmth or coldness, produce a disproportionate, 'radiating' effect on the way other traits are interpreted and the way impressions are formed on previously unknown traits. Asch (1946)

interpreted these effects in terms of a classical Gestalt theoretical orientation. He argued that perceivers are not merely passive information processors of incoming information. Rather, they are actively and subconsciously engaged in constructing judgments that are coherent and consistent with expectations, and result in a holistic impression characterized by good shape or form ('Gestalt').

Classic halo experiments

Physical attractiveness halo effects

Perhaps the most common and best known halo effect is associated with physical attractiveness, first demonstrated by Dion, Berscheid, and Walster (1972). It is rather interesting that the psychological consequences of such an obvious and immediately observable characteristic as a person's appearance were not investigated until the 1970s. Dion and her colleagues in their classic study presented participants with photos of young females who were previously rated as physically attractive, average, or unattractive. Participants who viewed these images were then asked to form impressions about, and rate these targets on a variety of characteristics such as their personality, expected future personal and occupation success, life satisfaction and happiness. Of course, if people were completely rational and unbiased information processors free of any cognitive illusions, they should simply not be able to perform this task. After all, how could one properly judge the personality or happiness of an individual, based on nothing more than how they happen to look on a photo!?

In fact, participants in Dion et al.'s (1972) experiment had no difficulty at all in performing this apparently unreasonable impression formation task. The only way they could form impressions about personality based on physical appearance was by relying on the cognitive illusion of the halo effect: inferring the existence of other positive or negative qualities based only on how a person looks. Results showed that, as expected, physically attractive targets were in fact judged to have significantly more socially desirable individual characteristics on a variety of dimensions, including personality, expected future occupational success, and happiness (see Table 15.1).

As Table 15.1 shows, even though attractive-looking women were thought to have better personalities, to be more happy and competent, and more likely to marry – somewhat unexpectedly, it was average-looking women who were thought to make more competent parents! One might wonder why parenthood in particular was judged to be unrelated to good looks? Perhaps judges might have thought that attractive women might find it easier to find alternative relationships, compromising their future parental competence?

Dion et al.'s (1972) study thus suggests that halo effects do not spread uniformly, equally and indiscriminately to all other inferred characteristics. Rather, the halo effect is moderated by the nature and content of the judgmental dimensions used. This pattern of trait-specificity suggests that halo effects also depend on judges' 'implicit theories of personality', what they already know and expect about the

TABLE 15.1 The effects of a person's physical attractiveness on perceptions of other characteristics

Rated characteristics	*Physical appearance of target*		
	Unattractive	*Average*	*Attractive*
Social desirability of personality	56.31	62.42	65.39
Occupational status	1.70	2.02	2.25
Marital competence	0.37	0.71	1.70
Parental competence	3.91	4.55	3.54
Social and professional happiness	5.28	6.34	6.37
Total happiness	8.83	11.60	11.60
Likelihood of marriage	1.52	1.82	2.17

Note. Higher numbers correspond to more positive judgments. (Adapted from "What is Beautiful is Good," by Dion et al., 1972, *Journal of Personality and Social Psychology*, 24, p. 288. Copyright 1972 by the American Psychological Association.)

relationship between different personality traits. Subsequent research also confirmed that the influence of halo effects is partly dependent on the particular judgmental dimension studied (e.g., Forgas, O'Connor, & Morris, 1983; Sigall & Ostrove, 1975).

Much subsequent work on the halo effect in social psychology has focused on this attractiveness halo or the 'what is beautiful is good' effect (see Eagly, Ashmore, Makhijani, & Longo, 1991, for a review). Numerous studies confirmed that the physical attractiveness of a target can reliably exert a significant influence on impressions on such unrelated characteristics as perceived social and intellectual competence, happiness and success, and even ascribed responsibility for transgressions (Eagly et al., 1991).

Physical attractiveness has many more surprising consequences as well, and good-looking people often seem to get preferential treatment in a variety of areas as a result of halo effects. For example, Landy and Sigall (1974) found that the same essay will be evaluated more positively by men when the writer is depicted as an attractive rather than an unattractive-looking woman (Table 15.2). This halo effect occurred for both good and poor quality essays. However, good-looking females received even more benefit from their markers when their essay happened to be rather poor rather than strong!

Physical attractiveness can also exert a halo effect on how negative behaviors and transgressions are evaluated, and the seriousness of the preferred punishments. In general, good-looking people are less likely to be held responsible for their transgressions than are unattractive people. Dion (1972) reported evidence suggesting that a transgression committed by an unattractive child was judged as more serious and the child held more responsible and the transgression was thought more likely to occur again than the same transgression when committed by an attractive-looking child. The suggested class demonstration proposed in Text box 15.1 was designed to reproduce very similar results.

TABLE 15.2 The effects of a female's physical attractiveness on male judges' ratings of an essay written by her

Variable	*Physical attractiveness of the writer*			
	Attractive	*Control*	*Unattractive*	*Total*
Ratings of essay quality				
Good essay	6.7	6.6	5.9	6.4
Poor essay	5.2	4.7	2.7	4.2
Total	6.0	5.5	4.3	
Ratings of writer's overall ability				
Good essay	6.4	6.3	6.0	6.2
Poor essay	5.7	4.7	3.4	4.6
Total	6.5	5.6	4.7	

Note. Higher numbers indicate more positive evaluations on a scale from 1 to 10. (Adapted from "Beauty is Talent: Task Evaluation as a Function of the Performer's Physical Attractiveness," by D. Landy and H. Sigall, 1974, *Journal of Personality and Social Psychology*, 29, p. 302. Copyright 1974 by the American Psychological Association.)

TEXT BOX 15.1 CLASSROOM DEMONSTRATION

A simpler classroom demonstration of the halo effect can also rely on the manipulation of the physical attractiveness of a target person to influence subsequent impressions on unrelated qualities. In the demonstration suggested here, participants will be asked to read a brief description of an everyday incident involving a transgression by a child, and then form impressions about the target, who is presented as either a physically attractive or a physically unattractive individual. Participants (should be about 20–25 persons per condition to produce a reliable effect) can be either students in a class or people interviewed at random in public places. They are randomly assigned to either the positive halo condition or the negative halo condition.

Materials and procedure

Participants are presented with a one-page description of a brief incident describing a child causing injury to another child. Halo effects are manipulated by describing the perpetrator as either physically attractive, with a common positively evaluated name, or physically unattractive with an unusual name. Photos of attractive or unattractive-looking children may also be used as an additional halo manipulation. Impressions about the incident and the perpetrator are then assessed as a function of the halo manipulations, as

illustrated by the following questionnaire that can be used directly to demonstrate halo effects:

Below you will find a brief description of an episode involving a small child. After reading the story, please answer the questions below.

[*Positive halo condition:*] Susie is a very attractive-looking four-year-old girl. She has beautiful blonde hair and lovely blue eyes.

[OR]

[*Negative halo condition:*] Manga is a rather unattractive-looking four-year-old girl. She has greasy hair and close set eyes.

The other day when she was playing with a neighbor's little four-year-old boy, she threw a stone at him that hurt his arm so badly that he had to be taken to a hospital. Imagine for a moment that you have just witnessed this incident. Do you think that in these circumstances . . .

Susie / Manga intended to hurt the little boy?	Yes 1__2__3__4__5__6 No
She should be punished?	Yes 1__2__3__4__5__6 No
She is likely to do it again?	Yes 1__2__3__4__5__6 No
She is likely to be an intelligent child?	Yes 1__2__3__4__5__6 No
Would you allow your child to play with her?	Yes 1__2__3__4__5__6 No

Analysis

Mean differences in ratings between the positive and the negative halo conditions can be easily compared along each of the judgmental dimensions. The significance of the halo effect can also be tested using simple *t*-tests of individual scales or a *t*-test of combined average judgments across all rating scales.

The physical attractiveness halo effect also appears to emerge fairly early in childhood, supporting the possibility that a deep-seated evolutionary mechanism may also be involved in the 'what is beautiful is good' phenomenon. In a remarkable study, Dion (1973) found that 3- to 6 1/2-year-old pre-schoolers could reliably discriminate between the physically attractive and unattractive facial photographs of peers who by adult standards were considered attractive or unattractive. Also, these very young participants already showed a significant preference for attractive children as potential friends and a corresponding dislike of unattractive children. The children also thought that more attractive children were more likely to behave pro-socially, while unattractive children were perceived as more likely to exhibit antisocial behaviors.

In another experiment illustrating judicial halo effects, Efran (1974) asked university students to play the role of members of a university disciplinary court. They were asked to make decisions about another student charged with misconduct, such

as cheating in an exam. In a clear illustration of halo effects, the judges were less inclined to believe the charges, and awarded less severe punishment when the defendant was good-looking rather than plain-looking.

However, there are also some interesting limits to halo effects produced by physical attractiveness, as the work of Dion et al. (1972) already suggested when assessing judgments of parental competence. In some contexts (such as perceptions of parenting skill) physical attractiveness appears to have limited halo effects. Indeed, the 'what is beautiful is good' expectation may not work and may even produce counterproductive effects when a person appears to be relying on his/her physical attractiveness to commit an antisocial act. In a 1975 study, Sigall and Ostrove found that a physically attractive person was held more responsible and actually given more severe punishment when she potentially could have used her attractiveness to commit a crime (swindle). However, a positive halo was still found and the same person was treated more leniently than others when her crime, although more serious (burglary), was not related to her physical attractiveness (see Figure 15.2).

Physical attractiveness is an enduring personal quality, so it is perhaps not surprising that it can exert such a powerful and reliable halo effect on impressions in so many areas. Would more fleeting and temporary attractiveness signals, such as smiling or not smiling, also produce similar halo effects? Forgas et al. (1983) tested this possibility, and found that smiles can also reliably trigger halo effects. Smiling targets were rated more favorably on unrelated dimensions, and were also given less severe punishments for a transgression than were the same people when not smiling (Figure 15.3). In addition to physical attractiveness and smiling, other and more diffuse 'attractiveness' signals can also produce halo effects. For example, Nisbett and

FIGURE 15.2 The effects of a female's physical attractiveness on sentences for crimes in which attractiveness did (swindle) or did not (burglary) play a role. (Adapted from "Beautiful but dangerous: Effects of offender attractiveness and nature of crime on juridic judgments," by H. Sigall and N. Ostrove, 1975, *Journal of Personality and Social Psychology*, Volume 31, p. 412. Copyright © 1975 by the American Psychological Association. Adapted with permission of the publisher.)

FIGURE 15.3 The effects of facial expression on responsibility attribution and the evaluation of a person who committed a transgression. Higher numbers correspond to less responsibility attributed and more positive evaluation. (Adapted from "Smile and punishment: The effects of facial expression on responsibility attributions by groups and individuals," by Forgas et al., 1983, *Personality and Social Psychology Bulletin*, Volume 9, p. 591. Copyright © 1983 by the American Psychological Association. Adapted with permission of the publisher.)

Wilson (1977) showed that the appearance, mannerisms, and accent of targets were judged more positively when they behaved in a warm and friendly rather than a cold, distant way.

The general robustness of physical attractiveness halo effects is further illustrated by research suggesting that attractiveness halo effects may also be transferable to persons other than the original target. Sigall and Landy (1973) reported an interesting study where the physical attractiveness of a good-looking female partner was found to exert a positive influence on impressions formed about her male consort. It seems than that the 'radiating beauty' halo effect is not limited to the actual target person, but may also spread to others intimately associated with an attractive target.

Other halo effects

Several further intriguing demonstration of halo effects have been reported using a variety of attributes. For example, Wilson (1968) found that information about the high or low academic status of a previously unknown guest lecturer (described as either a professor or a student) exerted a significant halo effect on observers' estimates of his physical height. Results showed that estimates of the lecturer's perceived physical "height increased with increasing academic status" (p. 99), even though observers had ample opportunity to actually observe the guest lecturer in action. This experiment can also be readily adapted for class demonstration, by asking

participants to guess the height of a person shown in a photo who is described as having either high or low academic or social status.

Interestingly, usual or unusual personal names can also produce remarkable halo effects (see the class demonstration in Text box 15.1). In one study, Harari and McDavid (1973) asked school teachers to grade compositions written by children who had either regular, positively evaluated names (David, Michael), or had unusual, negatively evaluated names (Elmer, Hubert). Although the essays were identical, 'Elmer' and 'Hubert' received significantly worse marks. Name halos have also been shown to influence judgments of likeability and electability in a recent study by Laham, Koval, and Alter (2012).

Halo effects induced by names can even distort the kinds of professional judgments one would expect to be immune from such cognitive biases. Luke Birmingham, a psychiatrist, asked over 400 British psychiatrists to diagnose a 24-year-old man who assaulted a conductor based on a written description of the case. When his name was shown as 'Mathew', over 75% of the psychiatrists gave him a sympathetic hearing, diagnosing him as suffering from schizophrenia and requiring treatment. When he was identified as 'Wayne', he was much more likely to be diagnosed negatively as a malingerer, a drug abuser and suffering from a personality disorder (reported in Gross, 2015).

Boundary conditions

However, there are also some important boundary conditions that limit the universality of halo effects. Dion et al.'s (1972) classic study already indicated that physical attractiveness halo effects do not spread equally to every judgmental scale (such as parenting competence). Such a pattern is also broadly consistent with the results of Solomon Asch's (1946) classic Gestalt experiments on impression formation, where the halo effect exerted by important central traits varied depending on other known traits, as well as the rating dimensions used.

Subsequently, Eagly et al. (1991) confirmed that the impact of physical attractiveness halos partly depends on the rating dimensions used. Halo effects were largest for ratings of social competence, followed by intellectual competence, with ratings of concern for others and integrity producing the smallest effects. It does appear then that at least in perceivers' minds, physical attractiveness does not seem to go hand in hand with high integrity and responsibility for others.

The availability of additional contextual information may not only limit but actually even reverse halo effects. For example, Sigall and Ostrove (1975) reported that the positive halo effects exerted by physical attractiveness cues when judging the seriousness of a transgression disappeared when judges had reason to believe that a transgressor actually used her good looks to commit a crime. Although physically attractive transgressors were judged more leniently when allegedly committing a burglary (a positive halo effect), physical attractiveness resulted in a reverse effect and more severe punishment when targets may have taken advantage of their physical attractiveness to commit a crime, such as a swindle.

Findings such as Sigall and Ostrove (1975) also suggest that the occurrence of halo effects may well depend on the kind of information processing strategy adopted by judges. When impressions are formed in a simple, automatic, heuristic manner, we may expect halo effects to flourish. However, more careful, systematic and attentive processing of the available information may well reduce halo effects. A recent experiment by Forgas (2011) produced specific evidence supporting this prediction. In this study, information processing style was manipulated indirectly, by inducing positive or negative mood in judges. Prior research showed that positive mood produces more heuristic, constructive processing, while negative mood recruits a more systematic and attentive thinking style. Results supported the predictions. A clear halo effect was found in positive mood, but negative mood reduced the halo effect.

Consequences and practical importance

The practical consequences of halo effects can be widespread and highly significant in everyday life. Once unjustified initial expectations are formed about a person, they can easily become self-perpetuating with serious implications for how a target is treated (Harari & McDavid, 1973). If we expect a person to have positive characteristics, we may selectively look for and find such features from the rich array of information available (a self-fulfilling prophecy), and positive impressions may in turn lead to preferential treatment in a range of domains: interpersonal relations, the work place, the health and legal systems, and even for decision making and consumer choices. As we have seen, even halos elicited by such obviously irrelevant cues as a person's name can influence liking, preferential educational assessments, and even psychiatric diagnoses.

Physical attractiveness halos, in particular, are extremely common and salient and may have implications for how people are treated in the workplace. Even from early childhood, more attractive people are more likely to be perceived positively (Dion, 1973), and as adults, they are more likely to be hired than less attractive people (Watkins & Johnston, 2000) and are more likely to be paid more (Hammermesh & Biddle, 1994).

Of particular relevance to the judicial system are studies that show that attractive perpetrators are judged less guilty and receive less severe punishments for cheating than are less attractive perpetrators (Efran, 1974; Forgas et al., 1983). Within the health system, Martin, Friedmeyer, and Moore (1977) found that more attractive inpatients suffering from schizophrenia were rated as emotionally better adjusted than were less attractive fellow patients. Further, Napoleon, Chassin, and Young (1980) found that less attractive patients received more severe diagnoses and had longer lengths of stay as inpatients.

One of the most common examples of the operation of halo effects in real life is when the opinions and views of well-known people and celebrities are sought out on topics utterly unrelated to their domains of competence. There seems to be unrelenting public and media interest in the pronouncements and opinions on almost anything by celebrities, famous athletes, pop singers, or actors. Why do

people assume that just because a person is well-known for some achievement (or indeed, for no achievement at all, in the case of many celebrities!), their views on almost anything are still worth knowing? We can understand this common phenomenon in popular culture in terms of the operation of halo effects. Consumers of such celebrity news demonstrate a halo effect when they assume that their idols, having achieved celebrity in one area, should also automatically be experts in other, unrelated areas.

Theoretical explanations and psychological mechanisms

Over the decades, a number of convergent theoretical frameworks have been advanced to explain halo effects. Gestalt theories of perception, applied to social information processing, provided the first such explanatory framework (Asch, 1946). The expectation that positive characteristics should signal the existence of other positive features is consistent with constructive theories of social cognition, and especially with Gestalt theories of perception, suggesting that human perceivers are universally motivated to construct coherent, consistent impressions that show good shape and form. The Gestalt principle offers a broad and parsimonious, albeit rather non-specific theoretical framework that applies to several impression formation effects. Many other constructive impression formation biases demonstrated in the literature, such as primacy effects, salience effects, or 'central trait' effects (Asch, 1946), bear considerable conceptual similarity to halo effects, in that they all demonstrate the radiating influence of some personality traits on others, consistent with the Gestalt principle that perceivers always seek to construct coherent and integrated impressions.

But what exactly constitutes a good 'Gestalt'? What determines the quality and direction of how perceivers extrapolate from one trait to others in halo effects? In other words, what determines the nature of what perceivers see as a meaningful, well-rounded impression? Halo effects may also be thought of as reflecting universal and culturally determined 'implicit theories of personality' we all share, allowing us to link one known personality characteristic to various unknown others, based on shared implicit representations about human nature. Implicit personality theories are helpful in explaining where people's ideas of what a well-formed impression is. For example, research findings suggesting that physical attractiveness is not necessarily related to parental competence (Dion et al., 1972), or that attractive people may not always be champions of social integrity and concern for others (Eagly et al., 1991; Sigall & Ostrove, 1975) can be explained as reflecting shared implicit theories of personality in our culture. Different cultures may well have developed quite different notions about how personality traits are related, suggesting that there is considerable scope for further cross-cultural research on the nature of halo effects in other than individualistic Western societies.

A complementary and even more relevant and parsimonious explanation of many spontaneous halo effects can be based on associative network theories and

the notion of spreading activation among related attributes. In functional terms, our accumulated experience of the world (and that includes our implicit personality theories as well) leads certain attributes or qualities to become more or less strongly associated with each other in memory. As a consequence, the presentation and activation of one attribute (e.g., physical attractiveness) will automatically and selectively prime, through spreading activation, access to other, previously associated attributes (e.g., more positive personality traits, etc.).

Halo effects can also be linked to various dichotomous process theories of social cognition. The tendency to make unwarranted inferences from one observation to others is most likely to be promoted by a kind of simplified, heuristic processing strategy human information processors typically adopt as their effort-minimizing fall-back strategy. Several recent experiments suggest that when a more elaborate, systematic and analytic processing style is triggered (for example, as a result of induced negative affect), halo effects tend to disappear (Forgas, 2011).

Other information processing variables, such as processing fluency, may also play a role in halo effects. Laham et al. (2012), for example, showed that the ease with which names are pronounced cast a halo over a range of impression formation judgments, from likeability to electability. Fluent processing experiences (such as pronouncing easy names) feel subjectively better than more difficult experiences, biasing subsequent judgments (see Chapters 10, 13, and 14 in this volume).

We should also note that many halo effects (e.g., physical attractiveness) also show a close resemblance to the well-known effects of stereotypes and prototypes in social judgments (e.g., Eagly et al., 1991). Just as the activation of a stereotype (e.g., 'Germans') or a person prototype (e.g., 'professor') spontaneously gives rise to a host of expectations about people so categorized, single personality traits ('physically attractive') seem to operate in a similar manner in the case of halo effects. Thus, halo effects can best be understood as a special case of a large class of more general constructive cognitive biases, where expectations are triggered by single traits, rather than by higher-level category labels.

The apparently universal human tendency to categorize and rely on cognitive shortcuts allows us to generalize from one piece of known information to other, unknown qualities. The ubiquity of this phenomenon even suggests that halo effects, as is the case with many other constructive biases, may have a fundamental evolutionary origin (e.g., Langlois et al., 2000). The fact that halo effects already emerge in early childhood certainly seems to support such an evolutionary explanation (Dion, 1972). It is most likely that these different theoretical explanations are complementary, and it is also probable that different psychological processes may be relevant to explaining qualitatively different halo effects.

Moderators and bias reduction

Halo effects appear extremely reliable and pervasive, and it seems that explicit interventions and instructions are not particularly effective in reducing their occurrence. For example, Wetzel, Wilson, and Kort (1981) tried to eliminate the halo effect

using Nisbett and Wilson's (1977) procedure, by (1) asking participants to pay close attention to what they were feeling, (2) giving participants a prior description of the halo effect and telling them to try avoid committing it, or (3) giving a description of the halo effect and instructing participants to commit it. None of these interventions had an influence on the size of the halo effect.

However, more subtle and subconscious manipulations designed to influence the kind of information-processing strategy adopted by judges may well be more effective. In one recent study, Forgas (2011) showed that the halo effect was eliminated when negative mood was induced. This result is also consistent with prior evidence indicating that negative moods reliably recruit a more analytical, systematic and externally focused processing strategy.

Summary

- The halo effect is a cognitive bias in impression formation that occurs when perceivers make unwarranted inferences about the positive or negative qualities of a person based on information about other, unrelated characteristics.
- Halo effects can be triggered by a variety of social characteristics, such as physical attractiveness, social status, having an unusual name, interpersonal style, etc.
- A number of theories of social cognition are relevant to explaining halo effects, such as Gestalt theory, implicit personality theories, dual-process theories and theories of associative learning and memory.
- Halo effects appear difficult to control through explicit instructions, but may be eliminated when more systematic and attentive information processing is adopted.
- Halo effects can have important real-life consequences, creating unwarranted positive or negative expectations about a person that may lead to self-fulfilling prophecies and potentially discriminatory treatment.

Further reading

For classic demonstrations of the halo effect, see Dion et al. (1972) and Wilson (1968). For meta-analytic summaries of the attractiveness halo, see Eagly et al. (1991) and Langlois et al. (2000). For a recent study illustrating the influence of different processing strategies on the halo effect, see Forgas (2011). For a review of the related halo error, see Cooper (1981).

References

Asch, S. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258–290.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*, 218–244.

Dion, K. K. (1972). Physical attractiveness and evaluation of children's transgressions. *Journal of Personality and Social Psychology, 24*, 207–213.

Dion, K. K. (1973). Young children's stereotyping of facial attractiveness. *Developmental Psychology, 9*, 183–188.

Dion, K. K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24*, 285–290.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but . . .: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin, 110*, 109–128.

Efran, M. G. (1974). The effects of physical attractiveness on judgments in a simulated jury task. *Journal of Research in Personality, 8*, 45–54.

Forgas, J. P. (2011). She just doesn't look like a philosopher . . .? Affective influences on the halo effect in impression formation. *European Journal of Social Psychology, 41*, 812–817.

Forgas, J. P., O'Connor, K., & Morris, S. L. (1983). Smile and punishment: The effects of facial expression on responsibility attribution by groups and individuals. *Personality and Social Psychology Bulletin, 9*, 587–596.

Gross, R. (2015). *Psychology: The science of mind and behavior* (7th ed.). London: Hodder Education.

Hammermesh, D. S., & Biddle, J. E. (1994). Beauty and the labor market. *American Economic Review, 84*, 1174–1194.

Harari, H., & McDavid, J. W. (1973). Name stereotypes and teachers' expectations. *Journal of Educational Psychology, 65*, 222–225.

Laham, S. M., Koval, P., & Alter, A. L. (2012). The name pronunciation effect: Why people like Mr. Smith more than Mr. Colquhoun. *Journal of Experimental Social Psychology, 48*, 752–756.

Landy, D., & Sigall, H. (1974). Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. *Journal of Personality and Social Psychology, 29*, 299–304.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin, 126*, 390–423.

Martin, P. J., Friedmeyer, M. T., & Moore, J. E. (1977). Pretty patient – healthy patient? Study of physical attractiveness and psychopathology. *Journal of Clinical Psychology, 33*, 990–994.

Napoleon, T., Chassin, L., & Young, D. (1980). A replication and extension of "Physical attractiveness and mental illness". *Journal of Abnormal Psychology, 89*, 250–253.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence of unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*, 250–156.

Sigall, H., & Landy, D. (1973). Radiating beauty: The effects of having a physically attractive partner on person perception. *Journal of Personality and Social Psychology, 28,* 218–224.

Sigall, H., & Ostrove, N. (1975). Beautiful but dangerous: Effects of offender attractiveness and nature of crime on juridic judgments. *Journal of Personality and Social Psychology, 41*, 410–414.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25–29.

Watkins, L. M., & Johnston, L. (2000). Screening job applicants: The impact of physical attractiveness and application quality. *International Journal of Selection and Assessment, 8*, 76–84.

Wetzel, C. G., Wilson, T. D., & Kort, J. (1981). The halo effect revisited: Forewarned is not forearmed. *Journal of Experimental Social Psychology, 17*, 427–439.

Wilson, P. R. (1968). Perceptual distortion of height as a function of ascribed academic status. *Journal of Social Psychology, 74*, 97–102.

16

OVERCONFIDENCE

Ulrich Hoffrage

When the editor of this book asked me whether I would be interested in contributing a chapter on overconfidence, my first question was "What's the deadline?" When Rüdiger said, "In half a year – end of October" I replied, "Impossible. With some luck this is when I could start writing. If the end of November is okay with you, I'm on board, otherwise I have to say no." At the time my confidence of providing the chapter by the end of November was about 80% and my confidence of being done before Christmas was almost 100%.

Christmas came and I had not yet even started. Milder forms of this "planning fallacy" (Buehler, Griffin, & Ross, 1994) are probably known to many of us, yet, it is ironic that such a thing occurred with this chapter on overconfidence, as this is just one illustration of this phenomenon (and it is embarrassing to admit that something similar also happened for the second edition of this book). More generally, overconfidence occurs if our confidences related to our judgments, inferences, or predictions are too high when compared to the corresponding accuracy. For the current review, I commence with (1) a brief overview of the three most frequently used tasks and measures, then (2) present classroom demonstrations, (3) summarize some major findings, (4) introduce and evaluate models and theoretical accounts, (5) discuss the functional value of overconfidence, (6) briefly mention its relevance in applied settings, and (7) conclude with some final remarks, a summary, and a list of further literature.

Overconfidence: Tasks and measures

The term *overconfidence* has several meanings, reflecting the different measures that are used to compare subjective beliefs and reality. The three most typical are (1) calibration, (2) the precision of numerical estimates, and (3) people's placement of their own performance relative to others. Moore and Healy (2008) refer to the corresponding types of overconfidence as *overestimation* (average confidence judgments

exceed the percentage of correct statements or choices), *overprecision* (confidence intervals around person's best estimates of numerical variables are too narrow), and *overplacement* (placement in rank orderings is higher than justified), respectively (see also Benoit & Dubra, 2011).

Calibration: Subjective confidences exceed objective accuracy

In an early study, Adams and Adams (1961) asked participants to state their confidence (on a full scale of 0 through 100%) in their recalls of nonsense syllables after 1, 2, 4, 8, or 16 study trials on a list of 30 syllables. The confidence scale was explained in terms of expected percentages of correct recalls, that is, the participants were instructed to assign confidence judgments such that across all instances, in which a confidence of x has been stated, x% of the recalls should be correct. Plotting the percentage of correct recalls separately for each confidence category yields the so-called calibration curve. Data points lined up along the diagonal are said to be well calibrated. Points below indicate overconfidence, that is, unwarrantedly high confidences. For instance, when participants (in one condition) said they were 100% certain that the syllable they recalled was on the list, they were correct in only about 85% of the cases, and when their confidence was 80–90%, they were correct in 55%. Data points above the diagonal indicate underconfidence. For confidences in the truth of the statement "A is larger than B" that range between 0 and 50%, the opposite is the case: Here a percentage of correct answers above the diagonal indicates overconfidence because the stated confidences have been too extreme – conversely, data points below the diagonal indicate underconfidence.

Many, if not most, of the subsequent studies on calibration have focused on general knowledge questions in which participants have to choose which of two alternatives is correct (two-alternative forced-choice tasks) (e.g., "what is absinthe: a precious stone or a liqueur?"). After participants made their choice they are asked to state their confidence in having chosen the correct answer, usually on a half scale of 50 to 100%, with 10% increments.

Measures

For both the full and the half scale, the appropriateness of participants' responses can be evaluated with respect to internal and external criteria. Whereas internal criteria focus on consistency, the external criterion is simply reality. The correspondence between beliefs and reality can be measured with a proper scoring rule, the most popular being the *Brier score* (*B,* named after Brier, 1950):

$$B = \frac{1}{N}\sum_{i=1}^{N}(r_i - c_i)^2 \tag{1}$$

where r is the response on the probability scale, c is the correctness of the statement (1 if correct, 0 if wrong) to which this probability has been attached, and N is the

number of items. The lower the score, the better: If a full scale is used, optimally all correct statements are rated with a subjective probability of 1, and all wrong statements with 0, resulting in a Brier score of 0. If one is uncertain about the truth of a statement (or the correctness of one's choice), the lowest expected Brier score is achieved by stating the true subjective probability.

Several decompositions of the Brier score have been suggested, the most popular by Murphy (1972), which, in the most common case of one response per item, reduces to

$$\begin{aligned} B &= \bar{c}(1-\bar{c}) + \frac{1}{N}\sum_{t=1}^{T} n_t (r_t - \bar{c}_t)^2 - \frac{1}{N}\sum_{t=1}^{T} n_t (\bar{c}_t - \bar{c})^2 \\ &= \textit{Knowledge} + \textit{Calibration} - \textit{Resolution} \end{aligned} \tag{2}$$

where $\bar{c}$ is the percentage correct across all items, r_t is the probability assigned to category t (usually these are fixed by the experimenter: 50%, 60%, . . ., 100%), and $\bar{c}_t$ is $\bar{c}$ computed within confidence category t. These terms can be determined for each participant separately and can be interpreted as follows: *Knowledge* measures how much a participant knows about the particular domain, *calibration* measures the (weighted average of the) correspondence between a particular confidence judgment and the proportion of correct statements in this category, and *resolution* reflects the assessor's ability to sort the events into subcategories for which the percentage correct is different from the overall percentage correct (for a discussion of other measures, see Keren, 1991; Lichtenstein, Fischhoff, & Phillips, 1982).

The measure that is most frequently used in the literature is *over/underconfidence*, which is sometimes also referred to as *calibration-in-the-large* (in contrast to *calibration* as introduced above, which is then referred to as *calibration-in-the-small*). It is simply defined as the difference between percentage correct (across all items) and mean confidence (across all items), specifically:

$$\textit{Over/underconfidence} = \bar{r} - \bar{c} \tag{3}$$

The precision of numerical estimates: Subjective confidence intervals are too narrow

"In what year was the first flight of a hot air balloon?" (Soll & Klayman, 2004). Participants are asked to provide their best guess, that is, the estimate for which they think chances that the true value is above or below this estimate are 50:50. Having received the response, the experimenter continues, "Imagine I tell you that your estimate was too high – what is now your best estimate?" and then "Now imagine that your first estimate was too low – what is now your best estimate?" This is followed by "Now consider again the full range of possible answers. Give me your lowest (highest) estimate – such that your subjective probability for the true value being below (above) this boundary is 1%."

Measures

Let us denote the answers to these questions as x_{50}, x_{25}, x_{75}, x_1, and x_{99}, respectively. The *interquartile index* is the proportion of true values lying between x_{25} and x_{75}, derived from a series of similar estimation tasks, and the *surprise index* is the proportion of true values lying either below x_1 or above x_{99} (occasionally, similar measures such as the proportion of true values between x_{10} and x_{90} are used). A person is well calibrated if the interquartile index is 50% and the surprise index is 2%. Typically, the interquartile index is too low (39% across 27 conditions of several studies reviewed by Lichtenstein et al., 1982) and the surprise index is too high (30% across those conditions). Thus, the confidence intervals are too tight, which indicates that people tend to think their estimates are closer to the true value than they actually are.

Placements in rank orderings: Better-than-average

"Each author thanks his co-author for doing 10 percent of the work on this project." These assessments (and the hidden "I did 90 percent of the work") can be found in the acknowledgement of Santos-Pinto and Sobel (2005, p. 1386). The general finding that these authors seek to explain is that people tend to overestimate their own contribution, performance, or skills relative to others. When asked to assess where they stand relative to a reference population, people exhibit the *better-than-average effect* (see also Chapters 10 and 17). For instance, Svenson (1981) reported that "In the US group 88% and in the Swedish group 77% believed themselves to be safer than the median driver", and "In the US sample 93% believed themselves to be more skillful drivers than the median driver and 69% of the Swedish drivers shared this belief in relation to their comparison group" (p. 146). Variants of the following list of further examples can be found in numerous textbook chapters, internet sites and essays on overconfidence:

- 19% of people think that they belong to the richest 1% of the population.
- 82% of people say they are in the top 30% of safe drivers.
- 80% of students think they will finish in the top half of their class.
- 68% of lawyers in civil cases believe that their side will prevail.
- 86% of my Harvard Business School classmates say they are better looking than their classmates.

Measures

The attentive reader may have noticed that all examples given above refer to a specific percentile of the entire reference population (richest 1%, top 30%, top half) which makes the comparison of the percentage of people who believe they are in this percentile with the percentile itself meaningful. Indeed, if exactly those 50% of the population who believe they are above the median are, in fact, also

above the median, then everyone's belief is accurate. The effect, however, is usually referred to as "better-than-average". Whereas the median splits a distribution in two equally sized parts, this is not necessarily the case for the average. To the contrary, if the distribution is skewed, the average splits the distribution in parts that are necessarily (and can be markedly) different from 50:50. As an example, consider the number of accidents per driver which is obviously not normally distributed but can be approximated better by a Poisson distribution. And indeed, 57% of 440 car drivers in Germany and 80% of 7,842 car drivers in the U.S. were involved in less accidents than the mean number of car accidents per person (Finkelstein & Levin, 2001). If "better-than-average driving skill" is defined as "less accidents than the mean", then these 80% of the U.S. car drivers can be perfectly rational in their belief that they are better than average. Many authors are aware of this terminological problem, many mention it explicitly, and many formulate their instructions in a way that either unambiguously refers to the median, or that makes this confusion between average and median not relevant. Still, the problematic label persists in the literature, and one should take it with a smile if someone points to people's irrationality and biases and then provides the "better-than-average" effect as an example.

Related methods and measures

For quite another way of structuring the methods, measures, and phenomena related to overconfidence the reader is referred to Alba and Hutchinson (2000), who classified the findings they reviewed into the major sections "remembering the past", "interpreting the present", and "predicting the future", thereby discussing many phenomena such as eyewitness testimony, incidental learning, belief polarization, the reiteration effect, and meta-memory (feeling of knowing) in their relation to overconfidence. Another phenomenon that is sometimes related to overconfidence is the *illusion of control*, that is, people's tendency to overestimate the degree of control that they have over processes or outcomes (Glaser & Weber, 2007; see also Chapter 7). Moreover, many authors use terms semantically close to overconfidence, such as *optimism* (which is conceptualized as a more stable individual trait to have positively biased expectations; Weinstein, 1980), *overoptimistic beliefs*, or *hubris* (see also Chapter 17).

Classroom demonstrations

It is easy to demonstrate overconfidence in the classroom. Text box 16.1 describes an experiment that uses the calibration paradigm. Text box 16.2 provides instructions and 10 items for numerical estimates. I use these items regularly in my master class "Managerial Decision Making" to demonstrate overconfidence in the form of overprecision of such estimates and observed so far, aggregated across 1,322 intervals provided by 140 students over several years, an average of only 5.04 instead of 9 hits, as should be.

TEXT BOX 16.1 CALIBRATION IN THE CLASSROOM

Method

Calibration is obtained in each of two item sets that Gigerenzer, Hoffrage, and Kleinbölting (1991) referred to as *representative* and *selected* set. For the *representative* set, Gigerenzer et al. (1991) used a complete paired comparison between 25 (Study 1; 21 for Study 2) German cities that had been randomly drawn from the set of all German cities with more than 100,000 inhabitants. The task was to choose the city with the larger population, and to provide a confidence rating for this choice. It is not necessary to realize a complete paired comparison, but it is important that each object in a given pair has been randomly drawn from a well-specified reference class. It is also not necessary to have as many comparisons as we had in the original studies, but if smaller samples are used, every participant should get a unique, independent random sample. It can be illustrative to use more than one representative set (for various demonstrations and a discussion of why the reference class, and in particular its size, affects overconfidence, see Hoffrage, 2011; Hoffrage & Hertwig, 2006).

The *selected* set can be obtained from the author upon request; however, it is recommended to adapt the procedure used by Juslin (1994), who asked people to generate items they considered appropriate to be included in general knowledge tests. Getting involved in selecting some "good general knowledge items" (p. 236) may help students to understand the results and the theory.

After completing choices and confidence ratings for an item set, participants should estimate the number of correct choices they achieved in a particular set.

Results

Figure 16.1 displays the major result of the original studies. The left panel shows that participants were well calibrated in the representative item set, whereas calibration for the selected item set was poor. The right panel shows that there was no overconfidence for the representative item set, yet overconfidence for the selected item set was substantial. In addition, the right panel also displays the differences between mean frequency estimates (which have been transformed to percentages to be comparable) and percentage correct. With this measure, the picture changed: For the selected set, the frequency estimates matched the percentage of correct answers, whereas participants underestimated their performance in the representative set (for more details, see Gigerenzer et al., 1991).

FIGURE 16.1 Left: Calibration curves for representative and selected item sets. Right: The graphs labeled *Confidence* depict mean confidence judgments minus mean percentage correct, and those labeled *Frequency* depict mean frequency estimates (which have been transformed to percentages to be comparable) minus percentage correct. Positive differences denote overconfidence; negative differences denote underconfidence. Data taken from Gigerenzer et al. (1991).

Discussion

This experiment has contributed two new insights to the field: Overconfidence depends on, first, the sampling procedure, and second, on how performance is assessed. Specifically, in the representative set, confidences were well calibrated, but frequency estimates showed underconfidence; in the selected set, confidences showed overconfidence whereas frequency estimates were well calibrated. Given these results it is too shortsighted to state that people are overconfident in general. These results instead call for models of the cognitive processes that are able to capture the quite differentiated picture displayed in Figure 16.1. The theory of "probabilistic mental models" (Gigerenzer et al., 1991) provides such an approach and will be introduced below.

TEXT BOX 16.2 PRECISION OF NUMERICAL ESTIMATES IN THE CLASSROOM

The following instruction and 10-item list has been taken from Russo and Shoemaker (2004): "For each of the following questions, provide a low and high guess such that you are 90 percent sure the correct answer falls between the two. Your challenge is to select a range that is neither too narrow

(overconfident) nor too wide (underconfident). If you succeed, you should have nine hits and only one miss" (p. 79).

- What is the weight of an empty Airbus A340–600 (in tons)?
- In what year did John Steinbeck win the Nobel Prize for Literature?
- What is the distance (in kilometers or miles) from the Earth to the Moon?
- What is the air distance (in kilometers or miles) from Madrid to Baghdad?
- In what year was the construction of the Roman Colosseum completed?
- What is the height (in meters or feet) of the Aswan High Dam?
- In what year did Magellan's crew complete the first naval circumnavigation of the globe?
- In what year was Mohandas K. Gandhi born?
- What is the surface (in square kilometers or miles) of the Mediterranean Sea?
- What is the gestation (i.e., pregnancy) period of the great blue whale (in days)?

The correct solutions are: 240 tons; 1962; 384,400 km; 4,308 km; AD 80; 114 m; 1522; 1869; 2,510,000 km^2; 335 days.

Typical findings

This section reviews some well-established results. I focus (but not exclusively) on studies using the calibration paradigm. Comprehensive reviews have been provided by Alba and Hutchinson (2000), Keren (1991), Lichtenstein et al. (1982), O'Connor (1989), and Yates (1990); for two reviews that focus more on models than on findings, see McClelland and Bolger (1994) and Juslin and Olsson (1999).

The overconfidence effect

The dominant finding is overconfidence: (1) mean confidence in the correctness of one's answers tends to exceed percentage correct, (2) subjective confidence intervals around numerical estimates are too narrow, and (3) subjective placements on rankings are, on average, too high. Nevertheless, calibration curves usually have a positive slope and, as Alba and Hutchinson (2000) point out, there is still a positive correlation between individuals' reported mean confidences and mean percentages correct (albeit rather low, in one study even as low as .2, but in other studies up to .75). This suggests that using confidences is still predictive when it comes to assessing the quality of the knowledge that led to these confidences.

The hard-easy effect

Overconfidence covaries with item difficulty. In the calibration paradigm, hard item sets (percentage of correct answers of about 75% or lower) tend to produce

overconfidence, whereas easy sets (percentage correct of about 75% or higher) tend to produce underconfidence (Juslin, Winman, & Olsson, 2000). A *hard-easy effect* has also been observed for overplacement. Moore and Cain (2007) reported that the better-than-average effect turns into worse-than-average placements on skill-based tasks that people believe to be difficult. But note that the direction is reversed! Whereas in the calibration paradigm difficult items lead to overconfidence, they lead to underconfidence when participants are asked to rate their performance relative to others (for an explanation of this counterintuitive finding, see below).

The importance of sampling

In sets of items that have been representatively drawn from a natural environment of the participants, confidence judgments tend to be well calibrated. In contrast, in sets of items that have been selected to be hard (or that have been nominated by participants as "good general knowledge items"), confidence judgments tend to be too high, resulting in poor calibration and in overconfidence. Sampling procedure is confounded with item difficulty: Representative item sets tend to be easier than such selected item sets. Nevertheless, the hard-easy effect and sampling are not only conceptually different, they can also be separated on an empirical basis: In a meta-analysis, Juslin et al. (2000) conducted a review of 95 independent data sets with selected items and 35 sets in which items had been sampled representatively. Across all selected item sets, overconfidence was 9%, and across all representative sets it was 1%. The authors pointed out that this difference could not be explained by differences in percentage correct: When they controlled for the end effects of the confidence scale and the linear dependence between percentage correct ($\overline{c}$) and the over/underconfidence score ($\overline{r} - \overline{c}$), the hard-easy effect almost disappeared for the representative item sets.

The base-rate effect

Closely related to the effect of sampling is the so-called *base-rate effect*. Whereas the former is obtained for two-alternative forced-choice tasks and confidences given on a half scale, the latter is obtained for statements and a full scale: When the experimenter manipulates the base rate of true statements, participants are unable to adjust their confidences accordingly, showing severe underconfidence (overconfidence) for item sets with a high percentage of true (wrong) statements (Ferrell & McGoey, 1980; Lichtenstein et al., 1982).

The confidence-frequency effect

In Figure 16.1 (right panel), one can separately compare confidences with accuracy and frequency estimates with accuracy. The direct comparison between these two forms of assessing one's own performance constitutes the *confidence-frequency effect*: In our studies, confidences exceeded frequency estimates by about 15 percentage

points (Gigerenzer, et al., 1991; see also Allwood & Montgomery, 1987; Schneider, 1996). For a similar finding with the interval method, see Cesarini, Sandewall, and Johannesson (2006).

The expertise effect

Although there are exceptions, many studies have shown that experts are well calibrated, at least in their domain of expertise. For instance, Murphy and Winkler (1977) found that weather forecasters were almost perfectly calibrated across the whole range of their subjective probabilities from 0 to 100%. Experts' good calibration seems to be domain specific. That is, in domains outside of their expertise, they are indistinguishable from other people; conversely, other people who had been tested in the experts' field of expertise also fared poorly (for more details, see the reviews mentioned above).

Betting

A variant of the calibration method is to assess people's willingness to bet. Fischhoff, Slovic, and Lichtenstein (1977) asked participants whether they were willing to gamble against the experimenter. In the participant's urn were all items for which he estimated the odds of having answered correctly as 50:1 or higher, and in the experimenter's urn were 100 white and 2 red balls. Whenever the participant drew a question from his urn for which he made the wrong choice, he had to pay $1 to the experimenter, and whenever the experimenter drew a red ball from her urn, she had to pay $1 to the participant. Of 42 participants, 64% decided to gamble (which indicates that they believed the proportion of correctly answered questions was greater than 50:1) and almost all lost. In contrast, Blavatskyy (2009), who implemented a similar incentive-compatible mechanism, classified 65% of his 48 participants as underconfident, and Johnson and Bruce (2001) found that bids from a naturalistic setting, namely the horserace betting market, were almost perfectly calibrated, suggesting that the good calibration of experts seems to be robust across different ways of expressing confidences (probabilities versus odds).

Format dependence

"The population of Bulgaria exceeds 30 million: true or false?" When participants make their own choices for items like this and subsequently state their confidence in the correctness of this choice on a half scale, the calibration curve for this and similar tasks cuts the diagonal at about 75% (with underconfidence in lower and overconfidence in higher confidence categories). However, when the experimenter (randomly) selects one of the alternatives and participants state the confidence in the correctness of that choice on a full scale, the calibration curve cuts the diagonal at about 50% (Juslin, Wennerholm, & Olsson, 1999). Finally, participants can be asked to

provide the smallest confidence interval within which they are, say, 80% certain that the population of Bulgaria lies – with this method participants tend to be grossly overconfident.

Underconfidence and error independence in sensory-discrimination tasks

Interestingly, calibration of confidences in choices that require sensory discrimination ("Which line is longer?") yields underconfidence (Juslin, Olsson, & Winman, 1998). Moreover, unlike sets of general knowledge questions that may contain items that are misleading for a majority of participants (thus leading to less than 50% correct choices), solution probabilities for sensory-discrimination tasks range between 50 and 100%. These findings bring to mind an observation Brunswik (1949) made by comparing two different versions of a size-constancy task. When presented as a perception task, the responses were compact and fairly normally distributed, with relatively few perfect answers, but also few drastic errors. In contrast, when presented as an arithmetic reasoning problem, many answers were exactly correct but there was also a substantial proportion of crude errors that revealed confusion and sometimes way-off and bizarre mistakes. Taken together, this strongly suggests that models of overconfidence must take the processes underlying the task-specific responses into account (see next section).

Theoretical accounts

The present section reviews some theories that try to identify causal mechanisms underlying the phenomenon, as well as moderators, mediators, and boundary conditions.

Heuristics and biases

During the 1970s and 1980s, when research in judgment and decision making was dominated by the *heuristics and biases program* spearheaded by Amos Tversky and Daniel Kahneman, the overconfidence phenomenon was considered one of the cornerstones that illustrate shortcomings in human information-processing capacities, thereby marking human irrationality. It has been taken as a psychological reality and explained in terms of the anchoring and adjustment heuristic (people anchor on 100% confidence and adjust insufficiently downward, or they anchor on their best guess when making numerical estimates which will, subsequently, result in confidence intervals that are too narrow; Block & Harper, 1991; see also Chapter 12), or in terms of a confirmation bias (for a critical discussion of this explanation see Gigerenzer et al., 1991; see also Chapter 5). However, this program cannot account for most of the findings summarized in the last section.

A signal-detection approach

Ferrell and McGoey (1980) proposed a computational approach based on signal-detection theory, which allowed for fitting calibration curves. Beliefs in the truth of true statements and beliefs in the truth of false statements were conceptualized as overlapping probability distributions along a continuous decision variable. By partitioning this continuum into segments, confidences were seen as a function of the area from the distribution of true statements and that of false statements that fell into the particular segment. Although quite specific on the level of the response function, this model is mute about the cognitive processes that are responsible for mapping a particular statement to a particular value on the decision variable. Moreover, given the many free parameters of the model ($n-1$ cut-off points for n segments), it is quite flexible and thus difficult to falsify (for a critical discussion of this model, see Juslin et al., 1998).

Ecological models

Independently, Gigerenzer et al. (1991) with their theory of probabilistic mental models (PMMs) and Juslin (1994) developed what was later termed *ecological models* (McClelland & Bolger, 1994). When solving a task such as "Which city has more inhabitants, A or B?" people construct a PMM (unless they have direct knowledge or can deduce the answer with certainty, which Gigerenzer et al. called a "local mental model"). By searching for probabilistic cues that discriminate between the two alternatives, the question is put into a larger context. For example, imagine that a search, prompted by a city-population comparison, hits on the soccer-team cue: City A has a soccer team in the major league and City B does not. Based on literature about automatic frequency processing, PMM theory posits that people are able to estimate the ecological validity of cues (as long as the objects belong to their natural environment, which would also explain the expertise effect mentioned above). This validity is defined by the relative frequency of cases in the environment for which the cue indicates the correct answer. If participants choose the alternative to which the cue points and report the cue validity as their confidence, they should be well calibrated. This, however, is only true if the cue validities in the item sample reflect the cue validities in the population. If researchers do not sample general-knowledge questions randomly, but over-represent items in which cue-based inferences would lead to wrong choices, overconfidence will occur. Such overconfidence does not reflect fallible reasoning processes but is an artifact of the way the experimenter sampled the stimuli and ultimately misrepresented the cue-criterion relations in the ecology.

How does PMM theory explain the confidence-frequency effect? While making inferences based on cues and stating cue validities as their confidences, people are unaware that the cue validities are eventually lower than in the corresponding reference class, that is, in the population from which the items have been drawn. The question "How many of the last 50 items did you answer correctly?" activates

another reference class, namely, one's performance in similar testing situations. And because people anticipate that general-knowledge items are typically difficult, they adjust their frequency estimates accordingly. As a consequence, if items are typical with respect to testing situations, that is, if they are selected to be difficult, people's frequency estimates are well calibrated. In contrast, if items are randomly drawn from a specified reference class and if cue validities in the sample are thus representative of those in the population, these items are, at the same time, untypical for general-knowledge questions. People's frequency estimates for these items – which are easier than expected – are hence too low, that is, they reveal underconfidence.

Error models

While Gigerenzer et al. (1991) focused on the cognitive processes and on the impact of item sampling, they did not attempt to elaborate explicitly on the impact of stochastic components of the judgment process. This was achieved in subsequent publications by Erev, Wallsten, and Budescu (1994) and Pfeiffer (1994), who demonstrated that even unbiased response error would deteriorate calibration, simply due to regression effects. To see why, assume that the overt response is a result of true confidence plus (unbiased) error. In the case of overt responses of 100%, the error can only pull in one direction, namely, downward. In fact, the representative set in Figure 16.1 shows such regression effects at the ends of the scale.

Yet, this is only half of the story. If two variables are imperfectly correlated, regression to the mean can be obtained for each of them. And indeed, Erev et al. (1994) also demonstrated that one simply has to plot mean subjective probabilities for specific events (e.g., outcomes of basketball games) against their objective probabilities to reverse the typical pattern, that is, to obtain underconfidences at the right end of the scale (events that are correctly predicted by all participants have a subjective probability below 100%) and overconfidences at the left end of the scale.

A similar explanation has been given by Moore and Cain (2007) to account for the sibling of the hard-easy effect in the ranking paradigm: better-than-average placements (i.e., overconfidence) for skill-based tasks that people believed to be easy, and worse-than-average placements (underconfidence) for tasks they believe to be difficult. These authors argued that a simple Bayesian explanation can account for these findings: "On skill-based tasks, people generally have better information about themselves than about others, so their beliefs about others' performances tend to be more regressive (thus less extreme) than their beliefs about their own performances" (p. 197). To illustrate, imagine someone believes, a priori, that the average performance in a group is, say, 70%, and also that her own performance is 70%. After having taken the test and been told that her performance is 90% (50%), she may now, a posteriori, believe that the average performance of the group is 80% (60%). These estimates are regressed towards the prior of 70%, and the contrast between own performance and estimated group performance results in better-than-average (worse-than-average) placements.

Combined error models

Whereas Erev et al. (1994) emphasized the role of error without specifying the cognitive processes (as, for instance, PMM theory did), other authors have brought these approaches together (Björkman, 1994; Juslin & Olsson, 1997; Juslin et al., 1999; Soll, 1996). Building on PMM theory, Juslin and his colleagues have called the mismatch between cue validities for a sample and for the population the "Brunswikian error", and the unsystematic response error introduced by Erev et al. the "Thurstonian error". Note that these two errors correspond to two sources of uncertainty. Brunswikian uncertainty is external and reflects less-than-perfect predictability of unknown states of the world (criterion) given known states (cues). Thurstonian uncertainty, in contrast, is internal and reflects less-than-perfect reliability of the information-processing system itself (Juslin & Olsson, 1997). By combining Brunswikian and Thurstonian errors in one single model they could, for instance, explain not only why the calibration curves look different for different sampling procedures (due to Brunswikian error) but also why they look different for the half and the full scale (due to Thurstonian error), thus accounting for the format dependence of confidence judgments (Juslin et al., 1999). For quite another account on overconfidence – one that also combined these two kinds of errors, but now in a model that assumes an instance-based memory representation – see Dougherty's (2001) application of the MINERVA-DM model.

The sensory sampling model

A special case of Thurstonian uncertainty is neural noise in an organism facing a sensory-discrimination task. If two lines have about the same length, then the impression of which one is longer will most likely vary over time. In their sensory sampling model, Juslin and Olsson (1997) suggested that people (a) decide which line is longer based on the proportion of impressions made within a given short-term memory window that speak for each of the two alternatives, and (b) state this proportion as their confidence. At first glance this reminds one of the rules suggested by PMM theory for stating choice and confidence based on uncertain cues. There are, however, two important differences. First, knowledge about cues is shared by many people and therefore their errors are dependent: If a cue is misleading, it misleads the majority of people. Stochastic fluctuations in people's sensory systems, in contrast, are independent of each other, as are their errors (see Brunswik, 1949, as mentioned above). Second, if a probabilistic cue used in general-knowledge items makes a correct prediction in 80% of the cases, and confidence is 80%, then a perfect calibration will result in the long run. In a sensory-discrimination task, however, the sensory sampling model predicts underconfidence.

To see why, imagine a participant has to choose which of two lines is longer. While comparing these lines, she will experience a series of sensory impressions, some of which suggest that the "red line is longer" and some that the "blue line is longer." These impressions can be thought of as balls, randomly sampled from an urn with, say, 60% red and 40% blue. According to the sensory sampling model, the choice is made in favor of the line that has the higher proportion of supporting

impressions, and the confidence is directly determined by this proportion. For instance, if 60% of the impressions for a given comparison favor the red line, then the red line is chosen and 60% is given as confidence. The proportion of correct choices is determined across comparisons. The model predicts underconfidence because the average confidence is expected to be lower than the proportion of correct choices. The expected confidence is, in our example, 60% (for each comparison and hence also averaged across comparisons). In contrast, it can be expected that more than 60% of the comparisons have a majority of red balls (for instance, if each sample of impressions for a given comparison is huge, say, 500, almost 100% of the samples will have more red balls than blue balls). Statistically speaking, choices capitalize on the law of large numbers, but confidences do not.

Rational beliefs based on imperfect measurement

I now turn to explanations that emerged from the literature on overplacement. Similarly to what we have seen for the calibration paradigm, researchers using this paradigm seemed to agree for a long time that overconfidence is real and widespread. Challenging this view, Benoît and Dubra (2011) argued and proved that a Bayesian who is uncertain about her true ability and who uses probabilistic information to infer this true ability may appear to be overconfident even though there is nothing wrong or irrational about her reasoning or calculations. To illustrate (I simplify the authors' example), let safe driving be defined by the number of caused accidents, assume there are only two types of drivers, skilled and less-skilled, and assume both groups are equally large. Moreover, the probability of causing an accident in a given year is, by definition, lower for the skilled than for the less-skilled drivers. Should someone who has caused no accident during the last three years believe that she belongs to the group of skilled or less-skilled drivers? Benoît and Dubra show that it is possible that p(skilled | no accident in 3 past years) > p(unskilled | no accident in 3 past years), while, at the same time, the number of people who have caused no accident within these years exceeds the number of skilled drivers. The base rates of accidents can be so low that some of the unskilled drivers may have been lucky and caused no accident, despite of their elevated propensity to do so. As a result, all drivers with no accident – including some of those who are in fact less-skilled – will place themselves, as Bayesians, as being better than average and also in the top 50%. But "rather than being overconfident, which implies some error in judgment, the drivers are simply using information available to them in the best possible manner" (Benoît & Dubra, 2011, p. 1592f). The authors conclude that "the simple truism that most people cannot be better than the median does not imply that most people cannot rationally rate themselves above the median" (p. 1592).

Subjective multidimensional assessment

I once took a guided tour through the main building of the University of Mannheim, which was built as a castle during the Baroque Era. The guide proudly explained that it is the biggest castle in Europe – and added after a while "in terms of total

window size." And he explained that other castles claim to be the biggest for other reasons. Lesson: If being biggest is not clearly defined, everyone can be the biggest.

Santos-Pinto and Sobel (2005) built a formal model based on this insight and demonstrated that ambiguity about the set of indicators and their relative weights can easily lead to a better-than-average effect. To illustrate, consider a group of three scientists. The first is convinced that the number of peer-reviewed articles contributes more than anything else to success and reputation, the second is convinced that the number of citations is the most important contributor, and the third believes that books count most. Not surprisingly, they will invest in developing different skills to maximize "their" dimensions (here, writing many articles, good articles, and books, respectively). Further assume that, due to their different skill investments, each will outperform the others on his or her dimension. As a result, when applying their own (idiosyncratic) weighting scheme to combine the various sub-dimensions of success, everybody will find out that his or her overall score will be above average. Similarly, it can well be that each of the two authors, Santos-Pinto and Sobel, have good reasons to believe that he contributed 90% to their paper (see above). For instance, one had 90% of the ideas and the other did 90% of the writing. Note, by the way, that the reversed causality is able to offer a plausible account for the better-than-average effect as well: Not only what we consider to be important may determine which skills we train, but, conversely, our skills and achievements may influence, in a self-serving manner, what we deem to be important. One way or another, if people differ with respect to how they think a multidimensional criterion should be measured, this variance can easily account for the better-than-average effect.

The functions of overconfidence

Overconfidence may have an instrumental value and people may display it because it has some desired consequences. Note that this perspective is quite distinct from the explanations discussed so far that treated overconfidence as a dependent variable – in the present section it is seen as an independent variable and the focus is on its effects. So what is overconfidence good for? I will discuss three potential functions: Overconfidence might have a consumption value, a motivation value, and a signaling value (Bénabou & Tirole, 2002).

Consumption: Feeling good

Overconfidence may have the same immediate effect as eating chocolate, or listening to our favorite music: We enjoy these activities. Likewise, we appreciate receiving positive feedback, praise, and approval from others. We aim to be good and competent, and we like it when others tell us that we are – and to some extent we can replace those others and tell ourselves how great we are. As Bénabou and Tirole (2002) put it, "people may just derive utility from thinking well of themselves, and conversely find a poor self-image painful" (p. 872).

Those of us who enjoy eating chocolate or potato chips when watching TV know that this may develop its own dynamics. Likewise, there is always the temptation to go

too far on the slippery slope from self-serving perceptions to self-serving biases. Such self-serving biases (Greenwald, 1980) often go along with egocentrism (Kruger, 1999) and with blind-spots when it comes to monitoring reality. Over time, people may get immune to feedback and it becomes harder and harder for them to notice any mismatch between beliefs and reality. Not only eating too much chocolate can have severe disadvantages; likewise, cheating ourselves and cultivating unjustified beliefs can backfire (see the section on applied settings below).

Motivation: Moving ahead

Unrealistic beliefs can be dangerous and lead to wrong decisions, but we often know only with the benefit of hindsight (see Chapter 23) whether a certain decision or activity was, overall, good or bad. But in order to find out, one has to move ahead – which can sometimes be advantageous, even if it was overconfidence that made us move. Whether justified or not, high confidence in his abilities and efficacy can help the individual undertake more ambitious goals and persist in the face of adversity. As an example, consider young children's belief that they can master even the most difficult tasks. As long as they grow up in an environment in which they are protected from harming themselves as a result of such overconfidence, this lack of meta-cognitive abilities is likely to increase the chances that they will attempt such tasks, thereby gaining experience and acquiring skills. Due to this self-fulfilling prophecy (even if only partial) they get an advantage over their peers who have a more realistic – and more pessimistic – view of their own competence (Bjorklund, 1997).

Closely related to the adaptive function of children's overconfidence is the planning fallacy. Had I been realistic about my time schedule and my time management abilities, I would not have committed myself to contribute this chapter. It was thus the overconfidence phenomenon itself that gave me the opportunity to write about it.

Signaling: Convincing others

Finally, believing – justified or not – that one can achieve some goals or possesses certain features helps to convince others of it. My own overconfidence did not only lead to committing myself, it also led Rüdiger to accept me as an author – and he might not have done so had I been realistic about my ability to meet the deadline. To the extent that having a chapter in this book is an honor (and I hasten to add that it is), and to the extent that potential authors compete for this honor, being overconfident can make all the difference. As Klayman, Soll, Gonzalez-Vallejo, and Barlas (1999) put it: "In a world in which competence is hard to measure, confidence often wins the day" (p. 243) (for evidence suggesting that signaling can better account for overconfidence than rational beliefs or consumption, see Burks, Carpenter, Goette, & Rustichini, 2013).

Convincing others is obviously easier if we are convinced of ourselves in the first place – and possibly even more than would be justified (Trivers, 2011). Consider a

physician who is overconfident that a particular treatment will benefit her patient. Showing high confidence that it will help may be essential for a placebo effect to occur. If the objective chances that the treatment will help are, a priori, 30%, and if they increase, objectively, to 60% a posteriori (i.e., after the physician expressed a very high confidence of, say 80%) who wants to blame her for being overconfident? After all, it helped her to be convincing which, ultimately, helped the patient. Interestingly, this self-fulfilling prophecy even reduced her overconfidence – simply by changing the reality against which the beliefs are measured.

What is the lesson to be learned from most children's overconfidence, some authors' overconfidence, and our fictitious physician's overconfidence? Unwarranted optimism about future developments may function to positively affect those developments.

Relevance in applied settings

As the last section suggested, overconfidence may have some advantages: We enjoy feeling confident, it makes us move, and we are more convincing in competitive situations. On the other hand, nurturing one's own fictions, moving into wrong directions, and engaging in fights one should better avoid can be very painful. Overconfidence is a double-edged sword. Moreover, it is often not easy to reveal as such, in particular with small samples. Studying overconfidence in the wild has its own challenges. In this section I will, very briefly, discuss three settings in the real world in which overconfidence and its effects are often investigated.

Finance and trading

Many studies in behavioral finance relate overconfidence to excessive trading. For instance, Barber and Odean (2000) analyzed the financial performance of 66,465 households who hold an account at a large discount broker and found that those who trade the most fare much worse than the market, and also much worse than the average household. They suggest that overconfidence is responsible for high trading level and the resulting poor performance. Likewise, Glaser and Weber (2007) found that online broker investors who think their investments skills are above average trade more, and Biais, Hilton, Mazurier, and Pouget (2005) found that miscalibrated traders perform worse. Relatedly, overconfidence has been discussed as a driving force in the formation of bubbles in financial markets (Scheinkman & Xiong, 2003). For a detailed overview of studies that focus on overconfidence in finance and trading, see Baker and Wurgler (2011).

Entrepreneurship and market entry

Overconfidence has often been proposed as an explanation of why individuals enter competitive markets. In a seminal study, Camerer and Lovallo (1999) ranked participants in two different ways, randomly and according to their performance

(determined either by logic puzzles or trivia questions). In 12 rounds, groups of 12–16 students were first informed about the capacity *c* of the market (*c* ranged between 2 and 8) and which type of ranking will be used in a given round to determine payoffs. Subsequently, participants had to decide – simultaneously – whether or not to enter the market. Among all who decided to enter, the *c* top-ranked participants shared $50, with higher earning for better ranks, and all others lost $10. The authors found that significantly more participants entered the market if the payoffs were based on skills (rather than on random rankings). Moreover, participants entered the competitive, skill-based market excessively even though they correctly predicted that more than *c* would enter. Why did they still dare to do so? They believed that such excessive entry was not a problem for them – but only for the others who were, unlike them, overconfident in their skills.

Åstebro, Herz, Nanda, and Weber (2014) provided an excellent overview of studies – most of them with real entrepreneurs – that focus on the role of overconfidence in entrepreneurship and market entry decisions. These authors structured their overview according to the distinction between the three different types of measuring overconfidence introduced above. They concluded that overestimation, overplacement, and optimism "all lead to positively biased perceptions of expected returns and hence should foster entrepreneurial entry", whereas "the effects of overprecision are less clear" (p. 60).

Management and CEOs

Both folk theory and scientific evidence suggest that the incidence of overconfidence is likely to be greater among top executives. One of the possible explanations is a sample selection bias: Goel and Thakor (2008) showed that overconfident individuals are more likely to win the competition within the firm and are more likely to be promoted to CEO. This preselection effect will likely be amplified in their career later on: Malmendier and Tate (2005) argued that "Individuals are the most optimistic about outcomes which they believe are under their control . . . And individuals are more prone to overestimate outcomes to which they are highly committed (Weinstein, 1980). Top corporate managers are likely to satisfy both of these pre-conditions" (p. 1736).

Some studies used one or several of the measures introduced above; others derived more task-specific measures. Ben-David, Graham, and Harvey (2013) combined both and found that Chief Financial Executives (CFOs) showed overconfidence in terms of overprecision: In a sample of more than 13,300 expected stock market return probability distributions, realized market returns were within the CFOs' 80% confidence intervals only 36.6% of the time. These authors also found "that CFO overprecision appears to be related to corporate decision making". For instance, "the overprecision measure, based on predicting the S&P 500 returns, is significantly correlated with overprecision in own-firm investment return predictions" and there is modest evidence that, on average, "firms with miscalibrated or optimistic executives invest more and have more debt" (p. 1577).

Final remarks

The field of overconfidence has seen major changes and developments. Today we can recognize several precise and highly formalized models of the cognitive processes involved in forming beliefs and generating confidences, indications of the theoretical progress that has been made in the recent past. These models have focused on the psychological mechanisms underlying choices, confidence judgments, beliefs, and eventually also overconfidence. Overconfidence is and will always remain a fascinating research topic as it confronts our subjective beliefs with the objective reality. Given the amount of interesting empirical findings, the stimulating theoretical accounts, and the ubiquitous practical relevance that we have seen in the past and continue to see in the present, I do not think I am overconfident when I predict, with a high confidence, more interesting findings and ideas to be published in the near and distant future.

Summary

- Overconfidence can mean (1) miscalibration, be it in the form of too-low percentages of correctly answered items for a given confidence category (calibration-in-the-small) or mean confidences exceeding mean percentages of correct answers across all items (calibration-in-the-large), (2) too-narrow confidence intervals around a numerical estimate, and (3) overplacement when ranking own performance, contribution, or skill relative to others.
- Several effects are well established, including the overconfidence effect, the hard-easy effect, the importance of sampling, the base-rate effect, the confidence-frequency effect, the expertise effect, and format effects.
- Ecological models have successfully explained overconfidence (and some other effects) as a consequence of item-selection procedures that distort cue validities. In this view, overconfidence is an artifact created by the experimenter.
- Error models have successfully attributed miscalibration to statistical regression of otherwise unbiased information processing.
- Several models are available that combine the strengths of the ecological and the error approaches.
- Overconfidence may have beneficial functions: a consumption value, a motivation value, and a signaling value.
- Overconfidence is a double-edged sword. There are constellations in which the benefits of being overconfident outweigh the costs, whereas in others it may lead to disaster, such as lost court trials, ruinous investments, or fatal car accidents.

Further reading

For review articles about the overconfidence phenomenon see Lichtenstein et al. (1982), O'Connor (1989), Yates (1990), Keren (1991), and Alba and Hutchinson (2000). More specific reviews and discussions of overconfidence focus on

computational models (Juslin & Olsson, 1999; McClelland & Bolger, 1994), on judgmental forecasting (Lawrence, Goodwin, O'Connor, & Önkal, 2006), on entrepreneurship and market entry (Åstebro, et al., 2014), on behavioral finance, trading, and investment (Baker & Wurgler, 2011), and on management (Malmendier & Tate, 2005).

References

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgments. *Psychological Review*, *68*, 33–45.

Alba, J. W., & Hutchinson, J. W. (2000). Knowledge calibration: What consumers know and what they think they know. *Journal of Consumer Research*, *27*, 123–156.

Allwood, C. M., & Montgomery, H. (1987). Response selection strategies and realism of confidence judgments. *Organizational Behavior and Human Decision Processes*, *39*, 365–383.

Åstebro, T., Herz, H., Nanda, R., & Weber, R. A. (2014). Seeking the roots of entrepreneurship: Insights from behavioral economics. *The Journal of Economic Perspectives, 28*(3), 49–69.

Baker, M., & Wurgler, J. (2011). Behavioral corporate finance: An updated survey. *Handbook of the Economics of Finance, 2*(A), 357–424.

Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance*, *LV*(2), 773–806.

Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *Quarterly Journal of Economics*, *117*(3), 871–915.

Ben-David, I., Graham, J. R., & Harvey, C. R. (2013). Managerial miscalibration. *The Quarterly Journal of Economics*, *128*(4), 1547–1584.

Benoît, J. P., & Dubra, J. (2011). Apparent overconfidence. *Econometrica*, *79*(5), 1591–1625.

Biais, B., Hilton, D., Mazurier, K., & Pouget, S. (2005). Judgmental overconfidence, self-monitoring, and trading performance in an experimental financial market. *The Review of Economic Studies*, *72*(2), 287–312.

Bjorklund, D. F. (1997). The role of immaturity in human development. *Psychological Bulletin, 122,* 153–169.

Björkman, M. (1994). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational Behavior and Human Decision Processes*, *58,* 386–405.

Blavatskyy, P. R. (2009). Betting on own knowledge: Experimental test of overconfidence. *Journal of Risk and Uncertainty, 38*(1), 39–49.

Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes, 49*(2), 188–207.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78,* 1–3.

Brunswik, E. (1949). Discussion: Remarks on functionalism in perception. *Journal of Personality*, *18*(1), 56–65.

Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the 'planning fallacy': Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, *67,* 366–381.

Burks, S., Carpenter, J., Goette, L., & Rustichini, A. (2013). Overconfidence and social signaling. *Review of Economic Studies, 80*(3), 949–983.

Camerer, C., & Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *American Economic Review*, *89*(1), 306–318.

Cesarini, D., Sandewall, Ö., & Johannesson, M. (2006). Confidence interval estimation tasks and the economics of overconfidence. *Journal of Economic Behavior & Organization*, *61*(3), 453–470.

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General*, *130,* 579–599.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*, 519–527.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance*, *26,* 32–53.

Finkelstein, M. O., & Levin, B. (2001). *Statistics for lawyers.* New York: Springer Science & Business Media.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3,* 552–564.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98,* 506–528.

Glaser, M., & Weber, M. (2007). Overconfidence and trading volume. *The Geneva Risk and Insurance Review*, *32*(1), 1–36.

Goel, A. M., & Thakor, A. V. (2008). Overconfidence, CEO selection, and corporate governance. *The Journal of Finance, 63*(6), 2737–2784.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American Psychologist*, *35*(7), 603–618.

Hoffrage, U. (2011). Recognition judgments and the performance of the recognition heuristic depend on the size of the reference class. *Judgment and Decision Making, 6*(1), 43–57.

Hoffrage, U., & Hertwig, R. (2006). Which world should be represented in representative design? In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 381–408). Cambridge, NY: Cambridge University Press.

Johnson, J. E. V., & Bruce, A. C. (2001). Calibration of subjective probability judgments in a naturalistic setting. *Organizational Behavior and Human Decision Processes*, *85,* 265–290.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, *57,* 226–246.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, *104,* 344–366.

Juslin, P., & Olsson, H. (1999). Computational models of subjective probability calibration. In P. Juslin & H. Montgomery (Eds.), *Judgment and decision making: Neo-Brunswikian and process-tracing approaches* (pp. 67–95). Mahwah, NJ: Lawrence Erlbaum.

Juslin, P., Olsson, H., & Winman, A. (1998). The calibration issue: Theoretical comments on Suantek, Bolger, & Ferrell (1996). *Organizational Behavior and Human Decision Processes*, *73,* 3–26.

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1038–1052.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, *107,* 384–396.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, 77, 217–273.

Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216–247.

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology,* 77, 221–232.

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*(3), 493–518.

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: Cambridge University Press.

Malmendier, U., & Tate, G. A. (2005). Does overconfidence affect corporate investment? CEO overconfidence measures revisited. *European Financial Management, 11*(5), 649–659.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980–94. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). Chichester, UK: Wiley.

Moore, D. A., & Cain, D. M. (2007). Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. *Organizational Behavior and Human Decision Processes*, *103*(2), 197–213.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

Murphy, A. H. (1972). Scalar and vector partitions of the probability score (Part 1): Two-state situation. *Journal of Applied Meteorology*, *12,* 595–600.

Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *National Weather Digest*, *2,* 2–9.

O'Connor, M. (1989). Models of human behavior and confidence in judgment: A review. *International Journal of Forecasting*, *5,* 159–169.

Pfeiffer, P. E. (1994). Are we overconfident in the belief that probability forecasters are overconfident? *Organizational Behavior and Human Decision Processes*, *58,* 203–213.

Russo, J. E., & Shoemaker, P. J. H. (2004). *Winning decisions.* New York: Doubleday/ Random House Inc.

Santos-Pinto, L., & Sobel, J. (2005). A model of positive self-image in subjective assessments. *American Economic Review, 95*(5), 1386–1402.

Scheinkman, J. A., & Xiong, W. (2003). Overconfidence and speculative bubbles. *Journal of Political Economy*, *111*(6), 1183–1220.

Schneider, S. (1996). Item difficulty, discrimination, and the confidence-frequency effect in a categorical judgment task. *Organizational Behavior and Human Decision Processes*, *61*, 148–167.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, *65*, 117–137.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 299–314.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, *47*(2), 143–148.

Trivers, R. (2011). *Deceit and self-deception: Fooling yourself the better to others*. Allen Lane, Penguin Books.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39*(5), 806–820.

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice-Hall.

Author's note

I would like to thank Gerd Gigerenzer, Wolfgang Hell, Peter Juslin, Rüdiger Pohl, and Luis Santos-Pinto for comments on earlier drafts, and Justin Olds and Anita Todd for their careful editing.

17

POLLYANNA PRINCIPLE

Margaret W. Matlin

About 40 years ago, I published a book with David Stang called *The Pollyanna Principle: Selectivity in Language, Memory, and Thought* (Matlin & Stang, 1978). That book was named after Pollyanna Whittier, the child heroine of a book by Eleanor Porter (1913). This book sold more than one million copies (Levine, 2007). Soon, "Pollyanna" became a fairly common label for optimists. A Pollyanna is someone who looks on the bright side of every misfortune, who remembers only the happy events, and who believes that the world and its inhabitants are truly wonderful.

In 1969, Jerry Boucher and Charles E. Osgood published a journal article called "The Pollyanna Hypothesis," which described observations about several linguistic phenomena. For example, in a variety of different languages, people use evaluatively positive words more frequently than evaluatively negative words. Stang and I argued, however, that Pollyanna tendencies are more widespread. These tendencies include additional positive biases in language, as well as in perception, memory, and judgment. Accordingly, we decided that the Pollyanna Hypothesis was so pervasive and powerful that it deserved to be upgraded to the "Pollyanna Principle."

In brief, the Pollyanna Principle states that pleasantness predominates: We typically process pleasant items more accurately and efficiently than unpleasant or neutral items, and we tend to make positive judgments about a wide variety of people, events, situations, and objects. (Even distilled water is judged to have a somewhat pleasant taste.) The present chapter emphasizes the Pollyanna Principle in two areas: memory and judgment.

It's worth noting, incidentally, that the Pollyanna Principle is related to another topic discussed in this handbook: the mere exposure effect (Chapter 14). Stang and I had previously published research on this particular phenomenon (e.g., Matlin, 1970, 1971; Stang, 1974, 1975). However, in the mere exposure effect, "frequency of use" is typically the independent variable and "positive evaluation" is typically

the dependent variable. In contrast, in the Pollyanna Principle, "positive evaluation" is the independent variable and "frequency of use" is typically the dependent variable.

Nearly four decades have passed since the publication of the book, *The Pollyanna Principle.* The world situation in the 21st century deserves the darkest possible pessimism. However, the psychological research on cognitive illusions focuses on the thought processes of individual people, who spend much more time contemplating their personal experiences than international events. The current chapter explores how the Pollyanna Principle is relevant in two cognitive domains: memory and judgment. In addition, Text box 17.1 describes a research study that focuses on judgment.

TEXT BOX 17.1 POLLYANNA RESEARCH STUDY

One attractive feature of many components of the Pollyanna Principle is that they are relatively easy to demonstrate. The research study described in this section is a variant of a study by Justin Kruger (1999), who examined a phenomenon called "the Lake Wobegon effect." This term requires a brief explanation for those residing outside North America. For several decades, public radio stations in the United States and Canada have broadcast a program called "A Prairie Home Companion." Every week, the program's host – Garrison Keillor – provides a monologue about a charming but fictional small town in rural Minnesota, which he calls "Lake Wobegon." Each monologue ends with the sentence, "And that's the news from Lake Wobegon, where all the women are strong, all the men are good looking, and all the children are above average."

In the real world, do people believe that they are genuinely above average? To test the generality of the "Lake Wobegon effect," Kruger obtained a list of 14 different skills, which had been rated on a 4-point difficulty scale by students participating in an earlier College Board survey. He then asked Cornell University undergraduates to rate themselves on each skill, compared to other students in their own classroom. The students used a scale from 1 (very unskilled) to 10 (very skilled). Indeed, these students did rate themselves as being "above average" on most of these characteristics.

Kruger also discovered that the difficulty rating of each skill was negatively correlated ($r = -0.81$) with the percentage of students who rated themselves above average on that skill. For example, 89% of the students rated themselves above average on the ability to get along with others (rated only 3.3 on the 10-point difficulty scale), whereas only 38% rated themselves above average on the mechanical scale (rated 6.3 on the difficulty scale).

To provide a wider range of items, I selected five skills from Kruger's list and added 16 additional skills, primarily based on skills listed on several

graduate-school recommendation forms. Then the 25 students in my Fall 2002 course in Cognitive Psychology evaluated their personal skill on these 21 items, using a 10-point rating scale to indicate their own "perceived difficulty." The students' mean difficulty ratings for these skills are shown in Table 17.1.

TABLE 17.1 List of skills, mean difficulty rating, mean bias score, number below average, and number above average (data obtained from my Cognitive Psychology Class, Fall 2002)

Skill	*Mean difficulty rating*	*Mean skill rating*	*Number below average*	*Number above average*
1. Ability to understand theoretical issues	5.6	6.6	6	19*
2. Ability to work independently	2.8	7.5	2	23*
3. Ability to work with a research team	4.6	6.1	9	16
4. Compassion for other people	4.6	7.7	2	23*
5. Computer skills	4.5	5.4	13	12
6. Concern about social justice	4.8	7.4	6	19*
7. Creativity	4.6	6.2	9	16
8. Critical thinking	5.2	6.4	7	18*
9. Discussion participant	5.6	5.9	14	11
10. Dependability	4.6	7.4	2	23*
11. Getting along with others	3.4	7.0	1	24*
12. Knowledge about psychology	5.3	6.4	4	21*
13. Leadership	5.9	6.7	6	19*
14. Motivation to achieve in psychology	4.9	6.8	4	21*
15. Oral communication before a group	6.5	6.5	12	13
16. Research design	6.6	5.6	12	13
17. Skill in mathematics	6.1	6.0	9	16
18. Statistics	6.4	6.3	8	17
19. Time management	6.2	6.6	8	17
20. Trustworthiness	4.1	7.8	1	24*
21. Written communication	4.3	6.8	4	21*

* The chi-square calculation for each of these items is significant, $p < 0.05$; on these items, students were significantly more likely to rate themselves "above average," rather than "below average." These items therefore demonstrate the Lake Wobegon effect.

Method

Photocopy the list of items in the Appendix onto a single sheet of paper; also include the rating scale and the instructions that appear at the top of the page. After all students have received a copy, they should be reminded, "Please be sure to compare yourself with the average student in this classroom." They should also be instructed that they must rate all 21 items on the list. When the students appear to be finished, tell them to fold their sheet in half so that the answers are not visible; then collect all the sheets.

Results

For each of the 21 skills, tally the number of students who consider themselves "above average" (ratings of 6, 7, 8, 9, and 10) and the number who consider themselves "below average" (ratings of 1, 2, 3, 4, and 5). Then the complete data set can be analyzed. First, count (A) the number of skills where the majority of students classified themselves "above average" and (B) the number of skills where the majority of students classified themselves "below average." (For students in my Cognitive Psychology class, there were 19 skills in Category A – demonstrating the Lake Wobegon effect – and 2 skills in Category B; see Table 17.1.) Next conduct a chi-square analysis – comparing the number of skills in Category A with the number in Category B. Did the class demonstrate an overall Lake Wobegon effect? In other words, did they believe that they were "above average" on a greater number of skills, compared to the number of skills where they were "below average"?

You can also determine which of those 21 skills are especially likely to generate the Lake Wobegon effect. Calculate the number of people who would need to rate themselves "above average" on a given item in order to obtain a significant chi-square at the $p < 0.05$ level of significance. (For a class of 40 students, at least 27 students would have to rate themselves "above-average" and 13 or fewer would have to rate themselves "below-average.") Make a list of the skills where the Lake Wobegon effect clearly operates, as well as a list of skills that do not demonstrate this effect. (Your class may even show a statistically significant "reverse Lake Wobegon effect" on some items.) Table 17.1 lists the data for my class. As you can see, 12 of the 21 skills show a significant Lake Wobegon effect, and none show a significant "reverse Lake Wobegon effect."

Next, ask the class members to speculate about which category of skills is most likely to generate the Lake Wobegon effect. Presumably someone will suggest that the easier tasks are especially likely. Everyone should then inspect the skill-difficulty ratings to see whether the class data seem to confirm this hypothesis. You can calculate a correlation between (1) an item's mean difficulty rating, as shown in Table 17.1 and (2) the number of students who

rated themselves above average in your classroom. In my class, the correlation was –0.53, when all 21 of the characteristics were included.

Depending on the time frame, the class members can gather ratings for these 21 skills on dimensions other than skill difficulty (e.g., social desirability or gender-congruent skills). The following overview of the Pollyanna Principle may suggest additional factors.

Overview of the Pollyanna Principle

This chapter examines, in some detail, the research about the Pollyanna Principle in two important domains, memory and judgment. In addition, Text box 17.2 illustrates the pervasiveness of this principle in two additional areas, perception and language. The Matlin and Stang (1978) book provides more specific information on each of these points, as well as other phenomena.

TEXT BOX 17.2 EXAMPLES OF OF THE POLLYANNA PRINCIPLE IN PERCEPTION AND LANGUAGE

Perception

1 Many studies show that people judge pleasant stimuli to be larger in size than unpleasant or neutral stimuli (Matlin & Stang, 1978).
2 People typically see more desirable objects as being closer to them than less desirable objects (Balcetis & Dunning, 2010).
3 Older adults generally prefer to look at positive or neutral images, rather than negative images (Allard & Isaacowitz, 2008; Reed, Chan, & Mikels, 2014).

Language

1 When people are asked to describe an individual, they tend to emphasize that person's positive characteristics, rather than her or his negative characteristics (Bergsieker, Leslie, Constantine, & Fiske, 2012).
2 Pleasant words usually have higher frequencies of occurrence, compared to less pleasant words. For instance, Dodds and his coauthors (2015) studied 100,000 words in 10 different languages, representing a variety of word sources. These included sources such as articles in English in the *New York Times*, Indonesian movie subtitles, and samples of Spanish tweets. Native speakers of each language then rated each word on a 9-point scale, from negative to positive. Altogether, the data represented

5 million individual scorings for these words. The results showed – in all 10 languages – that people use positive words significantly more often than negative words.

3 In antonym pairs, the more pleasant member typically entered the English language prior to the unpleasant member (Matlin & Stang, 1978).

4 When people construct a list of items, they usually place the pleasant items before the less pleasant items. For example, they list people they like before people they dislike. They also list their favorite television programs, actors, colors, and desserts before they list the items they dislike (Matlin & Stang, 1978).

It is important to emphasize that some of these phenomena – such as the Lake Wobegon Effect – are clearly cognitive illusions. For example, suppose that students have just completed their first examination in a class, and they are asked to judge their performance on this exam. If students who score in the bottom 20th percentile of the class estimate that they have performed at the 60th percentile, on average, then these estimations represent a cognitive illusion (Schlösser, Dunning, Johnson, & Kruger, 2013). Other phenomena – such as looking on the bright side of life events – seem to provide adaptive mechanisms for coping with tragedy, enhancing motivation, and facilitating social interactions.

Memory[1]

In general, the research on memory demonstrates selective recall. We tend to remember the delightful and forget the dismal, and we also remember praise better than condemnation. Furthermore, as the saying goes, "Time heals all wounds." As our memory for an event fades, the positive components of that event seem *slightly less* positive, whereas the negative components seem *much less* negative. As a result, the overall evaluation of that event becomes somewhat more positive.

Accuracy of recall for daily events

During the 1930s and the 1940s, psychologists published about a dozen studies that focused on people's recall of their daily experiences. In general, the research in that era supported the Pollyanna Principle; people usually recalled pleasant items somewhat more accurately than unpleasant items.

However, these early researchers also identified a second factor – the Intensity Principle – which operates independently of the Pollyanna Principle. According to the Intensity Principle, intense items often predominate over more neutral items (Skowronski, Betz, Thompson, & Shannon, 1991). For example, we are more likely to accurately recall an intensely positive stimulus, compared to a mildly positive

stimulus. We are also more likely to accurately recall an intensely negative stimulus, compared to a mildly negative stimulus. In order to avoid the confounding variable of intensity, researchers must compare positive and negative stimuli that are matched in their intensity. In other words, it would be unfair to compare (a) the unexpected death of a friend with (b) recall for receiving a high score on a quiz in a relatively easy course.

Accuracy of recall for personal information

If we recall pleasant events about our lives more accurately than unpleasant events, do we also recall pleasant information about ourselves more accurately than unpleasant information? In a representative study, Sedikides and Green (2000) told participants that they would take a computerized personality test. The 45 items on this bogus test included statements such as "I sometimes go to people I consider wise for advice" and "I don't mind visiting places where I have never been before" (p. 912). During the next phase of the study, the participants were told that they would receive highly accurate feedback about their personality, based on their responses on the test. Half of the feedback statements were positive (e.g., "X would follow through on a promise made to friends") and half were negative (e.g., "X would make fun of others because of their looks"). Later, the participants recalled a significantly greater number of positive descriptions of themselves, compared to the number of negative descriptions.

Memories for events tend to grow more positive over time

According to another component of the Pollyanna Principle, our memory for unpleasant events tends to grow substantially less negative (in other words, closer to neutral) with the passage of time. In contrast, our memory for pleasant events tends to grow only slightly less pleasant. The dependent variable in these studies is not the *number* of items recalled, but the *qualitative* nature of specific events in our lives.

This topic was originally studied by Walker, Vogl, and Thompson (1997). They asked undergraduate students to record one personal event every day for about 14 weeks and to rate both the pleasantness and the intensity of each event. After 3 months, the participants returned for a second session. A researcher read each event from the previous list, and the participants were instructed to rate the pleasantness of that event.

According to the results of this study, the events originally considered to be pleasant were now considered to be slightly less pleasant. In contrast, the events originally considered to be unpleasant were now considered to be substantially more pleasant (that is, substantially less unpleasant). Consistent with the Pollyanna Principle, the change in pleasantness ratings was significantly larger for the unpleasant items than for the pleasant items. Walker et al. (1997) called this phenomenon the "fading affect bias."

As we will see in the section on judgment, people believe that the events in their lives are generally pleasant. When combined with the fading affect bias, these two

factors support the title of an article by Walker, Skowronski, and Thompson (2003, p. 203): "Life is pleasant – and memory helps to keep it that way!" The remaining discussion of research on memory examines several recent studies on the fading affect bias, with respect to autobiographical memory (i.e., memory for experiences and information related to oneself).

Let's first consider the research about the fading affect bias, with respect to undergraduate participants. For example, Ritchie, Skowronski, Hartnett, Wells, and Walker (2009) conducted four relevant studies. In one of these studies, for example, the researchers asked students to describe one representative life event for each of four different emotions: sad, angry, calm, and happy. For each emotion, they were instructed to recall one event that had occurred about six months earlier and one that had occurred about one year ago. During this initial session, they wrote an event description for each relevant category. They also rated the intensity of their emotional reaction about this event, at the current time.

One month later, the participants re-read their own description of each event, but they did not see their earlier rating for that event. Once more, they rated the intensity of their current emotion for each of these events. According to an analysis of variance, the two kinds of unpleasant emotions showed significantly greater emotional fading than the two kinds of pleasant emotions. Several additional studies have varied the details of the memory conditions, and they also demonstrate the fading affect bias (Ritchie et al., 2009; Walker & Skowronski, 2009).

Other research has explored how rehearsal is related to the fading affect bias. Walker, Skowronski, Gibbons, Vogl, and Ritchie (2009) studied a total of 337 students at three different universities. The students were asked to recall a total of either four or six memories for positive and negative events in their lives. They also estimated the number of times that they had rehearsed each of these events. The students reported more frequent rehearsal for positive events, compared to negative events. Without rehearsal, their memory for those negative events is more likely to fade, potentially contributing to the fading affect bias.

As you might suspect, researchers have identified individual differences in the fading affect bias. For example, Walker and his coauthors found that undergraduates who are nondepressed are more likely than depressed undergraduates to show the fading affect bias (Walker & Skowronski, 2009; Walker et al., 2003).

The age-related positivity effect in memory

The previous section focused on the emotions of undergraduate students, most often in their late teens and early twenties. What happens when older individuals try to remember pleasant and unpleasant information? In general, the research shows that older adults are even more likely than younger adults to recall the emotionally positive events in their lives. This effect is called the age-related positivity effect. These findings are definitely not consistent with popular stereotypes about aging, which emphasize a "doom and gloom" perspective about the lives of elderly people (Reed & Carstensen, 2012; Reed, Chan, & Mikels, 2014).

For example, Reed and his colleagues (2014) conducted a meta-analysis of research-based studies about the age-related positivity effect. To study the positivity effect, Reed and his colleagues located published research that had compared younger and older adults, with respect to their long-term memory for positive and negative information. The results of the meta-analysis showed that older adults typically remembered more positive information than negative information. In contrast, younger adults typically remembered more negative information than positive information, a puzzling conclusion that is not consistent with the Pollyanna Principle.[2]

Why should older adults remember more positive information than negative information? One likely explanation is that older people increasingly realize that life is fragile and they have relatively little time left. Accordingly, they frequently choose to ignore negative information and seek emotional satisfaction (Carstensen, 2006; Kennedy, Mather, & Carstensen, 2004; Reed, Chan, & Mikels, 2014).

In related research, Berntsen, Rubin, and Siegler (2011) asked adults in their 60s to think of one highly positive life event, as well as one highly negative life event. They were also asked to indicate how central each event was to their life story. The results showed that people judged the highly positive event to be more central to their life story, compared to the highly negative event. They also considered the positive event to be more central to their personal identity. Note that this judgment may actually be correct. For instance, the positive event may actually be a greater distance from neutral, compared to the negative event. I will return to this topic when discussing people's judgments of their quality of life in the next section.

Judgment

Two components of the Pollyanna Principle in judgment have been especially active areas of research in recent years. This research demonstrates that people typically judge their lives to be happy. Furthermore, they tend to hold inflated opinions about their abilities, their personal attributes, and their future success.

Rating our happiness and quality of life

In *The Pollyanna Principle,* Stang and I reported a number of indications that people believe their lives are generally pleasant. For instance, people list a larger number of positive life events than negative life events (consistent with the selective recall discussed in the section on memory). People also report that they are optimists, and they report – on demographic surveys – that they are happy. In addition, they rate other people as being happy (Matlin & Stang, 1978). Let's consider three variables that could be related to happiness. These variables are age, personal wealth, and country of residence.

It is important to emphasize, however, that the research about these three categories is based on correlational data; we cannot manipulate any of these three variables. Furthermore, the data may reflect helpful coping strategies. For example, an elderly

person may reason, "My income may be relatively low, but money isn't everything, and I have many good friends."

1. Age

We saw in the previous section that older adults tend to recall more positive events than negative events (Reed et al., 2014). Löckenhoff and Carstensen (2007) conducted a related study on decision making. These researchers instructed younger participants (aged 22 to 39) and older participants (aged 62 to 93) to examine information about different physicians. Compared to the younger participants, the older participants spent a greater percentage of their time examining the positive information about the physicians.

Other studies about age and happiness are consistent with this research. Specifically, those older adults who have good social support – and who feel that they have control over their lives – typically report a higher sense of well-being than younger adults (Charles & Carstensen, 2010). Interestingly, elderly people who have strong social support actually live an average of five years longer than elderly people without strong social support (Pressman & Cohen, 2007). Again, however, it is important to emphasize that these data are correlational, rather than causal.

We might expect that elderly widows would give themselves relatively low ratings. After all, they are likely to experience poverty, poor health, and loneliness. Furthermore, elderly women are especially likely to receive ageist reactions from other people (Matlin, 2012). Nevertheless, Neill and Kahn (1999) found that elderly widows gave themselves an average rating of 19 on a life-satisfaction scale, in which the most negative score was 0 and the highest was 26. Several factors help to explain why these women might be reasonably happy. Specifically, they have learned how to cope effectively with negative emotions. They may also choose to spend time on activities they consider enjoyable. Furthermore, they have learned to maintain a positive view of themselves, even when faced with disappointments (Dark-Freudeman, 2010; Whitbourne, 2010; Whitbourne & Sneed, 2002). In general, then, the research defies the stereotype that happiness decreases as people grow older. In fact, elderly individuals are typically satisfied with their lives.

2. Personal wealth

Many people believe that money can buy happiness. However, the correlation between income level and well-being is complex. Yes, in order to be happy, people require a certain minimal income to cover the basic requirements of food, housing, and essential services. But beyond that minimum, the correlation between wealth and well-being is not strong when comparing people who live in the same country (Blanchflower & Oswald, 2004; Diener & Seligman, 2004). In fact, compared with income, strong social relationships with relatives and friends are much better predictors of happiness (Powdthavee, 2008).

3. Country of residence

Does your happiness depend on the country where you live? The residents of all but a few countries in a cross-national survey rated themselves on the positive end of the question about life satisfaction (Seligman, 2002). Here's another point: When a country has good medical care and social services, its residents are likely to be relatively happy (Easterlin, 2013). Once again, however, these data are correlational, rather than causal, so the conclusions are somewhat ambiguous.

Rating our abilities and personal characteristics

We have seen that people typically rate themselves above average on scales of well-being and happiness. What happens when people rate their own abilities and personal characteristics? According to the discussion of this topic in *The Pollyanna Principle*, people are remarkably confident about themselves. In fact, they typically rate themselves above average on a wide variety of attributes (Matlin & Stang, 1978). This tendency is certainly consistent with both the overconfidence effect and the hindsight bias (Chapters 16 and 23 in this book).

Research on this topic has blossomed in recent years. It is now known by other names, such as the "better-than-average effect" and the "optimistic bias" as well as the "Lake Wobegon effect." The Pollyanna research study by Justin Kruger (1999) is featured in Text box 17.1. This exercise is based on a classic study, which emphasizes that people consider themselves better than average on a wide variety of academic skills and personal characteristics.

Consider some additional research on the Lake Wobegon effect, which was conducted by Kruger and Dunning (1999). Their first study assessed people's ability to judge which jokes other people would regard as humorous. This ability to identify humor requires subtle knowledge about other people's tastes. Students at Cornell University were asked to rate the humor of 30 jokes, obtained from a variety of joke books. The students then evaluated themselves on their own ability to recognize humor, compared to the average student at Cornell. The students used a percentile scale ranging from 0 ("I'm at the very bottom") to 50 ("I'm exactly average") to 99 ("I'm at the very top").

Each of these jokes was also assessed by professional comedians, to provide an "expert opinion" about the humor level of every joke. Then, the researchers calculated – for every student – an objective percentile score, in terms of the correspondence between that student's ratings of the jokes and the professional comedians' ratings of the jokes.

By definition, the average of the students' percentile scores should be the 50th percentile. However, the average percentile score – which the students provided for themselves – was the 66th percentile. Now consider just the bottom 25% of the sample, whose humor judgments were *least* similar to the experts' judgments. These individuals gave themselves an average rating at the 58th percentile, rather than a rating below the 25th percentile. Therefore, on a variety of different judgment tasks, people do not know that they do not know (Dunning, 2011).

Additional studies have provided a broad perspective on the Lake Wobegon effect. Text box 17.3 illustrates some representative observations.

TEXT BOX 17.3 REPRESENTATIVE EXAMPLES OF THE LAKE WOBEGON EFFECT

1 Students enrolled in a large mid-level psychology class completed their first examination of the term (Schlösser et al., 2013). In one condition, the students judged their performance in terms of their percentile, based on the entire class. Here, the students in the bottom 20th percentile estimated that they had performed at the 60th percentile, on average. In another condition, the same students judged their performance in terms of their raw score, where a perfect score was worth 44 points. Here, these same students overestimated their scores by an average of only 4 points. Apparently, it is easier to admit receiving a low absolute score than to admit receiving a score that is substantially lower than the score of one's peers.

2 Just as people inflate their own virtues, they deflate their own risks of encountering a negative outcome. For instance, people believe that they are personally less likely than other people to experience automobile accidents, crimes, unwanted pregnancies, and a variety of psychological and physical health problems (Helweg-Larsen & Shepperd, 2001).

3 Students from several Eastern European universities were enrolled in a microeconomics course in the Czech Republic. After they completed their first examination, they estimated their actual scores and the percentage of people in the class who would perform better than themselves on this exam. The results showed the typical Lake Wobegon effect: Students overestimated their actual scores. Each student then received feedback about his or her score. The most overconfident students showed the greatest improvement in their predictions following the second examination (Ryvkin, Krajc, & Ortmann, 2012). A dose of reality may have reduced the students' overconfidence, although regression toward the mean may explain at least part of the results.

4 The Lake Wobegon effect is even more pervasive than many psychologists had imagined. Specifically, people rate other members of their group as being better than the average for that specific group, a phenomenon called the "everybody is better than their group's average." For example, suppose that people in a well-established small group (e.g., students in a psychology class) are sitting in a circle, and they are randomly assigned one other member of that group to evaluate, relative to the average group member. The results showed that they systematically rate the target person as being better than the group average (Klar, 2002;

Klar & Giladi, 1997). In fact, they even demonstrate the Lake Wobegon effect when they are asked to rate an anonymous person in a group, who is identified only by an ID number! In other words, the Lake Wobegon effect does not simply operate in an egocentric fashion. We apparently abandon our logical and mathematical skills (even though our skills are presumably well above average), and we proclaim that this anonymous person is in fact above average.

Unrealistic optimism effect

A related phenomenon is called "the unrealistic optimism effect." According to this effect, people frequently predict futures that are very unrealistically positive. Notice that this effect resembles the Lake Wobegon effect. However, the unrealistic optimism effect applies to situations that might happen in the *future*, rather than to current situations (Shepperd, Klein, Waters, & Weinstein, 2013; Shepperd, Waters, Weinstein, & Klein, 2015).

Taylor, Phan, Rivkin, and Armor (1998) conducted the classic study of the unrealistic optimism effect. This study should be especially useful for professors who assign writing projects in their courses. Professors often lament that most students are overly optimistic about the amount of time that they will require to complete a class assignment, such as a review of the literature on a particular topic. This phenomenon is also called the "planning fallacy," and Chapter 16 provides more information about this topic.

Taylor et al. (1998) studied students' behavior with respect to completing a specific class project on time. These researchers found that only 14% of students in a control condition completed their class project on time. However, students in an experimental condition were instructed to envision every step in the process of completing this project, for example, gathering the resources and organizing the project's basic structure. These students were also instructed to rehearse these steps for 5 minutes each day. In this condition, 41% completed their projects in time. Clearly, both of these groups of students displayed unrealistic optimism. However, the students in the experimental condition benefitted from envisioning the specific steps required by the project.

Shepperd et al. (2013, 2015) discussed a large number of situations in which people exhibit unrealistic optimism. In some research studies, people believe that their own personal outcome will be more positive than some objective standard. For example, college students typically believe that they will earn starting salaries that are significantly higher than *actual* starting salaries in the workplace. In other research studies, people believe that their personal outcome will be significantly more positive than the outcome experienced by their peers. For example, most people believe that they are less likely than other individuals to develop a fatal heart attack in the future.

Exceptions for the Pollyanna Principle in judgment

Despite the prevalence of the Lake Wobegon effect and the unrealistic optimism effect, some factors reduce its power. For example, as emphasized in the research study earlier in the chapter, the better-than-average effect is strong for abilities that are considered relatively easy. In contrast, people may actually rate themselves *below* average on tasks considered to be highly challenging (Kruger, 1999). Furthermore, if people know that they will be required to justify their self-evaluation to someone else – especially someone with expertise in the relevant area – the Lake Wobegon effect is much weaker (Sedikides et al., 2002).

Individual differences also play a role in the strength of these positive biases. As you might expect, people with clinical depression are relatively immune to the Lake Wobegon effect (Helweg-Larsen & Shepperd, 2001). In addition, people raised in a Japanese culture were less likely to inflate their positive attributes; presumably, other Eastern cultures might be similarly reluctant to exaggerate their self-ratings (Heine, Lehman, Markus, & Kitayama, 1999).

Furthermore, people do not always show unrealistic optimism. For example, people are less likely to demonstrate unrealistic optimism about events that they cannot control (Shepperd et al., 2015).

Explanations for the Pollyanna Principle in judgment

A variety of factors combine to produce a relatively robust tendency for people to rate themselves – and also those in their own group – as better than average. For instance, people often pay attention to their own virtues, while ignoring their own vices (Sedikides et al., 2002). In addition, they supply an inflated anchor with respect to assessing their own ability, and they do not make sufficiently large adjustments about the abilities of other people (Kruger, 1999; see also Chapter 12 in this book). Furthermore, in a study about memory, most people recalled their virtues more often than their flaws, consistent with the selective recall tendency examined earlier in this chapter (Harrison & Shaffer, 1994).

How can we explain the "everybody in my group is above the group average" phenomenon? One reasonable explanation is that people may focus on the individual group member whom they are evaluating, but they do not pay sufficient attention to the others in the group. For instance, suppose that Susan is asked to make judgments about the friendliness of Rachel, a randomly selected member of her cognitive psychology class. Susan may reason, "Well, Rachel is really friendly," but she may fail to appreciate that most members of this particular class are friendly (Klar, 2002).

The Pollyanna Principle and psychology's emphasis on the negative

Anyone who has glanced through psychology journals would certainly contend that the appropriate muse is not young Pollyanna Whittier, but grumpy Ebenezer

Scrooge from Charles Dickens's *A Christmas Carol*. For example, I recently examined the psychological research on the topic of pregnancy. Using the resource PsycINFO, I identified all English-language references published in the first three months of 2015. A total of 44 articles focused on negative topics; the four most popular negative topics that were related to pregnancy were carcinogens, health problems, sexually transmitted diseases, and postpartum depression. Eight other articles focused on emotionally neutral issues. However, not one article looked at women's excitement during a planned pregnancy, their joyous anticipation of holding a newborn baby, or any genuinely positive component of this 9-month time span of pregnancy.

Interestingly, however, a movement called "positive psychology" has become popular in recent years. Seligman and Csikszentmihalyi (2000) were among the first to introduce the term. According to the positive psychology perspective, psychologists must stop focusing on human pathology and emphasize positive features that make human lives meaningful and hopeful. The approach has been applied widely. For instance, a PsycINFO search for the term "positive psychology" – between January 2010 and May 2015 – produced 427 entries. Representative citations include these three titles: "Positive psychology in Latin America" (Castro Solano, 2014), "The state of positive psychology in higher education" (Parks, 2011), and "Extending family nursing: Concepts from positive psychology" (Skerrett, 2010).

Clearly, we need to be aware of the "dark side" of the Pollyanna Principle. This awareness can help us devise appropriate precautions, so that excessive optimism can be transformed into a realistic perspective that is sadder, but also wiser.

Summary

- According to the Pollyanna Principle, we typically process pleasant items more accurately and efficiently than other items, and we typically provide positive judgments about a wide variety of targets. This chapter focuses on memory and judgment, although it briefly describes other cognitive processes.
- The Pollyanna Principle operates in perception; for example, people judge pleasant stimuli to be larger than unpleasant or neutral stimuli.
- This principle is also illustrated in a variety of linguistic phenomena; for instance, in a variety of languages, pleasant words have higher frequencies than less pleasant words.
- In memory, people typically recall pleasant information more accurately than similarly intense unpleasant information. Also, unpleasant memories tend to grow much more neutral with the passage of time, compared to pleasant memories; this effect is called "the fading affect bias." This tendency is especially strong in older adults, a phenomenon called the "age-related positivity effect."
- People typically judge themselves to be happy; older people tend to judge themselves to be relatively happy, compared to younger people. However, personal happiness is not strongly influenced by two other demographic variables, wealth and country of residence.

- People tend to rate themselves above average with respect to personal abilities and characteristics; this optimistic-bias tendency is often called the "Lake Wobegon effect."
- People also rate a randomly specified group member as being above average for that group.
- People also tend to predict futures for themselves that are overly positive, a phenomenon called the unrealistic optimism effect.
- Psychology's emphasis on the negative has encouraged some psychologists to adopt a perspective that emphasizes positive psychology.

Further reading

To trace the development of the Pollyanna Principle, read Boucher and Osgood's (1969) article about Pollyanna tendencies in language. Matlin and Stang's (1978) book *The Pollyanna Principle* provides an overview of the phenomena in a variety of cognitive and social areas. An article by Ritchie et al. (2009) is an excellent example of the Pollyanna Principle in memory. For illustrations of the Pollyanna Principle in judgment, consult the research by Kruger and Dunning (1999), by Dunning (2011), and by Shepperd et al. (2013).

Notes

1 Very few journal articles examine the effects of pleasantness on working memory (also called "short-term memory"). Two exceptions are articles by Majerus and D'Argembeau (2011) and by Mikels, Larkin, Reuter-Lorenz, and Carstensen (2005). Research on recognition memory is also uncommon. Two exceptions are articles by Eschrich, Münte, and Altenmüller (2008) and by Stalinski and Schellenberg (2013).

2 I inspected five recent, large-sample articles cited in the study by Reed and his colleagues (2014). Each of these articles had shown greater recall for the negative items than for the positive items. This inspection revealed a wide variety of methodologies. In fact, only two of the research methods specifically focused on the straightforward recall of life events.

References

Allard, E. S., & Isaacowitz, D. M. (2008). Are preferences in emotional processing affected by distraction? Examining the age-related positivity effect in visual fixation within a dual-task paradigm. *Aging, Neuropsychology, and Cognition, 15*, 725–743.

Balcetis, E., & Dunning, D. (2010). Wishful seeing: More desired objects are seen as closer. *Psychological Science*, *21*, 147–152.

Bergsicker, H. B., Leslie, L. M., Constantine, V. S., & Fiske, S. T. (2012). Stereotyping by omission: Eliminate the negative, accentuate the positive. *Journal of Personality and Social Psychology, 102*, 1214–1238.

Berntsen, D., Rubin, D. C., & Siegler, I. C. (2011). Two versions of life: Emotionally negative and positive life events have different roles in the organization of life story and identity. *Emotion, 11*, 1190–1201.

Blanchflower, D. G., & Oswald, A. J. (2004). Well-being over time in Britain and the USA. *Journal of Public Economics, 88*, 1359–1386.

Boucher, J., & Osgood, C. E. (1969). The Pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior*, *8*, 1–8.

Carstensen, L. L. (2006). The influence of a sense of time on human development. *Science, 312,* 1913–1915.

Castro Solano, A. (2014). *Positive psychology in Latin America.* New York: Springer Science and Business Media.

Charles, S. T., & Carstensen, L. L. (2010). Social and emotional aging. *Annual Review of Psychology, 61,* 383–409.

Dark-Freudeman, A. (2010). Successful aging. In J. C. Cavenaugh, C. K. Cavenaugh, J. M. Berry, & R. L. West (Eds.), *Aging in America* (Vol. 1, pp. 255–259). New York: Praeger.

Diener, E., & Seligman, M. E. P. (2004). Beyond money: Toward an economy of well-being. *Psychological Science in the Public Interest, 5*, 1–31.

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., . . . Danforth, C. M. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, *112*, 2389–2394.

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 44, pp. 247–296). Amsterdam: Elsevier.

Easterlin, R. A. (2013). Happiness, growth, and public policy: WEAI 2012 presidential address. *Economic Inquiry, 51*, 1–15.

Esrich, S., Münte, T. F., & Altenmüller, E. O. (2008). Unforgettable film music: The role of emotion in episodic long-term memory for music. *BMC: Neuroscience*, *9*, 48.

Harrison, D. A., & Shaffer, M. A. (1994). Comparative examinations of self-reports and perceived absenteeism norms: Wading through Lake Wobegon. *Journal of Applied Psychology*, *79*, 240–251.

Heine, S. J., Lehman, D. R., Markus, H. R., & Kitayama, S. (1999). Is there a universal need for positive regard? *Psychological Review, 106*, 766–794.

Helweg-Larsen, M., & Shepperd, J. A. (2001). Do moderators of the optimistic bias affect personal or target risk estimates? A review of the literature. *Personality and Social Psychology Review, 5*, 74–95.

Kennedy, K. M., Mather, M., & Carstensen, L. L. (2004). The role of motivation in the age-related positivity effect in autobiographical memory. *Psychological Science, 15*, 208–214.

Klar, Y. (2002). Way beyond compare: Nonselective superiority and inferiority biases in judging randomly assigned group members relative to their peers. *Journal of Experimental Social Psychology, 38*, 331–351.

Klar, Y., & Giladi, E. E. (1997). No one in my group can be below the group's average: A robust positivity bias in favor of anonymous peers. *Journal of Personality and Social Psychology, 73*, 885–901.

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability adjustments. *Journal of Personality and Social Psychology,* 77, 221–232.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology,* 77, 1121–1134.

Levine, M. (2007). Pollyanna and the glad game: A potential contribution to positive psychology. *The Journal of Positive Psychology, 2*, 219–227.

Löckenhoff, C. E., & Carstensen, L. L. (2007). Aging, emotion, and health-related decision strategies: Motivational manipulations can reduce age differences. *Psychology and Aging, 22,* 134–146.

Majerus, S., & D'Argembeau, A. (2011). Verbal short-term memory reflects the organization of long-term memory: Further evidence from short-term memory for emotional words. *Journal of Memory and Language, 64*, 181–197.

Matlin, M. W. (1970). Response competition as a mediating factor in the frequency-affect relationship. *Journal of Personality and Social Psychology, 16*, 536–552.

Matlin, M. W. (1971). Response competition, recognition and affect. *Journal of Personality and Social Psychology, 19*, 295–300.

Matlin, M. W. (2012). *Psychology of women* (7th ed.). Belmont: Wadsworth/Cengage.

Matlin, M. W., & Stang, D. J. (1978). *The pollyanna principle: Selectivity in language, memory, and thought.* Cambridge, MA: Schenkman.

Mikels, J. A., Larkin, G. R., Reuter-Lorenz, P. A., & Carstensen, L. J. (2005). Divergent trajectories in the aging mind: Changes in working memory for affective versus visual information with age. *Psychology and Aging, 20*, 542–553.

Neill, C. M., & Kahn, A. S. (1999). The role of personal spirituality and religious social activity on the life satisfaction of older widowed women. *Sex Roles, 40*, 319–329.

Parks, A. C. (2011). The state of positive psychology in higher education. *The Journal of Positive Psychology, 5*, 429–431.

Porter, E. H. (1913). *Pollyanna.* Boston: The Page Company.

Powdthavee, N. (2008). Putting a price tag on friends, relatives, and neighbours: Using surveys of life satisfaction to value social relationships. *The Journal of Socioeconomics, 37*, 1459–1480.

Pressman, S. D., & Cohen, S. (2007). Use of social words in autobiographies and longevity. *Psychosomatic Medicine, 69*, 262–269.

Reed, A. E., & Carstensen, L. L. (2012). The theory behind the age-related positivity effect. *Frontiers in Psychology, 3*, 1–9.

Reed, A. E., Chan, L., & Mikels, J. A. (2014). Meta-analysis of the age-related positivity effect: Age differences in preferences for positive over negative information. *Psychology and Aging, 20*, 1–15.

Ritchie, T., Skowronski, J. J., Hartnett, J., Wells, B., & Walker, W. R. (2009). The fading affect bias in the context of emotion activation level, mood, and personal theories of emotion change. *Memory, 17,* 428–444.

Ryvkin, D., Krajc, M., & Ortmann, A. (2012). Are the unskilled doomed to remain unaware? *Journal of Economic Psychology, 33*, 1012–1031.

Schlösser, T., Dunning, D., Johnson, K. L., & Kruger, J. (2013). How unaware are the unskilled? Empirical tests of the "signal extraction" counter explanation for the Dunning-Kruger effect in self-evaluation of performance. *Journal of Economic Psychology, 39,* 85–100.

Sedikides, C., & Green, J. D. (2000). On the self-protective nature of inconsistency-negativity management: Using the person memory paradigm to examine self-relevant memory. *Journal of Personality and Social Psychology, 79*, 906–922.

Sedikides, C., Herbst, K. C., Hardin, D. P., & Dardis, G. J. (2002). Accountability as a deterrent to self-enhancement: The search for mechanisms. *Journal of Personality and Social Psychology, 83*, 592–605.

Seligman, M. E. P. (2002). *Authentic happiness: Using the new positive psychology to realize your potential for lasting fulfillment.* New York: Free Press.

Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist, 55*, 5–14.

Shepperd, J. A., Klein, W. M. P., Waters, E. A., & Weinstein, N. D. (2013). Taking stock of unrealistic optimism. *Perspectives on Psychological Science, 8,* 395–411.

Shepperd, J. A., Waters, E. A., Weinstein, N. D., & Klein, W. M. P. (2015). A primer on unrealistic optimism. *Current Directions in Psychological Sciences*, *24*, 232–237.

Skerrett, K. (2010). Extending family nursing: Concepts from positive psychology. *Journal of Family Nursing, 16,* 487–502.

Skowronski, J. J., Betz, A. L., Thompson, C. P., & Shannon, L. (1991). Social memory in everyday life: Recall of self-events and other-events. *Journal of Personality and Social Psychology, 60*, 831–843.

Stalinski, S. M., & Schellenberg, G. (2013). Listeners remember music they like. *Journal of Experimental Psychology, 39,* 700–716.

Stang, D. J. (1974). Methodological factors in mere exposure research. *Psychological Bulletin, 81,* 1014–1025.

Stang, D. J. (1975). Effects of mere exposure on learning and affect. *Journal of Personality and Social Psychology, 31*, 7–12.

Taylor, S. E., Phan, L. B., Rivkin, I. D., & Armor, D. A. (1998). Harnessing the imagination: Mental simulation, self-regulation, and coping. *American Psychologist, 53*, 429–439.

Walker, W. R., & Skowronski, J. J. (2009). The fading affect bias: But what the hell is it for? *Applied Cognitive Psychology, 23*, 1122–1136.

Walker, W. R., Skowronski, J. J., Gibbons, J. A., Vogl, R. J., & Ritchie, T. D. (2009). Why people rehearse their memories: Frequency of use and relations to the intensity of emotions associated with autobiographical memories. *Memory, 17,* 760–773.

Walker, W. R., Skowronski, J. J., & Thompson, C. P. (2003). Life is pleasant – and memory helps to keep it that way! *Review of General Psychology,* 7, 203–210.

Walker, W. R., Vogl, R. J., & Thompson, C. P. (1997). Autobiographical memory: Unpleasantness fades faster than pleasantness over time. *Applied Cognitive Psychology, 11,* 399–413.

Whitbourne, S. K. (2010). *The search for fulfillment*. New York: Ballentine.

Whitbourne, S. K., & Sneed, J. R. (2002). The paradox of well-being, identity processes, and stereotype threat: Ageism and its potential relationship to the self and later life. In T. D. Nelson (Ed.), *Stereotyping and prejudice against older persons* (pp. 247–273). Cambridge, MA: MIT Press.

APPENDIX

Instructions, rating scale, and list of skills to be used in the classroom study

Instructions to students: Below is a list of skills that might be relevant for a college student. Using the rating scale below, rate your skill on each item *in comparison to the average student in this classroom*. Place the rating in front of each item. When you are finished, fold each sheet in half and wait until everyone is done. (Do not write your name on this sheet.)

1	2	3	4	5	6	7	8	9	10
Very unskilled				Average				Very skilled	

Rating	*Skills*
_____	1. Ability to understand theoretical issues
_____	2. Ability to work independently
_____	3. Ability to work with a research team
_____	4. Compassion for other people
_____	5. Computer skills
_____	6. Concern about social justice
_____	7. Creativity
_____	8. Critical thinking
_____	9. Discussion participant
_____	10. Dependability
_____	11. Getting along with others
_____	12. Knowledge about psychology

Rating	*Skills*
_____	13. Leadership
_____	14. Motivation to achieve in psychology
_____	15. Oral communication before a group
_____	16. Research design
_____	17. Skill in mathematics
_____	18. Statistics
_____	19. Time management
_____	20. Trustworthiness
_____	21. Written communication

This page intentionally left blank

PART III

Memory

This page intentionally left blank

18

REVELATION EFFECT

André Aßfalg

Suppose you look through a magazine and encounter a full-page advertisement printed, remarkably, upside down relative to the rest of the magazine. This unusual print would force you to turn the magazine on its head in order to reveal what the advertisement is about. What might seem like an awkward mistake during the magazine's printing or design process could instead be an effective marketing strategy. At least in controlled laboratory experiments, upside-down stimuli that have to be identified first appear more familiar compared to normal upright stimuli (Watkins & Peynircioglu, 1990). Similarly, research participants rate brand names as more appealing when the brand name has to be reconstructed from an anagram compared to an ordinary, intact brand name (Kronlund & Bernstein, 2006).

The revelation-effect phenomenon

The effect that revealing information has on subsequent judgments, the *revelation effect*, is best documented for recognition judgments (Watkins & Peynircioglu, 1990). In a typical recognition experiment, participants first study a list of words. Later, in a recognition test, participants judge whether the test words in a mixed list of "old" (studied) and "new" (unstudied) words are old. Critically, a brief task directly precedes half of the "old" and half of the "new" words. When Watkins and Peynircioglu (1990) discovered the revelation effect, the preceding tasks that elicited the effect quite literally involved revealing a word. For example, in some experiments, the words appeared as fragments at first (e.g., _ _r_l_ _ _), and were then revealed letter by letter (a_r_l_n_, air_lan_, airplane). The participants had to identify the word based on the currently available fragment as soon as possible. After this letter-by-letter revelation, participants were more likely to claim that "old" and "new" words were old compared to the control condition without revelation. Later research identified tasks that also elicit a revelation effect without revealing the test

word, such as solving a simple addition problem (e.g., 245 + 318 = ?; Niewiadomski & Hockley, 2001). Put in general terms, a typical revelation-effect experiment includes a *task condition* in which the judgment directly follows a preceding task and a *no-task condition* in which the judgment occurs without a preceding task. Typically, these experiments use a within-subject design in which all participants provide several judgments in both conditions. However, the revelation effect also occurs in a between-subjects design in which one group of participants receives only task trials and another group receives only no-task trials: Participants in the task condition provide more "old" responses than participants in the no-task condition (Westerman & Greene, 1996).

Judgments

Table 18.1 lists judgments for which the revelation effect has been documented. Most research has focused on recognition judgments, perhaps because early attempts to elicit a revelation effect for other judgments failed. For example, Watkins and Peynircioglu (1990) tested the effect of a preceding task on judgments of word frequency in everyday language or typicality judgments for category exemplars (e.g., apple – fruit, dog – animal). The revelation effect was absent in both judgments. This led to the belief that the revelation effect is a genuine memory phenomenon (e.g., LeCompte, 1995). Indeed, various memory judgments are susceptible to the revelation effect. Preceding tasks such as an anagram not only increase familiarity judgments but also the judged frequency of a word in the study list (Bornstein & Neely, 2001) or the belief that an event occurred in one's own autobiography. For example, after solving an anagram (e.g., iwdnwo – window), participants are more likely to indicate that a statement such as "*I broke a window playing ball*" describes an event from the participants' childhood (Bernstein, Godfrey, Davison, & Loftus, 2004; Bernstein, Rudd, Erdfelder, & Loftus, 2009; Bernstein, Whittlesea, & Loftus, 2002).

Young et al. (2009) also found the revelation effect in metamemory. In their experiment, participants first studied cue-target word pairs (e.g., apple – cloud) and later tried to recall the target word (cloud) when being presented with the cue word (apple). Whenever participants did not recall the target word, they additionally provided a *feeling-of-knowing* (FOK) judgment, indicating their belief that they would later recognize the target word in a recognition test. Critically, when participants could not recollect the target word and had just solved an anagram of the cue word (eplap – apple), they provided higher FOK judgments than without anagram. However, this revelation effect did not actually indicate better memory for target words with anagram than without. Instead, in the recognition test, participants did equally well for target words whose cues had appeared as an anagram compared to target words with non-anagram cues. This observation corresponds with one of the criteria for cognitive illusions described in the introduction to this

TABLE 18.1 Judgments for which the revelation effect reportedly occurs, sometimes occurs, or does not occur, along with example items and a description of the effect

Judgment: Example	*Effect*
Judgments for which the revelation effect *occurs*	
Recognition: Did you see *airplane* in the study list?	More recognition claims and higher recognition confidence (e.g., Watkins & Peynircioglu, 1990)
Frequency: How often did *airplane* appear in the study list?	Higher frequency judgments (Bornstein & Neely, 2001)
Autobiographical: Before the age of 10, I broke a *window* playing ball.	Higher agreement ratings (e.g., Bernstein et al., 2002)
Feeling-of-knowing (FOK): How likely will you remember the word later?	Higher FOK ratings (Young, Peynircioglu, & Hohman, 2009)
Truth: Was Scorcese the director of "The Godfather"?	More "true" responses (Bernstein et al., 2002)
Preference: How much do you prefer this brand over another brand?	Higher preference ratings (Kronlund & Bernstein, 2006)
Judgments for which the revelation effect *sometimes occurs*	
Associative recognition: Did you see the word-pair *airplane-forensic* in the study list?	More recognition claims only under impoverished study conditions (Cameron & Hockley, 2000; Westerman, 2000)
Rhyme recognition: Did one of the study-list words rhyme with *spill*?	Higher agreement ratings only under impoverished study conditions (Mulligan, 2007)
Judgments for which the revelation effect *does not occur*	
Typicality: How typical is *apple* for the category *fruits*?	No revelation effect (Watkins & Peynircioglu, 1990)
Lexicality: Is *swerth* a word?	No revelation effect (Watkins & Peynircioglu, 1990)
Everyday frequency: How often have you encountered the word *airplane* in the past seven days?	No revelation effect (Watkins & Peynircioglu, 1990)
Personal relevance: Does the word *airplane* relate to you personally?	No revelation effect (Frigo, Reas, & LeCompte, 1999)
Plurality: Did you see *raindrop* (as compared to raindrops) in the study list?	No revelation effect (Westerman, 2000)

book: The preceding task increases familiarity ratings without any basis for this familiarity in "reality."

Although the revelation effect has been predominantly researched in the context of recognition judgments, it also occurs for other judgments. Bernstein et al. (2002) asked their participants to judge the truth of statements such as "*Scorcese*

was the director of 'The Godfather'." Participants provided higher truth ratings after solving an anagram of one of the words in the statement (e.g., escrosec – scorcese) compared to the no-task condition without anagrams. Kronlund and Bernstein (2006) found another instance of the revelation effect not directly related to recognition judgments. In their study, participants rated their preference for various brand names. Again demonstrating a revelation effect, brand names that participants first unscrambled in an anagram task received higher preference ratings than brand names that appeared without anagram.

Table 18.1 also lists judgments for which the revelation effect occurs only sometimes. These judgments are largely immune to the revelation effect because the default basis of these judgments is the recollection of the study event. Conversely, the revelation effect primarily occurs in recognition judgments that rely on vague feelings of familiarity (Landau, 2001; LeCompte, 1995). For example, associative recognition judgments require participants to recall whether two words appeared together in the study list. Even if both words appear familiar, this does not guarantee that the words actually appeared together. Similarly, rhyme recognition judgments require participants to recall the study-list words from memory and to judge whether any of these words rhyme with the test word (Mulligan, 2007). However, when memory for the study-list words is poor, the revelation effect also occurs for associative or rhyme-recognition judgments. Poor memory can occur, for example, when the presentation duration of the study words is very brief (e.g., 0.5 s per word). It has been argued that poor memory for study words increases the use of familiarity-based judgments, thus enabling the revelation effect (Cameron & Hockley, 2000; Mulligan, 2007; Westerman, 2000).

Preceding tasks

Research has identified several preceding tasks that cause a revelation effect (see Table 18.2). In most studies, the preceding task requires participants to identify a distorted word, for example, by solving an anagram or identifying a word that is revealed letter-by-letter (e.g., Watkins & Peynircioglu, 1990). Importantly, the revelation effect does not depend on the test item being part of the preceding task. For example, participants rate words as more familiar after a preceding task in which they solve a simple addition problem (e.g., 248 + 372 = ?) compared to the no-task condition (Niewiadomski & Hockley, 2001).

A common practice for revelation-effect studies is to ensure that participants either successfully complete the preceding task or to reveal the solution of the preceding task automatically before participants provide their judgment. For example, in a typical anagram task, participants receive a number code that will help them to solve the anagram (see Text box 18.1 for an example). To investigate whether the completion of the preceding task is a necessary condition for the revelation effect, Westerman and Greene (1998) gave their participants a 10 s deadline to identify fragmented words (e.g., ai_p_a_e). This led to a success rate of 49%. However, whether participants successfully completed the preceding task had no statistically significant effect on recognition judgments, and thus the revelation effect.

TABLE 18.2 Preceding tasks that reportedly cause a revelation effect along with task descriptions and example items

Task	*Description*	*Example*
Anagram	Participants solve an anagram of a word	ernpalia → airplane
Letter-by-letter	Participants identify a word that is revealed letter-by-letter	_ _r_ _a_e → a_r_a_e → a_rp_a_e
Addition	Participants solve an addition problem	248 + 372 = ?
Fragment or rotation	Participants identify a fragmented or rotated stimulus (e.g., a word or picture)	ai_p_a_e?
Memory span	Participants attempt to recall a sequence of digits or letters in their correct order	8, 3, 1, 6, 6, 9, 4, 1
Reverse typing	Participants type a word in reverse letter order	airplane → enalpria
Letter counting	Participants count the number of letters in a word following a specific rule	abandon → 2
Synonym generation	Participants generate a synonym for a word	abandon → leave

Measures

In the context of recognition experiments, participants judge their familiarity with a stimulus either with a dichotomous old-new response or a recognition-confidence judgment, for example, ranging on a 6-point scale from 1 ("sure new") to 6 ("sure old"). The revelation effect is the higher proportion of "old" judgments or the higher average recognition-confidence judgment in the task condition compared to the no-task condition.

Another common way to analyze the revelation effect involves *Signal Detection Theory* (SDT, e.g., MacMillan & Creelman, 2008). The basic components of SDT appear in Figure 18.1. The memory strength for targets ("old" test items) and lures ("new" test items) is modeled with two Gaussian distributions on a common memory strength dimension. Although the target distribution is placed higher on the memory strength dimension than the lure distribution, the two distributions may overlap because of imperfect memory for the targets. In SDT, a participant's recognition judgment for targets and lures is used to compute two latent variables: sensitivity and criterion. Sensitivity is the ability to discriminate targets from lures and, generally speaking, will be greater the less the target and lure distributions overlap. Conversely, the criterion represents a cut-off value on the memory strength dimension, above which a participant will call an item "old" irrespective of the item's true status. That is, only targets above the criterion and lures below the criterion will be classified correctly. See Text box 18.1 for equations on how to compute a participant's sensitivity and criterion.

FIGURE 18.1 The revelation effect for recognition judgments in the context of signal detection theory: The criterion in the task condition (c_{task}; intermittent line) is more liberal with regard to "old" responses compared to the criterion in the no-task condition ($c_{no\ task}$, continuous line). The figure depicts a typical scenario in which the variance of the target distribution is larger than that of the lure distribution.

TEXT BOX 18.1 A CLASSROOM DEMONSTRATION OF THE REVELATION EFFECT

Participants

In the proposed experiment, task presence (task vs. no task) is a within-subject manipulation. Assuming a medium-sized revelation effect ($f = 0.25$), balanced error probabilities ($\alpha = \beta = .05$) and a relatively high correlation between judgments in the task and no-task conditions ($r = .7$), the experiment requires 34 participants.

Material

The proposed experiment requires 70 words in total, of which 10 words serve as primacy and recency buffers in the study list, 30 words serve as study items, and 30 words serve as lures. A typical item set consists of common 8-letter nouns (e.g., *airplane*, *pleasure*, *railroad*).

Design

The experiment includes two experimental manipulations: whether a word appears during the study phase of the experiment and whether an anagram precedes the test word. Thus, the experiment consists of a 2 (word status: old vs. new) × 2 (task presence: task vs. no task) within-subject design with old-new judgments as the only dependent variable.

Procedure

The experiment begins with the presentation of the study list. In the classroom, it is perhaps most feasible to present the study list via projector to all participants simultaneously. In the literature, presentation times per item vary between 0.1 s and 5 s, with 2 s occurring most often. It is advisable to choose presentation times in the study list such that memory performance in the subsequent recognition test will be poor. Conversely, when participants remember the study list well, the preceding task has little to no effect on recognition judgments (Cameron & Hockley, 2000; LeCompte, 1995; Westerman, 2000). In addition to the 30 study words, the study list includes 5 primacy and 5 recency buffers which appear before and after the study words, respectively, and are not part of the recognition test.

After the study-list presentation, participants practice the anagram task. This ensures that all participants are familiar with the anagram task during the recognition test. An additional benefit of the anagram practice is that it increases the retention interval, thus further reducing memory performance and increasing the probability that the revelation effect will occur. Typically, a solution code appears below each anagram. The number in the solution code indicates the position of the letter in the anagram solution. Here is an example of an anagram and its solution code:

n e s n u h i s
3 8 1 7 2 5 6 4

In the example, the 1 below the letter *s* indicates the first letter of the anagram solution, the 2 below the letter *u* indicates the second letter of the anagram solution, and so forth. Applying this rule to all letters reveals the anagram solution *sunshine*. The anagram task can be practiced with the primacy or recency buffers, removing the need to add further items to the experiment.

In the classroom context, the recognition test is easily administered with a paper-and-pencil questionnaire listing the test words in a column (see the Appendix for example items and instructions). Thus, each row contains a test word or, in task trials, an anagram. The participants' task is to judge whether each test word is old, starting at the top of the page and working their way down the list. Participants can provide old-new judgments by circling the test words that they think are old. In task trials, participants have to solve the anagram first, write down the anagram solution, and then, immediately afterwards, circle the anagram solution if they think it had appeared in the study list. The recognition test is usually self-paced. However, participants should be instructed not to return to a previous item once they have made their judgment. Without this restriction, the classification of judgments into task and no-task conditions is difficult, if not impossible. Further, to minimize

item-specific effects, the assignment of words to the task condition should be counterbalanced across participants. To that end, it is advisable to distribute the test words into two lists: A and B. List A appears in the task condition and list B in the no-task condition for half of the participants and vice versa for the other half.

Analysis

Judgments are first aggregated across trials, separately for each participant, word status (old vs. new), and task condition (task vs. no task). For dichotomous old-new judgments, this yields the proportion of "old" responses for each participant and condition. The aggregated judgments are then subject to a repeated-measures analysis of variance with repeated-measures factors word status (old vs. new) and task presence (task vs. no task). The main effect of task presence indicates whether the anagrams elicited a revelation effect. However, in many studies, particularly those where the anagram solution is also the test word, as is the case here, the revelation effect is stronger for "new" words compared to "old" words. In the analysis of variance, this corresponds to a significant interaction effect between word status and task presence.

Another common way to analyze judgments in revelation-effect experiments is signal detection analysis. As outlined previously, when the anagram solution is also the test word, the preceding task tends to reduce the participants' sensitivity and criterion compared to the no-task condition. For dichotomous old-new judgments, sensitivity is $d' = z(H) - z(F)$ and the criterion is $c = -0.5\,[z(H) + z(F)]$, where H is the participant's hit rate (i.e., the proportion of "old" judgments for "old" words), F is the participant's false-alarm rate (i.e., the proportion of "old" judgments for "new" words) and z is the inverse of the standard-normal distribution function (MacMillan & Creelman, 2008). In SDT, the revelation effect corresponds to a more liberal – that is, numerically smaller – criterion in the task condition compared to the no-task condition, easily assessed with a *t*-test for dependent measures.

In a typical experiment, the preceding task decreases the criterion (c_{task} in Figure 18.1) compared to the no-task condition ($c_{no\ task}$ in Figure 18.1). Consequently, more targets and lures appear above the criterion in the task condition compared to the no-task condition. In contrast, the preceding task affects sensitivity only under specific circumstances. When the task item and the test item are identical – for example, when participants solve an anagram of the word *airplane* and then decide whether *airplane* is old – sensitivity tends to be lower in the task condition compared to the no-task condition. Conversely, when the task item and the test item differ – for example, when participants solve an anagram of the word *airplane* and

then decide whether the word *pleasure* is old – sensitivity is roughly equal in the task and no-task conditions (Verde & Rotello, 2004).

A classroom experiment

As a classroom demonstration of the revelation effect, I propose a typical revelation-effect experiment in which participants study a list of words and, directly afterwards, receive a recognition test for these words. In the recognition test, an anagram task precedes half of the test words. In the task condition, participants first solve the anagram and then judge whether the test word appeared in the study list. Conversely, in the no-task condition, the test word appears without anagram (e.g., Aßfalg & Nadarevic, 2015; Westerman & Greene, 1996, 1998). Text box 18.1 describes an experiment that follows this procedure.

Results

Experiments similar to the proposed classroom experiment elicit a revelation effect for recognition judgments and criterion estimates alike. Figure 18.2 depicts revelation effect sizes for experiments in which judgments include dichotomous old-new responses and the preceding task consists of solving anagrams (Hockley & Niewiadomski, 2001; Leynes, Landau, Walker, & Addante, 2005; Niewiadomski & Hockley, 2001; Peynircioğlu & Tekcan, 1993; Westerman, 2000). Because this exact

FIGURE 18.2 Revelation effect-sizes (standardized mean difference *g*; Hedges, 1981) as a function of word status (old vs. new) when judgments include dichotomous old-new responses and the preceding task consists of solving anagrams. Circles represent individual effect sizes and diamonds represent combined effect sizes for "old" and "new" words separately. The diamond widths represent the 95%-confidence intervals of the combined effect sizes. The dotted line marks the null effect.

combination is relatively rare, Figure 18.2 includes experiments in which either the anagram solution or an unrelated word serves as the test word. The diamond symbols in Figure 18.2 represent the combined effect sizes for "old" and "new" words, respectively, using a random-effects meta-analysis (Aßfalg, Bernstein, & Hockley, 2015). The combined effect sizes in Figure 18.2 translate to an average increase of 6% in hits and 14% in false alarms due to the preceding task.

Unfortunately, SDT results are not reported consistently in the literature and are therefore difficult to summarize. Different authors use different estimates of SDT measures and some studies neglect SDT altogether. However, of the studies contributing to Figure 18.2, some report SDT measures and find a significantly more liberal criterion in the task condition compared to the no task condition (Hockley & Niewiadomski, 2001; Leynes et al., 2005; Niewiadomski & Hockley, 2001).

Theoretical accounts

Test-item presentation time

Naturally, explanations of the revelation effect have evolved over time. Early explanations focused on memory and tried to explain why participants judge the test item as more familiar after revealing the test item compared to the no-task condition. Luo (1993) speculated that the added presentation time of the test item due to the preceding task in the task condition was responsible for the revelation effect. To test this assumption, the participants in Luo's experiment either judged the test item as soon as it appeared or after a delay of several seconds. However, this manipulation had no effect on recognition judgments, casting doubt on the hypothesis that the revelation effect results from varying presentation times.

Affirmation tendency

Westerman and Greene (1996) tested two competing hypotheses. According to the first hypothesis, the revelation effect is an affirmation tendency – that is, the preceding task induces a tendency to respond "yes" to the question whether the test item is old. According to a competing hypothesis, the preceding task elicits a feeling of familiarity and therefore causes a "yes" response as well. Westerman and Greene pitted these hypotheses against each other by changing the wording of the recognition question to "Is this word new?" If the preceding task induces an affirmation tendency, participants should answer "yes" irrespective of the wording of the question. However, contrary to this hypothesis, "no" judgments were more frequent in the task condition compared to the no-task condition. This result suggests that the preceding task induces a feeling of familiarity. Further support for this hypothesis came from another experiment by the same authors. Participants first studied two lists of words. In the subsequent recognition test, the participants had to judge whether the test word had appeared in List 1. If the preceding task induces an affirmation tendency, participants should answer this question with "yes." However, again, the

opposite pattern emerged. Participants responded "no" more often in the task condition compared to the no-task condition. This outcome is compatible with the familiarity hypothesis, assuming that words in the more recently studied list (List 2) appeared more familiar to the participants than the words in List 1. If revealed words appear more familiar than words in the no-task condition, they should be (mis-)attributed to the more recent, and therefore more familiar, study list. Thus, the results of Westerman and Greene's experiments suggest that the preceding task induces a feeling of familiarity rather than an affirmation tendency.

Global-matching hypothesis

The global-matching hypothesis postulates that the size of the revelation effect depends on the similarity between the task item and the study list. Specifically, the preceding task is thought to activate study-list memory traces. Based on global-matching models of memory, this activation is assumed to contribute to the perceived familiarity of the test item (Westerman & Greene, 1998).

A testable prediction of the global-matching hypothesis is that the similarity between the task item and study-list items determines the size of the revelation effect. Indeed, Westerman and Greene (1998) initially observed that the revelation effect occurs when the study-list items are words and the task item is a word – for example in an anagram task (enalpria – airplane) – but not when the task item is a number – for example, in an addition problem (e.g., 248 + 372 = ?). However, Niewiadomski and Hockley (2001) later found similar-sized revelation effects when the study list consisted of words and the preceding task involved either anagrams or addition problems. Similarly, Verde and Rotello (2003) had their participants solve anagrams of words (enalpria – airplane) or of numbers (372615 – 231657) in the preceding task. Both anagram conditions used a solution code to ensure that participants would change the letter or digit order as intended. Contradicting the global-matching hypothesis, the size of the revelation effect was roughly equal for both word and number anagrams when the study list consisted of words. Perhaps most problematic for the global-matching hypothesis is the study by Frigo et al. (1999). In this study, participants studied what they thought was a list of words, and then solved anagrams in the preceding task of a recognition test. As in many other experiments, a revelation effect emerged: After the preceding task, recognition claims increased compared to the no-task condition. Critically, however, this experiment lacked any real study list. Participants were instructed to study spoken words, heavily masked by noise which, in fact, lacked any actual words. In the absence of a study list, the activation of study-list memory traces could not have contributed to the revelation effect.

Decrement-to-familiarity hypothesis

According to the decrement-to-familiarity hypothesis, the preceding task activates competing solutions to the task. This activation still persists during the recognition judgment, increasing the difficulty of the recognition judgment and lowering the

familiarity of the test item. As a reaction to the increased judgment difficulty, it is assumed that participants employ a more liberal response criterion compared to the no-task condition. This liberal response criterion in turn causes the revelation effect. Hicks and Marsh (1998) proposed the decrement-to-familiarity hypothesis after testing the effect of the preceding task in two-alternative-forced-choice (2-AFC) studies. In Hicks and Marsh's 2-AFC studies, participants had to decide which of two words had appeared in the study list after one of the words was revealed letter-by-letter. When both items were old or when both items were new, participants preferred the word that was not revealed. This pattern of results supports the assumption that the preceding task actually reduces the familiarity of the revealed word.

As Niewiadomski and Hockley (2001) pointed out, one of the criticisms of the global-matching hypothesis also applies to the decrement-to-familiarity hypothesis: The similarity between the task item and the study-list items should influence whether activated information interferes with the recognition decision. As described earlier, evidence for this assumption is mixed. Support for the decrement-to-familiarity hypothesis also depends on whether the task and test items are identical. The 2-AFC procedure is limited to experiments in which the task and test items are identical. Thus, it is unclear whether test items also appear less familiar after a preceding task with an unrelated task item. Further, neurological correlates of familiarity indicate that familiarity with the test item reduces immediately following the preceding task (Azimian-Faridani & Wilding, 2004). However, this is only the case when the task and test items are identical, not when they are, for example, unrelated words (Leynes et al., 2005). Finally, the preceding task only increases the difficulty of the recognition judgment, as measured by the decreased ability to discriminate targets and lures (i.e., sensitivity), when the task and test items are identical – for example, when the solution of an anagram is also the test item – but not when they differ – for example, when the solution of an anagram is unrelated to the test item (Verde & Rotello, 2004). In sum, the empirical support for the familiarity-decrement hypothesis is mixed and depends on whether the task item is related to the test item.

Criterion-flux hypothesis

An interesting aspect of the revelation effect is that participants apply varying response criteria from trial to trial depending on whether the trial involves the preceding task. This happens, according to the criterion-flux hypothesis, because the preceding task affects working memory (Hockley & Niewiadomski, 2001; Niewiadomski & Hockley, 2001). According to the hypothesis, working memory is assumed to contain important test-list information such as the proportion of studied items in the recognition test. This information determines what criterion participants apply to the judgment. Further, the preceding task is assumed to temporarily remove this information from working memory, forcing participants to rely on default assumptions about the test list. In this situation, the hypothesis postulates

that participants apply a more liberal response criterion than usual, thus causing the revelation effect. Once the effect of the preceding task wears off, the original criterion is re-established, leading to a constant flux between criteria in task and no-task trials.

According to the hypothesis, the revelation effect should increase with the degree of working-memory involvement in the preceding task. However, when the preceding task is a classical working-memory task (i.e., memory span), the difficulty of the task has no bearing on the size of the revelation effect. Specifically, 3-letter versus 8-letter memory spans in the preceding task elicit roughly equal revelation effects, despite much better performance in the 3-letter-span task (95% correct responses) compared to the 8-letter-span task (7% correct responses) (Westerman & Greene, 1998).

Discrepancy-attribution hypothesis

It has been argued that fluency – the perceived ease and speed of information processing – serves as a heuristic cue for various judgments such as familiarity, truth, and aesthetical pleasure (e.g., Alter & Oppenheimer, 2009). Specifically, the more fluently a stimulus is processed, the more familiar, true, and likeable this stimulus will appear (see Chapters 10, 13, and 14 in this volume). Importantly, changes in fluency will not automatically influence judgments. Instead, the observed fluency has to violate an expectation about fluency (Whittlesea & Williams, 1998). In the context of a revelation-effect experiment, it is the preceding task that affects fluency. Bernstein et al. (2002) argued that, at first, the preceding task might appear hard to solve. However, if participants solve the preceding task relatively easily, a discrepancy between expected and observed fluency arises that is subsequently misattributed to familiarity with the stimulus. Additionally, in preceding tasks such as anagrams, participants likely process the anagram (e.g., ernpalia) less fluently on a perceptual and conceptual level than the anagram solution (airplane), which might further add to the discrepancy between expected and observed fluency (Aßfalg & Bernstein, 2012).

The discrepancy-attribution hypothesis has two major advantages over other extant explanations of the revelation effect. First, it addresses various judgment types, not just recognition judgments. Fluency is believed to inform judgments that are also susceptible to the revelation effect such as recognition, truth, or preference judgments (Alter & Oppenheimer, 2009). The general idea is that fluency is correlated with these judgments in everyday life, providing a valid heuristic during judgment formation. However, because fluency is correlated with numerous concepts, it is also easy to misattribute fluency to a false source such that fluency originating from the preceding task influences the following judgment (Oppenheimer, 2008).

The second advantage of the discrepancy-attribution hypothesis is that it offers a framework to explain reversals of the revelation effect. Aßfalg and Bernstein (2012) observed such a reversal when participants judged the familiarity of faces after

solving a face puzzle in the preceding task: Revealed faces received lower familiarity ratings than faces in the no-task condition. According to the over-discounting hypothesis, participants are able to discount fluency as a cue for their judgment if fluency is irrelevant for their judgment and if they become aware of the irrelevant fluency source (Oppenheimer, 2006). This could also explain why fluency has to be unexpected in order to influence judgments: Expected fluency can be easily discounted (Alter & Oppenheimer, 2009; Whittlesea & Williams, 1998). In a revelation-effect experiment, fluency induced by the preceding task is indeed irrelevant for the judgment because half of the "old" stimuli and half of the "new" stimuli appear in the task condition. However, in the absence of perfect calibration, participants are believed to over-discount, essentially reversing any fluency-based effect. In the case of the revelation effect, over-discounting could arise because some aspect of the experimental procedure increases the participants' awareness of the preceding task as an irrelevant source of fluency.

Although the discrepancy-attribution hypothesis can explain why the revelation effect occurs for various judgments and sometimes even reverses (Aßfalg & Bernstein, 2012; Bernstein et al., 2002; Kronlund & Bernstein, 2006), it has garnered criticism for conceptual and empirical reasons. Westerman and Greene (1996) argued that the discrepancy-attribution hypothesis is implausible conceptually, because it is unclear how the fluency change occurring during the preceding task should affect an unrelated judgment. Empirically, attempts to eliminate or even reverse the revelation effect by warning participants about the revelation effect – and thus induce over-discounting – failed. The revelation effect occurs whether or not participants are informed about the revelation effect and even occurs when participants demonstrably understand the instructions to suppress the effect and believe to have been successful (Aßfalg & Nadarevic, 2015). On a more general level, current versions of the discrepancy-attribution hypothesis include too many degrees of freedom, allowing for fluency to influence judgments, to be discounted (Whittlesea & Williams, 1998), or even to be over-discounted (Oppenheimer, 2006), thus, offering post-hoc explanations for any possible outcome of an experiment. This would not be a problem if a reliable, independent measure of fluency could be used to test changes in fluency over time. However, typical measures such as reaction times or self-reports have been criticized either because they are too task-specific and ignore that only discrepant fluency affects judgments or because of their unreliability (Oppenheimer, 2008). Thus, while the discrepancy-attribution hypothesis offers great potential to explain the revelation effect, it is not yet developed enough to do so convincingly.

Conclusions

From a theoretical perspective, the revelation effect is, if not an adaptive ability, then at least a by-product of human cognition. As such, the revelation effect offers a window into how judgments form. Two observations suggest that the cognitive mechanisms that elicit the revelation effect are very general in nature: Firstly,

the revelation effect occurs for various judgments including familiarity, frequency, metacognition, truth, and preference. Secondly, the revelation effect is triggered by various preceding tasks such as identifying a stimulus, the reverse typing of a word, or the mental calculation of a sum. Another intriguing aspect of the revelation effect is its low metacognitive accessibility: The revelation effect appears even when participants believe they successfully suppressed the effect.

Challenges for research include finding the source of the reversed revelation effect and an explanation why the revelation effect appears unreliably in some subpopulations. Specifically, the revelation effect is well-documented in young adults and has been replicated in children but could not be replicated consistently in older adults. Prull, Light, Collett, and Kennison (1998) found a revelation effect only for younger but not for older adults, whereas Thapar and Sniezek (2008) found the revelation effect for younger and older adults alike. However, the major challenge for future research remains finding a plausible explanation of the revelation effect. Extant hypotheses are either too vague, too narrow in their application, focus too much on recognition judgments while neglecting other judgments, or rely on assumptions that have been falsified.

Summary

- Directly following a preceding task such as solving an anagram (ernpalia → airplane) or an addition problem (248 + 372 = ?), stimuli appear more familiar, frequent, true, or preferable than without a preceding task – the revelation effect.
- The revelation effect occurs in a state of uncertainty: In the context of recognition experiments, the preceding task affects judgments of familiarity rather than recollection.
- Most evidence suggests that the similarity between the preceding task and the study or test items has little influence on the size of the revelation effect.
- The revelation effect also occurs even when participants think they successfully prevented the effect.
- No extant hypothesis can account for the entirety of revelation-effect findings.

Further reading

Currently, there is no published meta-analysis of the revelation effect (but see Aßfalg et al., 2015). However, there are some key findings in the literature that deserve special attention. Westerman and Greene (1996, 1998) ruled out several explanations of the revelation effect and discovered that the revelation effect also occurs if the preceding task and the judgment are semantically unrelated such that, for example, an anagram of the word *airplane* triggers a revelation effect for the word *pleasure*. Bernstein and his colleagues (Bernstein et al., 2002; Kronlund & Bernstein, 2006) established that the revelation effect is a more general cognitive illusion than previously thought by demonstrating the effect for judgments of preference and truth.

Acknowledgements

I am grateful to Daniel Bernstein, Lena Nadarevic, and Lena Jooß for helpful comments on earlier versions of this chapter. Some of the author's studies cited in this chapter have been supported by the German Research Foundation (DFG AS 427/1–1).

References

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.

Aßfalg, A., & Bernstein, D. M. (2012). Puzzles produce strangers: A puzzling result for revelation-effect theories. *Journal of Memory and Language*, *67*, 86–92.

Aßfalg, A., Bernstein, D. M., & Hockley, W. (2015). The revelation effect: A meta-analytic test of hypotheses. Manuscript submitted for publication.

Aßfalg, A., & Nadarevic. L. (2015). A word of warning: instructions and feedback cannot prevent the revelation effect. *Consciousness and Cognition, 34*, 75–86.

Azimian-Faridani, N., & Wilding, E. L. (2004). An event-related potential study of the revelation effect. *Psychonomic Bulletin & Review*, *11*, 926–931.

Bernstein, D. M., Godfrey, R. D., Davison, A., & Loftus, E. F. (2004). Conditions affecting the revelation effect for autobiographical memory. *Memory & Cognition*, *32*, 455–462.

Bernstein, D. M., Rudd, M. E., Erdfelder, E., Godfrey, R., & Loftus, E. F. (2009). The revelation effect for autobiographical memory: A mixture-model analysis. *Psychonomic Bulletin & Review*, *16*, 463–468.

Bernstein, D. M., Whittlesea, B. W., & Loftus, E. F. (2002). Increasing confidence in remote autobiographical memory and general knowledge: Extensions of the revelation effect. *Memory & Cognition*, *30*, 432–438.

Bornstein, B. H., & Neely, C. B. (2001). The revelation effect in frequency judgment. *Memory & Cognition*, *29*, 209–213.

Cameron, T. E., & Hockley, W. E. (2000). The revelation effect for item and associative recognition: Familiarity versus recollection. *Memory & Cognition*, *28*, 176–183.

Frigo, L. C., Reas, D. L., & LeCompte, D. C. (1999). Revelation without presentation: Counterfeit study list yields robust revelation effect. *Memory & Cognition*, *27*, 339–343.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*, 107–128.

Hicks, J. L., & Marsh, R. L. (1998). A decrement-to-familiarity interpretation of the revelation effect from forced-choice tests of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1105–1120.

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.

Kronlund, A., & Bernstein, D. M. (2006). Unscrambling words increases brand name recognition and preference. *Applied Cognitive Psychology*, *20*, 681–687.

Landau, J. D. (2001). Altering the balance of recollection and familiarity influences the revelation effect. *The American Journal of Psychology*, *114*(3), 425–437.

LeCompte, D. C. (1995). Recollective experience in the revelation effect: Separating the contributions of recollection and familiarity. *Memory & Cognition*, *23*, 324–334.

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.

Luo, C. R. (1993). Enhanced feeling of recognition: Effects of identifying and manipulating probes on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 405.

Macmillan, N. A., & Creelman, C. (2008). *Detection theory: A user's guide*. New York: Psychology Press.

Mulligan, N. W. (2007). The revelation effect: Moderating influences of encoding conditions and type of recognition test. *Psychonomic Bulletin & Review*, *14*, 866–870.

Niewiadomski, M. W., & Hockley, W. E. (2001). Interrupting recognition memory: Tests of familiarity-based accounts of the revelation effect. *Memory & Cognition*, *29*, 1130–1138.

Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly. *Applied Cognitive Psychology*, *20*, 139–156.

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, *12*, 237–241.

Peynircioğlu, Z. F., & Tekcan, A. I. (1993). Revelation effect: Effort or priming does not create the sense of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 382–388.

Prull, M. W., Light, L. L., Collett, M. E., & Kennison, R. F. (1998). Age-related differences in memory illusions: Revelation effect. *Aging, Neuropsychology, and Cognition, 5*, 147–165.

Thapar, A., & Sniezek, S. M. (2008). Aging and the revelation effect. *Psychology and Aging, 23*, 473–477.

Verde, M. F., & Rotello, C. M. (2003). Does familiarity change in the revelation effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 739–746.

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.

Watkins, M. J., & Peynircioglu, Z. F. (1990). The revelation effect: When disguising probes induces recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 1012–1020.

Westerman, D. L. (2000). Recollection-based recognition eliminates the revelation effect in memory. *Memory & Cognition*, *28*, 167–175.

Westerman, D. L., & Greene, R. L. (1996). On the generality of the revelation effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1147–1153.

Westerman, D. L., & Greene, R. L. (1998). The revelation that the revelation effect is not due to revelation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 377.

Whittlesea, B. W., & Williams, L. D. (1998). Why do strangers feel familiar, but friends don't? A discrepancy-attribution account of feelings of familiarity. *Acta Psychologica*, *98*, 141–165.

Young, K. D., Peynircioglu, Z. F., & Hohman, T. J. (2009). Revelation effect in metamemory. *Psychonomic Bulletin & Review*, *16*, 952–956.

APPENDIX

Example items and instructions for the paper-pencil recognition test in a classroom demonstration:

Please circle the words that appeared in the study list. Begin at the top of the page and work your way down the list. In rows with an anagram, first solve the anagram and write down the solution on the line next to the anagram. As practiced before, use the number code below the anagram to find the anagram solution. Then circle the anagram solution if it appeared in the study list. Importantly, please judge each word in the order provided and do not return to a word that you have already judged.

1 e r n p a l i a _____

8 3 7 4 1 5 2 6

2 p l e a s u r e

3 n e s n u h i s _____

3 8 1 7 2 5 6 4

4 r a i l r o a d

. . .

19

SURVIVAL PROCESSING EFFECT

Meike Kroneisen and Edgar Erdfelder

> *Memory is a gift of nature, the ability of living organisms to retain and utilize acquired information or knowledge. [. . .] Owners of biological memory systems are capable of behaving more appropriately at a later time because of their experiences at an earlier time, a feat not possible for organisms without memory.*
>
> Tulving (1995, p. 751)

Why do we remember? Research on memory is often concerned with the question of how our memory system works. It typically focuses on structural mechanisms, for example, how do we process and store information? How long can we hold information in our working memory? However, when thinking about the question of why human memory exists, it seems unlikely that it only evolved to learn, process, and store abstract information. If our memory system is a solution to adaptive problems, shaped by the process of natural selection, then its structural properties should reflect their functionality (Tooby & Cosmides, 1992). However, what are the adaptive problems human episodic memory was designed to solve? Again, it seems implausible that its function is only to remember the past. It seems more plausible that we need to remember the past to predict the likelihood of events occurring in the future (Suddendorf & Corballis, 1997; Tulving, 2002). Specifically, memory could be designed to retain information relevant to survival, for example, remembering the location of food or water.

The survival processing effect

Nairne, Thompson, and Pandeirada (2007) claimed that nature "tuned" our memory systems to process and remember fitness-relevant information. Our ancestors therefore had survival advantages and were more likely to reproduce. In order to test

this prediction, Nairne et al. (2007) asked participants to imagine a survival scenario (being stranded on the grasslands of a foreign country; see the classroom experiment in Text box 19.1 for the instructions) and then rate up to 30 words regarding their relevance to this situation. A second group of participants had to imagine a different

TEXT BOX 19.1 CLASSROOM EXPERIMENT

Material

Stimulus materials are taken from the updated Battig and Montague norms (Van Overschelde, Rawson, & Dunlosky, 2004) and consist of 30 typical members drawn from 30 unique categories (for a complete list, see the Appendix). All words are presented in random order.

Design

The simplest design is used here, with only one randomized between-subject factor. One half of the participants receives the survival, the other half the moving scenario. All participants are asked to rate the same words in one of the two rating scenarios. The rating task should be followed immediately by a short, 2-minute distractor task prior to a final unexpected free-recall task. Except for the rating scenarios, all aspects of the design, including timing, are held constant across participants. Proportion correct recall serves as the dependent variable.

Procedure

Depending upon experimental condition, the participants first read one of the following instructions:

> *Survival*
> *In this task, we would like you to imagine that you are stranded in the grasslands of a foreign land, without any basic survival materials. Over the next few months, you'll need to find steady supplies of food and water and protect yourself from predators. We are going to show you a list of words, and we would like you to rate how relevant each of these words would be for you in this survival situation. Some of the words may be relevant and others may not – it's up to you to decide.*
>
> *Moving*
> *In this task, we would like you to imagine that you are planning to move to a new home in a foreign land. Over the next few months, you'll need to*

locate and purchase a new home and transport your belongings. We are going to show you a list of words, and we would like you to rate how relevant each of these words would be for you in accomplishing this task. Some of the words may be relevant and others may not – it's up to you to decide.

After reading the scenario description, stimuli should be presented for 5 s each, and participants should be asked to rate the words according to their relevance to the relevant scenario on a 5-point scale, with 1 = totally irrelevant and 5 = extremely relevant. Following the rating task, a short distractor task is presented, for example, a digit-recall task. For this task, seven digits ranging between zero and nine should be presented sequentially for 1 s each, and participants are required to recall the digits in order by writing responses into a text box. Following the distractor task, instructions for the surprise free recall task should appear. Participants are instructed to write down the previously rated words, in any order, on a response sheet. The final recall phase proceeds for 10 minutes.

scenario, namely moving to a foreign country (see the classroom experiment for the instructions). The task of this group was to rate the relevance of words presented to them according to the moving scenario. A third group had to rate the pleasantness of the words only. Each word was presented for 5 seconds, and participants were asked to rate the words on 5-point scales, with 1 indicating totally irrelevant (or unpleasant) and 5 signifying extremely relevant (or pleasant). The material consisted of typical members from 30 unique categories of concrete words. Since Nairne and collaborators were specifically interested in how survival versus non-survival processing affects retention in general, the words were randomly selected (e.g., chair, aunt, door) and did not follow a specific structure, schema, or script. A surprise retention test two minutes later showed that survival-based processing yielded better subsequent retention than did the other encoding procedures. Figure 19.1 shows the results of the free recall test for all three conditions.

In addition, Nairne et al. (2007) were able to show that the survival recall advantage holds for both free recall and for recognition memory tests irrespective of whether experimental conditions are manipulated within or between subjects. Other researchers have shown that the effect still occurs, for example, when using pictures instead of words (e.g., Otgaar, Smeets, & van Bergen, 2010) or emotionally arousing control scenarios rather than the moving control scenario (e.g., Kang, McDermott, & Cohen, 2008). Furthermore, the survival processing effect also shows up in judgments of learning (Palmore, Garcia, Bacon, Johnson, & Kelemen, 2012), implying that participants are aware of the memorial benefits of survival processing during learning. The effect was replicated not only in various samples of young adults from all over the world (Howe & Derbish, 2010; Kroneisen & Erdfelder, 2011; Nairne et al., 2007; Otgaar et al., 2010) but also in children (Aslan & Bäuml, 2012; Otgaar & Smeets, 2010, Exp. 2) and in older adults (Nouchi, 2012).

FIGURE 19.1 Average proportion of correct recall, sorted by condition (survival scenario vs. moving scenario vs. pleasantness rating) in Experiment 1 from Nairne et al., 2007 (from "Adaptive memory: Survival processing enhances retention" by James S. Nairne, Sarah R. Thompson, and Josefa N. S. Pandeirada, 2007, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Volume 33, p. 265. Copyright © 2007 American Psychological Association. Adapted with permission of the publisher.).

Moreover, Weinstein, Bugg, and Roediger (2008) found the survival processing effect only when ancestral survival scenarios were used, not when the scenario was changed to fit into a modern world (i.e., by replacing the grasslands scenario with a modern city scenario). Hence, the survival processing advantage appears to be limited to scenarios consistent with environments typical for human evolution (Nairne, Pandeirada, Gregory, & Van Arsdall, 2009).

However, there is also considerable evidence that challenges the idea of an evolutionary nature of this effect. Howe and Derbish (2010) found that not only correct but also false recall is enhanced by survival processing. When both true and false recall and true and false recognition were taken into account, a null effect of the survival advantage was observed (see also Otgaar & Smeets, 2010). In addition, no survival processing benefit was found for memory of faces (Savine, Scullin, & Roediger, 2011), for indirect memory tests (Tse & Altarriba, 2010), and for source memory (Bröder, Krüger, & Schütte, 2011). When the standard survival scenario was compared with threatening fictitious scenarios (a zombie attack), recall was even better for this unrealistic (i.e., evolutionary irrelevant) zombie scenario (Soderstrom & McCabe, 2011). Furthermore, when the survival scenario was compared to other fitness- or evolutionary-irrelevant scenarios, like being lost in space, the survival processing advantage sometimes disappeared (Kostic, McFarlan, & Cleary, 2012). These results suggest that, in contrast to the results of Weinstein et al. (2008), the mnemonic advantage of survival processing is not limited to ancestral contexts and does not show up in all types of memory measures. Thus, explaining the full pattern of results observed with respect to survival processing is not a trivial exercise.

Possible mechanisms underlying the effect

When trying to explain the survival processing effect, it is important to distinguish between ultimate and proximate explanations. Scott-Phillips, Dickins, and West (2011) pointed out that, "ultimate explanations are concerned with the fitness consequences of a trait or behavior and whether it is (or is not) selected. In contrast, proximate explanations are concerned with the mechanisms that underpin the trait or behavior – that is, how it works. Put another way, ultimate explanations address evolutionary function (the 'why' question), and proximate explanations address the way in which that function is achieved (the 'how' question)" (p. 38).

Nairne and colleagues (Nairne, 2010; Nairne & Pandeirada, 2008, 2010, 2011; Nairne, Pandeirada, & Thompson, 2008; Nairne et al., 2007) claimed that the survival processing effect can be seen as evidence that human learning and memory systems have been selectively tuned during evolution to process and retain information that is relevant to fitness (selective-tuning hypothesis). Therefore, the survival processing advantage reveals the ultimate function of human episodic memory, namely, to enable encoding, storage, and retrieval of survival-relevant information in the first place. We can remember information that was previously evaluated with respect to its survival relevance especially well because this skill facilitates adaptive behaviors that help us survive.

However, this ultimate explanation does not suffice for a complete understanding of the survival processing effect. An answer to the question *why* there is a mnemonic advantage of survival processing does not provide us with an answer to the question *how* this advantage comes about. More precisely, knowing why our memory system contributes to inclusive-fitness maximization does not imply anything with regard to how memory works in specific encoding and retrieval contexts. In the following, we focus on proximate explanations of the survival processing advantage with the goal to uncover the memory mechanisms driving this effect. As several possible proximate mechanisms are already ruled out conclusively, we included only the most promising explanations in our review (for a more detailed overview, see Erdfelder & Kroneisen, 2014).

Proximate explanations of the survival processing effect

Planning

Klein, Robertson, and Delton (2010, 2011) saw memory as future-oriented. According to their theory, "memory can be viewed as the result of the complex interplay of a set of processes that enable the organism to draw on past experiences to guide current behaviour and plan for future contingencies" (p. 122). They explain the survival processing effect by assuming that survival processing encourages future-oriented thoughts of planning more strongly than control conditions do, thus serving the ultimate function to facilitate future decisions. In their experiments, Klein et al. (2011) compared different groups that either read scenarios involving (1) survival without planning, (2) planning without survival, or (3) both survival

and planning. In line with their prediction, recall performance was highest for groups engaged in planning (with and without the survival component) and lowest for that group engaging in survival processing without any planning component. The authors concluded that planning for future events is "a specific, evolved set of mechanisms designed to help solve a general problem – how to remain alive long enough to reproduce and care for one's offspring" (p. 135).

Affective explanations

One influential explanation of the effect maintains that the survival processing advantage is due to emotion or arousal. This idea was already discussed by Nairne et al. (2007). There is evidence that items or stimuli that represent threats (e.g., guns, snakes, or aggressive faces) affect attention and memory (e.g., Kensinger, 2007). The survival processing effect can be seen as a variant of emotional influences on memory: Survival processing stimulates negative emotions, which may in turn recruit additional resources that may support the encoding and retrieval of information (D'Argembeau & van der Linden, 2004; Doerksen & Shimamura, 2001). However, the effects of emotion on memory are not straightforward. Sometimes negative information can be remembered better (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001); sometimes positive information is prioritized (D'Argembeau & van der Linden, 2011).

The survival processing literature provides no evidence that the survival processing advantage is mediated by emotional arousal (Kang et al., 2008; Soderstrom & McCabe, 2011) or by stress (Smeets, Otgaar, Raymaekers, Peters, & Merckelbach, 2012). However, there is some support for the assumption that the survival processing effect is moderated by the perceived threat within the survival situation such that higher perceived threat boosts the survival processing advantage (Olds, Lanska, & Westerman, 2014; Soderstrom & McCabe, 2011).

Soderstrom and McCabe (2011), for example, constructed different control scenarios identical to Nairne et al.'s (2007) grasslands survival scenario except that two words were exchanged. In one set of conditions, the word "grasslands" was replaced by "city", transforming the ancestral survival scenario in a modern survival scenario. In a different set of conditions, the word "predator" was exchanged against "zombie", transforming the realistic survival scenario into an unrealistic threat scenario. Complete cross-classification of the scenario types resulted in four different scenarios, namely, survival-grassland, survival-city, zombie-grassland, and zombie-city. The two scenarios involving "zombies" were associated with higher arousal levels and were evaluated more negatively than the other conditions. Furthermore, the zombie scenarios produced highest levels of recall. However, the differences in arousal and valence ratings could not fully explain the observed pattern in recall rates. Even when the effects of arousal and valence ratings were statistically controlled, the mnemonic benefit of zombie processing compared to predator processing remained (Soderstrom & McCabe, 2011).

It is also possible that it is not arousal in general but specific aspects of arousal that explain the effect. Hart and Burns (2012) claimed that survival contexts evoke

the thought of death. There are several studies suggesting that death awareness results in emotional and behavioral changes, some of which may also affect encoding and retrieval (e.g., Landau, Greenberg, & Sullivan, 2009; Landau et al., 2004). In three experiments, Hart and Burns (2012) were indeed able to show that, compared to two different control conditions unrelated to death, priming of death-related thoughts prior to providing pleasantness ratings enhanced later recall significantly (DTR-, *dying-to-remember,* effect). However, they did not include the survival scenario as a control condition. When the survival processing scenario was compared to a death-related scenario, the results were somewhat mixed. Burns, Hart, Kramer, and Burns (2014) found that the DTR-effect was completely eliminated when thoughts of mortality were combined with an orienting survival processing task. They concluded that processing induced by mortality salience is redundant with that required by the survival orienting task. Bell, Röer, and Buchner (2013), in contrast, compared the standard survival scenario with the moving scenario and a clearly death-related control scenario (floating in outer space with dwindling oxygen supplies). Here, the death-related scenario did not improve recall. Similar results were observed by Klein (2012). He compared a modified death scenario with both the standard survival scenario and a pleasantness rating control condition. Recall rates in the death scenario condition did not differ significantly from the control condition but were significantly worse than in the survival processing condition. However, Burns, Hart, and Kramer (2014) conducted a series of studies using dying and survival scenarios and were able to demonstrate that a dying scenario is equivalent to the survival scenario when specific variables are controlled. In their experiments, the scenarios were matched on thematic structure, concreteness, and relevance. In addition, possible congruity effects (see next paragraph) were controlled. When matched on these dimensions, no difference between the survival and the dying scenario was found.

Item congruity

Nairne et al. (2007) already discussed the possibility that the survival advantage might be mediated by a congruity effect (Schulman, 1974). Historically, the term "congruity effect" refers to the phenomenon that participants remember word stimuli previously judged with respect to their fit into sentence frames better when the words received a "yes" response than when they received a "no" response, presumably because these words were linked to the conceptual framework in which they were encoded and therefore better retrieved later on (e.g., Craik & Tulving, 1975; Moscovitch & Craik, 1976). To avoid congruity-biased item pools, Nairne et al. (2007) used random word lists from a variety of semantic categories with no obvious links to any of the scenario conditions. In addition, they compared average ratings between experimental conditions to check whether the survival recall advantage was confounded with higher survival relevance ratings, giving rise to a congruity effect explanation. According to their findings, the survival processing effect emerged in recall rates even when the average relevance ratings did not differ between conditions.

However, different results were obtained by Butler, Kang, and Roediger (2009). In their first experiment, they not only found the survival processing advantage but also a correlation between relevance ratings and recall rates for both the survival and the moving scenarios. Similar results were repeatedly observed by other researchers (e.g., Kroneisen & Erdfelder, 2011; Kroneisen, Erdfelder, & Buchner, 2013; Kroneisen, Rummel, & Erdfelder, 2014; Palmore et al., 2012), suggesting that within-list congruity effects are involved in either condition. Moreover, Butler et al. (2009) compared the survival scenario with a robbery scenario based on list items that were either highly relevant (e.g., "alarm" or "car" for the robbery scenario; "lion" and "fire" for the survival scenario) or irrelevant (e.g., "couch" or "jazz") to the respective scenarios. Compared to recall performance in the robbery-scenario condition, the results showed a survival processing memory advantage for survival-related word lists, a survival processing disadvantage for robbery-related words and no differences between scenarios for irrelevant words (Butler et al., 2009, Exp. 2 and Exp. 3, respectively). Importantly, the recall rate of robbery-congruent words in the robbery scenario condition was not worse than the recall rate of survival-congruent words in the survival scenario condition (for similar results, see Palmore et al., 2012).

However, not all results of Butler et al. (2009) were replicated. Nairne and Pandeirada (2011) still found the survival processing advantage even when they used the word lists from Butler et al. (2009). More precisely, they detected a survival processing effect when using only scenario-incongruent items (Nairne & Pandeirada, 2011, Exp. 2), only scenario-congruent items (Nairne & Pandeirada, 2011, Exp. 3), and equal amounts of congruent and incongruent items (Nairne & Pandeirada, 2011, Exp. 4).

Richness of encoding

Another idea to explain the survival processing effect was already suggested by Nairne et al. (2007; Nairne & Pandeirada, 2008): the idea that survival processing is a form of "deep" processing that leads to enhanced elaboration. Following this reasoning, Kroneisen and Erdfelder (2011) argued that the survival processing effect can be traced back to the richness of encoding triggered by the relevance-rating task. According to Craik and Tulving (1975), one important factor in depth of processing is not merely semantic encoding of information but rather the richness and distinctiveness with which information is encoded. Kroneisen and Erdfelder (2011) maintained that, in the survival-relevance rating task, participants are implicitly encouraged to think about different possible uses of objects in a complex survival context, both standard and nonstandard uses. This process generates a highly distinctive memory representation of list items during encoding, providing a large number of powerful memory cues (i.e., thoughts about different object functions) for the retrieval situation later on. Kroneisen and Erdfelder (2011) called this the richness-of-encoding (RE) hypothesis.

In their first two experiments, Kroneisen and Erdfelder (2011) tested this hypothesis by comparing the standard survival and moving scenarios with a reduced

survival scenario. In this reduced scenario, only a single survival problem (i.e., lack of potable water) was addressed. The authors argued that, as a consequence of the less complex encoding context, distinctiveness of memory representations diminishes and fewer retrieval cues are available in the subsequent free recall test compared to the standard scenario that addresses three survival problems (i.e., lack of potable water, lack of food, and predators). In line with these ideas, the survival processing advantage vanished both in within-subjects and between-subjects designs for the reduced scenario (Kroneisen & Erdfelder, 2011, Exp. 1 and 2, respectively). As shown in their third experiment, similar effects can be achieved by dispensing with the relevance-rating task and asking participants instead to state a single object use (simple encoding context) versus four different object uses (complex encoding context) for the survival versus moving scenarios. As predicted, there was a strong survival processing advantage for the complex encoding context but no such effect for the simple encoding context (Kroneisen & Erdfelder, 2011, Exp. 3).

Notably, when replacing the relevance rating task by an interactive imagery task (i.e., imagining use of an object in the respective scenario and rating the ease of this interactive imagery), the survival processing effect was also shown to vanish (Kroneisen et al., 2013). More precisely, interactive imagery reduced recall performance particularly in the survival, not in the moving condition. According to Kroneisen et al. (2013), forcing participants to engage in interactive imagery appears to distract them from the type of information processing that underlies the survival processing effect in standard relevance rating tasks. Corroborating the RE hypothesis more directly, Röer, Bell, and Buchner (2013) showed that the strength of the survival processing effect is a function of the number of unique relevance arguments generated per item. Moreover, Bell, Röer, and Buchner (2015) recently demonstrated that rating the usefulness of objects compared to rating the dangerousness of objects in a survival scenario leads to better recall.

Domain-general encoding processes such as elaboration and distinctive processing involve consciously controlled forms of encoding that are cognitively demanding. Therefore, they require working memory capacity. Hence, the survival processing effect should diminish when working memory resources are scarce. This conflicts with the evolutionary view that maintains stability of the survival processing advantage irrespective of working memory resources. Two recent studies indeed showed that survival processing diminishes under working memory load (Kroneisen et al., 2014; Nouchi, 2013; see, however, Stillman, Coane, Profaci, Howard, & Howard, 2014, for a different result).

In sum, all these studies suggest that a domain-general encoding process previously discussed in the depth-of-processing literature – richness of encoding – appears to be an important determinant of the survival-processing advantage.

Single-item and relational processing

Item-specific processing is concerned with encoding individual characteristics of items, whereas relational processing is concerned with encoding of the

relationships between list items (Burns, Hwang, & Burns, 2011). It is known that item-specific processing enhances the distinctiveness of items within a memory trace, whereas relational processing provides a structure for organizing these items within the trace (Burns, 2006). The combination of both item-specific and relational processing may also play an important role for the survival advantage (Burns et al., 2011). It is conceivable that survival-processing tasks prompt participants to engage in both item-specific and relational processing at the same time, a combination that has previously been shown to boost episodic memory (Einstein & Hunt, 1980; Hunt & Einstein, 1981; Hunt & McDaniel, 1993).

If relational processing is an important aspect for the survival advantage, then, according to Nairne and Pandeirada (2008), the survival processing effect should diminish when semantically unrelated words are replaced by categorized word lists from different semantic categories. Since these categorized word lists invite relational processing by default, additional relational processing due to survival processing should not lead to significantly better recall. However, Nairne and Pandeirada (2008) still found the survival processing effect in their experiment and therefore concluded that relational processing does not play a role in this effect. Burns et al. (2011), in contrast, criticized these studies because the categorized item material of Nairne and Pandeirada (2008) was survival-related (i.e., fruits, four-footed animals, . . .) and their measure of relational processing (i.e., the categorized list) may thus have been insensitive. In their experiments, they aimed at demonstrating that survival tasks require participants to engage in both item-specific and relational processing. Burns et al. (2011) compared the survival task with control tasks that involve either item-specific or relational or both processing forms. According to their results, the survival processing advantage is robust when the control condition contains only one component (item-specific or relational processing). However, when the control task involves both processing forms, the survival processing effect disappeared in their study.

However, there are still open questions that cannot be answered by the experiments conducted so far. For example, why does the survival processing task promote more item-specific and relational processing than control tasks do? Is it possible to modify the survival processing task so that single-item or relational processing increases or decreases selectively? A complete theoretical account of the survival processing effect should provide answers to research questions like these.

Conclusions

It is still an open debate which cognitive mechanisms are involved in the survival processing effect. In this chapter, we discussed different proximate explanations. However, we renounced to review all explanations mentioned in the literature but focused on the most promising mechanisms discussed so far, namely, (1) planning, (2) arousal, (3) item congruity, (4) richness of encoding, and (5) the combination of single-item and relational processing. These mechanisms of human memory, most of them well-known from related research on long-term episodic

memory, have proven to be successful in accounting for many results on moderators and possible mediators of the survival processing effect documented in the empirical literature (for a similar conclusion, see Howe & Otgaar, 2013).

As outlined in the literature summarized in this chapter, the survival processing advantage in free recall and recognition data is a robust and general phenomenon. In fact, the data support the conclusion that it is one of the best encoding procedures yet identified in human episodic-memory research. Given this generality, we do not consider it likely that this effect is accidental. It appears to reflect an evolved adaptive function of human episodic memory that contributes to inclusive-fitness maximization (Nairne et al., 2007). Nevertheless, we believe that this ultimate explanation does not necessarily imply a domain-specific proximate explanation of the mechanisms driving this effect.

As discussed above, there is considerable evidence that well-known domain-general mechanisms of human episodic memory such as richness of encoding can account for variations in the strength of the survival processing effect. Therefore, these domain-general mechanisms are promising candidates for a complete proximate explanation of the survival processing advantage.

Summary

- Episodic memory helped our ancestors to solve adaptive problems related to survival. The operating characteristics of episodic memory should "bear the imprints of the specific selection pressures that shaped their development" (Nairne & Pandeirada, 2010, p. 977).
- The survival processing effect is a strong and rather general memory advantage for word-material processed in a survival-related context.
- Ultimate explanation: Nature "tuned" our memory systems to process and remember fitness-relevant information.
- Proximate explanations:
    - Planning: Survival processing encourages future-oriented thoughts of planning.
    - Affective explanations (arousal and mortality salience): The survival processing effect is partly determined by the perceived threat within the survival situation. However, the empirical evidence is inconsistent for mortality salience.
    - Item congruity: The survival processing effect benefits from congruity of item content with the scenario. Whereas Butler et al. (2009) found evidence for congruity effects of relevance ratings and memory performance, Nairne and Pandeirada (2011) did not.
    - Richness of encoding: The survival scenario can be seen as fostering distinctive processing and elaborative encoding. Survival relevance ratings encourage persons to think about various possible uses of objects in a survival context, both typical and atypical uses. Thoughts about object functions later serve as powerful retrieval cues.

- Single-item and relational processing: The survival scenario encourages both item-specific and relational processing, resulting in an extremely good memory performance.

Further reading

We recommend the original article of Nairne et al. (2007) as an introduction to the survival processing literature. A rather comprehensive and timely overview of the current literature can be found in Erdfelder and Kroneisen (2014). Howe and Otgaar (2013) provided a more concise overview on proximate explanations of the survival processing advantage. In addition, there is a special issue of *Memory* (Otgaar & Howe, 2014) that covers selected theoretical and empirical aspects of the survival processing effect.

References

Aslan, A., & Bäuml, K.-H. (2012). Adaptive memory: Young children show enhanced retention of fitness-related information. *Cognition*, *122*, 118–122.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*, 323–370.

Bell, R., Röer, J. P., & Buchner, A. (2013). Adaptive memory: The survival-processing memory advantage is not due to negativity or mortality salience. *Memory & Cognition, 41*, 490–502.

Bell, R., Röer, J. P., & Buchner, A. (2015). Adaptive memory: Thinking about function. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 1038–1048.

Bröder, A., Krüger, N., & Schütte, S. (2011). The survival processing effect should generalize to source memory, but it doesn't. *Psychology*, *2*, 896–901.

Burns, D. J. (2006). Assessing distinctiveness: Measures of item-specific and relational processing. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 109–130). New York: Oxford University Press.

Burns, D. J., Hart, J., & Kramer, M. E. (2014). Dying scenarios improve recall as much as survival scenarios. *Memory, 22*, 51–64.

Burns, D. J., Hart, J., Kramer, M. E., & Burns, A. D. (2014). Dying to remember, remembering to survive: Mortality salience and survival processing. *Memory, 22*, 36–50.

Burns, D. J., Hwang, A. J., & Burns, S. A. (2011). Adaptive memory: Determining the proximate mechanisms responsible for the memorial advantages of survival processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 206–218.

Butler, A. C., Kang, S. H. K., & Roediger, H. L., III. (2009). Congruity effects between materials and processing tasks in the survival processing paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1477–1486.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*, 268–294.

D'Argembeau, A., & van der Linden, M. (2004). Influence of affective meaning on memory for contextual information. *Emotion, 4*, 173–188.

D'Argembeau, A., & van der Linden, M. (2011). Influence of facial expression on memory for facial identity: Effects of visual features or emotional meaning? *Emotion, 11*, 199–202.

Doerksen, S., & Shimamura, A. P. (2001). Source memory enhancement for emotional words. *Emotion, 1*, 5–11.

Einstein, G. O., & Hunt, R. R. (1980). Levels of processing and organization: Additive effects of individual item and relational processing. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 588–598.

Erdfelder, E., & Kroneisen, M. (2014). Proximate cognitive mechanisms underlying the survival processing effect. In B. L. Schwartz, M. Howe, M. Toglia, & H. Otgaar (Eds.), *What is adaptive about adaptive memory?* (pp. 172–198). New York: Oxford University Press.

Hart, J., & Burns, D. J. (2012). Nothing concentrates the mind: Thoughts of death improve recall. *Psychonomic Bulletin & Review*, *19*, 264–269.

Howe, M. L., & Derbish, M. H. (2010). On the susceptibility of adaptive memory to false memory illusions. *Cognition*, *115*, 252–267.

Howe, M. L., & Otgaar, H. (2013). Proximate mechanisms and the development of adaptive memory. *Current Directions in Psychological Science, 22*, 16–22.

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior, 20*, 497–514.

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language*, *20*, 497–514.

Kang, S. H. K., McDermott, K. B., & Cohen, S. M. (2008). The mnemonic advantage of processing fitness-relevant information. *Memory & Cognition*, *36*, 1151–1156.

Kensinger, E. A. (2007). Negative emotion enhances memory accuracy: Behavioral and neuroimaging evidence. *Current Directions in Psychological Science, 16*, 213–218.

Klein, S. B. (2012). The effects of thoughts of survival and thoughts of death on recall in the adaptive memory paradigm. *Memory, 22, 65–75*.

Klein, S. B., Robertson, T. E., & Delton, A. W. (2010). Facing the future: Memory as an evolved system for planning future acts. *Memory & Cognition*, *38*, 13–22.

Klein, S. B., Robertson, T. E., & Delton, A. W. (2011). The future-orientation of memory: Planning as a key component mediating the high levels of recall found with survival processing. *Memory*, *19*, 121–139.

Kostic, B., McFarlan, C. C., & Cleary, A. M. (2012). Extensions of the survival advantage in memory: Examining the role of ancestral context and implied social isolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1091–1098.

Kroneisen, M., & Erdfelder, E. (2011). On the plasticity of the survival processing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1552–1563.

Kroneisen, M., Erdfelder, E., & Buchner, A. (2013). The proximate memory mechanism underlying the survival processing effect: Richness of encoding or interactive imagery? *Memory, 21*, 494–502.

Kroneisen, M., Rummel, J., & Erdfelder, E. (2014). Working memory load eliminates the survival processing effect. *Memory, 22*, 92–102.

Landau, M. J., Greenberg, J., & Sullivan, D. (2009). Defending a coherent autobiography: When past events appear incoherent, mortality salience prompts compensatory bolstering of the past's significance and the future's orderliness. *Personality and Social Psychology Bulletin, 35*, 1012–1020.

Landau, M. J., Johns, M., Greenberg, J., Pyszczynski, T., Martens, A., Goldenberg, J. L., & Solomon, S. (2004). A function of form: Terror management and structuring the social world. *Journal of Personality and Social Psychology, 87*, 190–210.

Moscovitch, M., & Craik, F. I. M. (1976). Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall. *Journal of Verbal Learning and Verbal Behavior*, *15*, 447–458.

Nairne, J. S. (2010). Adaptive memory: Evolutionary constraints on remembering. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and behavior* (Vol. 53, pp. 1–32). San Diego, CA: Academic Press.

Nairne, J. S., & Pandeirada, J. N. S. (2008). Adaptive memory: Is survival processing special? *Journal of Memory and Language*, *59*, 377–385.

Nairne, J. S., & Pandeirada, J. N. S. (2010). Adaptive memory: Ancestral priorities and the mnemonic value of survival processing. *Cognitive Psychology*, *61*, 1–22.

Nairne, J. S., & Pandeirada, J. N. S. (2011). Congruity effects in the survival processing paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 539–549.

Nairne, J. S., Pandeirada, J. N. S., Gregory, K. J., & Van Arsdall, J. E. (2009). Adaptive memory: Fitness relevance and the hunter-gatherer mind. *Psychological Science*, *20*, 740–746.

Nairne, J. S., Pandeirada, J. N. S., & Thompson, S. (2008). Adaptive memory: The comparative value of survival processing. *Psychological Science*, *19*, 176–180.

Nairne, J. S., Thompson, S. R., & Pandeirada, J. N. S. (2007). Adaptive memory: Survival processing enhances retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 263–273.

Nouchi, R. (2012). The effect of aging on the memory enhancement of the survival judgment task. *Japanese Psychological Research*, *54*, 210–217.

Nouchi, R. (2013). Can the memory enhancement of the survival judgment task be explained by the elaboration hypothesis? Evidence from a memory load paradigm. *Japanese Psychological Research, 55*, 58–71.

Olds, J. M., Lanska, M., & Westerman, D. L. (2014). The role of perceived threat in the survival processing memory advantage. *Memory, 22*, 26–35.

Otgaar, H., & Howe, M. L. (2014). What kind of memory has evolution wrought? *Memory, 22*, 1–8.

Otgaar, H., & Smeets, T. (2010). Adaptive memory: Survival processing increases both true and false memory in adults and children. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1010–1016.

Otgaar, H., Smeets, T., & van Bergen, S. (2010). Picturing survival memories: Enhanced memory after fitness-relevant processing occurs for verbal and visual stimuli. *Memory & Cognition*, *38*, 23–28.

Palmore, C. C., Garcia, A. D., Bacon, L. P., Johnson, C. A., & Kelemen, W. L. (2012). Congruity influences memory and judgments of learning during survival processing. *Psychonomic Bulletin & Review*, *19*, 119–125.

Röer, J. P., Bell, R., & Buchner, A. (2013). Is the survival processing memory advantage due to richness of encoding? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1294–1302.

Savine, A. C., Scullin, M. K., & Roediger, H. L. (2011). Survival processing of faces. *Memory & Cognition*, *39*, 1359–1373.

Schulman, A. I. (1974). Memory for words recently classified. *Memory & Cognition*, *2*, 47–52.

Scott-Phillips, T. C., Dickins, T. E., & West, S. A. (2011). Evolutionary theory and the ultimate-proximate distinction in the human behavioral sciences. *Perspectives on Psychological Science*, *6*, 38–47.

Smeets, T., Otgaar, H., Raymaekers, L., Peters, M. J. V., & Merckelbach, H. (2012). Survival processing in times of stress. *Psychonomic Bulletin & Review*, *19*, 113–118.

Soderstrom, N. C., & McCabe, D. P. (2011). Are survival processing memory advantages based on ancestral priorities? *Psychonomic Bulletin & Review*, *18*, 564–569.

Stillman, C. M., Coane, J. H., Profaci, C. P., Howard, J. H., & Howard, D. V. (2014). The effects of healthy aging on the mnemonic benefit of survival processing. *Memory & Cognition, 42*, 175–185.

Suddendorf, T., & Corballis, M. C. (1997). Mental time travel and the evolution of the human mind. *Genetic, Social, and General Psychology Monographs, 123*, 133–167.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). New York: Oxford University Press.

Tse, C.-S., & Altarriba, J. (2010). Does survival processing enhance implicit memory? *Memory & Cognition, 38*, 1110–1121.

Tulving, E. (1995). Introduction to the section on memory. In Michael S. Gazzaniga (Ed.). *The cognitive neurosciences* (pp. 751–753). Cambridge: MIT Press.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53*, 1–25.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the norms. *Journal of Memory and Language, 50*, 289–335.

Weinstein, Y., Bugg, J. M., & Roediger, H. L. (2008). Can the survival recall advantage be explained by basic memory process? *Memory & Cognition*, *36*, 913–919.

APPENDIX

Word material

Stimuli for experiments (taken from Nairne et al., 2007, p. 273)

truck	juice	silver	door	car	silk
diesel	shoes	orange	broccoli	sword	teacher
mountain	finger	whiskey	bear	apartment	pan
pepper	aunt	flute	cathedral	soccer	sock
book	chair	snow	screwdriver	emerald	eagle

20

LABELLING AND OVERSHADOWING EFFECTS

Rüdiger F. Pohl

A label is something that is affixed.

(Link & Phelan, 2001, p. 368)

The labelling effect describes cases in which a verbal label is affixed to a stimulus and then exerts its distorting influence in subsequent judgment or recall of that stimulus (cf. the anchoring effect in Chapter 12, the misinformation effect in Chapter 22, and the hindsight bias in Chapter 23). Text box 20.1 lists a number of examples demonstrating labelling effects in different domains. This list already suggests that labelling effects may be rather widespread in our lives. More examples, from more applied domains, are given in Text box 20.2.

TEXT BOX 20.1 EXAMPLES OF LABELLING EFFECTS FROM DIFFERENT EXPERIMENTAL DOMAINS

If a white wine was labeled as "sweet," students who tasted the wine assigned larger amounts of residual sugar to it, than when the same wine was labeled as "dry" (Pohl, Schwarz, Sczesny, & Stahlberg, 2003). An odour that was introduced as "pleasant" received higher hedonic ratings than the same odour introduced as "unpleasant" (Herz & von Clef, 2001). Ambiguous line drawings that were labeled with a concrete noun were in many cases reproduced more in line with what the given label suggested than what the original figure actually contained (Carmichael, Hogan, & Walters, 1932). A blue-green colour that was labeled as "bluish" ("greenish") was later recalled as more blue (green) than it was (Bornstein, 1976). Eyewitnesses who saw a

car accident on video and were then asked to estimate the speed of the cars involved were influenced by the wording of the question. They estimated a higher speed when the question asked, "About how fast were the cars going when they *smashed* into each other?" than when the question was, "How fast were the cars going when they *hit* each other?" (Loftus & Palmer, 1974).

TEXT BOX 20.2 APPLIED DOMAINS WHERE LABELLING EFFECTS MIGHT OCCUR

Apart from the cognitive paradigms reviewed above, labelling occurs in other, more applied contexts as well. For example, developmental psychologists study the effects of labelling on young infants' ability to categorize objects by transitive inference (e.g., Sloutsky, Lo, & Fisher, 2001; Welder & Graham, 2001). Or, social psychologists are interested on how labelling contributes to impression formation (Higgins, Rholes, & Jones, 1977) and may thus foster stereotyping, prejudice, and stigmatization (e.g., Link & Phelan, 2001). In clinical psychology, the acceptability of psychological treatment methods was found to depend on the specific labels used to describe these methods (Woolfolk, Woolfolk, & Wilson, 1977; but see Katz, Cacciapaglia, & Cabral, 2000). The same reasoning also applies to the large domain of advertisement research studying, for example, how product acceptability (and thus the decision to buy) may be reached through the use of favorable product labels (e.g., Woolfolk, Castellan, & Brooks, 1983). And finally, the question of how eye- or earwitnesses can be influenced (and even misled) by such simple changes as using different words in a question is of major importance to forensic psychologists. Thus, labelling appears to play a role in quite a number of situations (many more than are reviewed in this chapter), infiltrating much of our daily lives.

But how can a verbal label distort memory? According to Paivio's (1969, 1971) dual-coding theory, verbal labels should generally help to memorize stimuli that are presented in another code (e.g., visual). But this is possibly true only as long as verbal labels and stimulus match, for example, when the word "pine tree" accompanies the picture of a "pine tree". But as soon as both even only slightly mismatch, a detrimental effect of the verbal label may occur, for example, when the (more general) word "tree" accompanies the picture of a "pine tree". In that case, memory for the pine-tree picture could be distorted to more prototypical instances of trees. The reversed example of a distorting influence would be if the picture is ambiguous (e.g., a prototypical tree) and the verbal label suggests a specific interpretation of it (e.g., a pine tree).

Labelling effects and their theoretical accounts

This section presents four cognitive domains in which labelling effects have been found. I begin with the classical examples studying the effects of labelling on the memory for ambiguous line drawings and other visual objects. Using the misinformation design (see Chapter 22), other researchers investigated how labelling might affect the memory for colour or speed. And finally, another domain of labelling concerns judgments of taste and odour.

Memory for visual objects

In one of the most influential studies on the labelling effect, Carmichael et al. (1932) presented a set of relatively ambiguous line drawings. Immediately before seeing a figure, each participant received one of two words (Label 1 or 2) describing the figure that was presented next (see Table 20.1 for the complete material). These words had

TABLE 20.1 The 12 ambiguous visual stimuli with two lists of labels (Carmichael et al., 1932)

List 1	*Figures*	*List 2*
Bottle		Stirrup
Crescent moon		Letter "C"
Bee hive		Hat
Eye glasses		Dumbbells
Seven		Four
Ship's wheel		Sun
Hour glass		Table
Kidney bean		Canoe
Pine tree		Trowel
Gun		Broom
Two		Eight
Curtains in a window		Diamond in a rectangle

different meanings but could both be applied to name the same ambiguous figure. In the end, participants were asked to draw the figures from memory in any order.

The authors found that from all drawings that were most deviant from the original ones (i.e., 25.9%), 74% and 73% could be classified as being closer to the named object (Label 1 or 2, respectively) than to the original figure. This observation, compared to only 45% of such cases in a control group (with no labels), led Carmichael et al. (1932, p. 81) to state that "naming a form immediately before it is visually presented may in many cases change the manner in which it will be reproduced."

These results conform to the process of "assimilative perception" as proposed by the first experimental psychologist, Wilhelm Wundt, and later taken up by James Gibson who coined the term "figure assimilation" to denote the effect of verbal labels on the perception of form. These mental manipulations were supposed to provide a greater stability and precision to the perceived form, rendering it "as good as possible" conforming to the "Law of Prägnanz," as the Gestalt psychologists had claimed.

Text box 20.3 describes a classroom demonstration of labelling effects, which has been adapted from the described experiment by Carmichael et al. (1932). This

TEXT BOX 20.3 CLASSROOM EXPERIMENT DEMONSTRATING THE LABELLING EFFECT

As a classroom demonstration of labelling effects, this text box describes an adapted version of the classical experiment by Carmichael et al. (1932) on the effects of labelling on the reproduction of ambiguous line drawings.

Method

Materials and participants

The complete material is given in Table 20.1. It consists of a series of 12 ambiguous line drawings (as used in the original experiment), together with two lists of labels that serve to influence participants' memory. Although not reported in the original study, the observed effect size was apparently rather large, suggesting that a total sample size of $N = 30$ would suffice (χ^2-tests with $\alpha = 0.05$ and $\beta = 0.20$).

Design and analysis

The participants are randomly assigned to one of three groups (two experimental groups and one control group). One experimental group receives the labels of List 1, the other one those of List 2, and the control group no labels at all. If the number of available participants is limited, you may drop the control

group. If more than enough participants are available (i.e., at least 50), you may introduce a further manipulation and split the two experimental groups. You could introduce the label to half of the participants before presenting the stimulus (i.e., in the learning phase) and to the other half after presenting it (i.e., in the recall phase), thus investigating how strongly the label affects the figure's encoding vs. reconstruction processes (cf. Hanawalt & Demarest, 1939). The dependent measures are the frequencies of reproduced figures classified into four categories: (1) reproductions that resemble the original drawing, (2) those that resemble Label 1, (3) those that resemble Label 2, and (4) those that can't be assigned to any of the three preceding categories. Classification should be done by two (or more) independent judges who are blind to the individual experimental condition.

Procedure

The experiment is run in two sessions. In the first session, participants are instructed that they will see 12 line drawings that they should try to memorize in order to later recall them. Then all stimuli together with either Label 1, Label 2, or no label are presented, one at a time, for about 5 sec each. Each label is given immediately before showing the corresponding figure. After the presentation phase, a retention interval of at least one day (better one week) should follow because the labelling effect generally increases with increasing retention interval. In the second session, participants are asked to reproduce all figures. If the time of presenting the labels is varied, then only half of the participants in the two experimental groups receive the labels at the encoding phase (Session 1). The other half receives the label at the recall phase (Session 2), prompting the participants separately for each stimulus to reproduce "The figure that resembled *x*" (with *x* being replaced by the appropriate label from List 1 or 2).

Results

In line with the classical results of Carmichael et al. (1932), the number of reproductions resembling the labels of List 1 should be larger in the List-1 group than in the List-2 group, and vice versa for reproductions resembling the labels of List 2 (χ^2-tests with $df = 1$). In addition, the number of correct reproductions (i.e., those that were classified as resembling the original drawing) should be larger in the control group than in each of the experimental groups (again, χ^2-tests with $df = 1$). If the time of presenting the labels is varied, fewer correct recalls and more reproductions in line with the given labels should be observed when labels were presented at the recall phase than when they were presented at the encoding phase (cf. Daniel, 1972; Hanawalt & Demarest, 1939).

classroom demonstration is a simplified version and should be fairly easy to set up and to analyze.

Following up on Carmichael et al. (1932), Hanawalt and Demarest (1939) questioned whether the effect of verbal suggestions on the reproductions of visual forms was grounded in a change of the corresponding memory trace or not. They assumed that the labels did not change the memory representation of the perceived forms, but rather, that the labels were used as an aid in reconstructing forgotten parts of the figure. In the spirit of Frederick Bartlett's theory on remembering, Hanawalt and Demarest emphasized that "reproduction is a construction and hence only partially dependent upon the trace." (p. 160) In their experiment, the authors used the same materials, but a quite different procedure. First, all figures were shown without any label. Then, the sample was divided into three groups, receiving Label 1, Label 2, or no label (control). Participants in the two experimental groups were instructed, separately for each of the previously seen figures, to "Draw the figure which resembled *x*" (with *x* being replaced by the appropriate label). As a further between-subjects manipulation, the retention interval was either zero (immediately), two days, or seven days. In the two experimental groups, the percentage of drawings that resembled the suggested label increased significantly with the retention interval. Thus, the data clearly revealed that the suggested labels played an increasingly important role in helping to reconstruct the figure. In accordance with the authors' hypothesis, the results are best understood in assuming that the labels did not change the memory trace of the original figure, but rather were effective only in the reproduction phase.

Prentice (1954) also used the Carmichael et al. (1932) drawings and claimed that labels "operate on S's response in the testing phase (. . .) rather than on the process of memory that makes such reproductions possible." (p. 315) He based his conclusion on the finding that a recognition test that he applied was unaffected by the given labels.

Daniel (1972), however, found that labelling did affect recognition. He used different materials, though – namely, four series of 12 different shapes. In a later recognition test (immediately after exposure, or after a delay of 5 min, 20 min, or 2 days), the "recognized" stimulus shifted away from the original stimulus towards the object denoted by the given label. This effect increased with increasing retention interval. As this occurred in a recognition test (and not only in a reproduction test as before), the author concluded that "the introduction of the label very likely influenced Ss' initial encoding of the form" (p. 156). Daniel explained the difference of his findings to those of Prentice (1954) with differences in the recognition tests applied. While Prentice presented all test stimuli simultaneously on a large sheet, so that participants had ample time to compare all of them, Daniel presented each stimulus alone requiring a rating within a relatively short response interval so that recognition was more difficult than in the Prentice study.

In a further experiment, Nagae (1980) used relevant, irrelevant, or no labels affixed to abstract visual shapes, and manipulated in addition the retention interval. He further assumed that labels could exert two influences, namely a discriminating function at encoding and a categorizing function at retrieval. The main findings

were as follows: Both types of labels, relevant and irrelevant, improved performance in recognizing the original shapes, compared to the no-label control condition. However, both labels also led to larger percentages of false alarms (on highly similar distractor shapes) for longer retention intervals. Thus, labels could be seen as having a positive discriminating function in the short run, but also a distorting categorizing function that is most effective in the long run.

Feist and Gentner (2007) used line drawings of simple visual scenes of static objects (e.g., a marionette standing on a table, or a coin in the palm of a hand). For each scene, three slightly differing drawings were produced: One served as the standard; one was closer to what was described in an accompanying sentence; and one was further away. In the study phase, only the standards were shown either with a describing sentence (experimental) or not (control). In the test phase, all three variants were shown for each scene and participants had to select the originally seen one. The results showed that participants selected the (wrong) line drawing that better fit the sentence more often in the experimental than in the control condition, thus showing the detrimental effect of labels.

Lupyan (2008) investigated the effects of verbal labels on household objects. As materials, he selected pictures of chairs and lamps (all taken from the IKEA online catalogue). All objects were arranged in pairs of highly similar looking objects, so that one object in each pair served as target and the other as lure. In the study session, participants responded to each of the targets either by classifying it (as chair or lamp) or by judging their preference (like vs. don't like). The category names thus served as labels in the classification condition, but were not explicitly used in the preference condition. In the test phase, participants saw all pictures (targets and lures) and had to indicate for each whether it was old (i.e., already seen in the study phase) or new. The results showed that the proportion of pictures correctly judged as old was significantly lower in the classification condition than in the preference condition. In other words, the category label apparently distorted memory for the originally seen objects.

Memory for colour

Another group of researchers looked at the effects of labelling on the memory for colour. For example, Thomas and DeCapito (1966) found that an ambiguous blue-green colour (with 490 nm) that was called either "blue" or "green" by individual participants led to a differential generalization to bluer or greener colours, respectively.

Following this experimental idea, Bornstein (1976) first determined the point of subjective equality (PSE) for calling a colour "blue" or "green" for each of his participants. After one day, the PSE stimuli were presented again in a recognition test, in randomized order together with 8 distractors (with 4 items having shorter and 4 having longer wave lengths than the PSE stimulus). The recognition test was then repeated after one week. Both tests were highly similar and differed only in the label applied to the target stimulus. The instruction stated that a "bluish" (or

"greenish") standard stimulus will be presented for 30 sec and that this standard has to be recognized in a subsequent test with different stimuli (by responding "same" or "different"). The standard was always the individually devised PSE stimulus. The results showed a strong labelling effect. The distribution of mean probabilities of saying "same colour" moved towards the blue colours, when the label was "bluish," and towards the green colours, when the label was "greenish." Thus, the results confirmed that "memory for ambiguous hues can indeed come under the control of linguistic labels." (p. 275)

Administering a different method, namely a post-event misinformation design (see Chapter 22), Loftus (1977) let her participants first view a series of colour slides depicting an auto-pedestrian accident. One slide involved a green car that passed by the scene. Immediately after the end of the series, participants were asked to answer several questions with respect to the slides they saw. For half of the participants, one question was, "Did the blue car that drove past the accident have a ski rack on the roof?", thus introducing the misinformation "blue." The other half (serving as controls) received the same question without any colour term. After a 20 min filler task, a colour recognition test was applied to both groups. It consisted of 30 simultaneously presented colour strips covering the whole range of visible colours. Participants were asked to identify the colour for 10 objects that were depicted in the slides seen before. The critical item was the (originally green) car that passed by the accident. The percentage of correct colour choices was significantly lower in the experimental group. Relative to the controls, the participants with the "blue" misinformation selected significantly more blue and fewer green colours. Loftus interpreted these results as showing the blending of the original trace with the subsequently presented false information (see also Belli, 1988; Pohl & Gawlik, 1995, for more evidence on blends in memory as well as a critical discussion of theoretical conclusions).

In another misinformation experiment on colour memory, Braun and Loftus (1998) investigated the role of wrong advertisement information on a previously experienced product encounter. In the experiment, participants viewed a green wrapper around a chocolate bar in an alleged taste study. After a 15 min filler task, they either saw a blue wrapper around the chocolate bar (visual misinformation) or were told that it was blue (verbal misinformation) or received no such information (control). After another 15 min distractor task, the same recognition test as in the Loftus (1977) study was administered. Participants were asked to pick the correct colours from 30 colour strips for 10 previously seen objects, among them the wrapper of the chocolate bar as the critical item. Again, whereas 73% of control participants chose the correct (green) colour, only 34% and 32% did so in the visually and verbally misled groups, respectively. In addition, misled participants chose more often than controls blue or bluish-green hues, thus demonstrating the biasing influence of the misinformation. In accordance with Belli (1988), but unlike in the Loftus (1977) study, replacements (selecting blue instead of green) were observed more often than blends (selecting a mixture of blue and green), thus suggesting that original and misinformation had coexisted in memory.

Memory for speed

Estimating (or memorizing) speed represents another domain with stimuli that are difficult to verbalize and may thus be easily susceptible to labelling effects. In a now-famous study, Loftus and Palmer (1974) presented short films with traffic accidents. Following each film, the spectators were asked to write a short account of the accident and then to fill in a questionnaire. The critical question had five versions (between-subjects): "About how fast were the cars going when they contacted (hit, bumped, collided with, smashed into) each other?" The authors assumed that the specific verb used in the question would act as a label and thus influence the subsequent speed estimate. The data confirmed this expectation. The mean speed estimates significantly increased from 31.8 mph, when the verb was "contacted," to 40.8 mph, when it was "smashed into". In a second, highly similar experiment with only one film, only the labels "hit" and "smashed" were used and again led to significantly different speed estimates. After a retention interval of one week, the used verbs also led to different percentages of false memories that there had been broken glass in the film (when in fact there had been none).

Loftus and Palmer (1974) assumed that two kinds of information, namely, the perception of the original event and external information supplied after the event, may be integrated over time. As a consequence, it would be difficult (if not impossible) to disentangle the different sources of information: "All we have is one memory." (p. 588) In other words, the use of labels apparently changed the memory representation of the original event. That position, however, was a matter of heated debate in the 80s and 90s. The opponents argued that labels did not change memory, but rather only biased the retrieval from memory (cf. the arguments of Prentice, 1954, for memory of visual form; presented above). Nowadays, it appears that both views peacefully coexist and accept each other's position, acknowledging that both processes may add to the observed labelling effects (see the discussion in Chapter 22 and the SARA model in Chapter 23).

More critique on the Loftus and Palmer (1974) results, however, accrued when two attempts to replicate them failed (McAllister, Bregman, & Lipscomb, 1988; Read, Barnsley, Akers, & Whishaw, 1978). Presumably, labelling is not as robust as supposed, or perhaps its effects depend on the complexity of the material to which the labels are affixed. The other studies (referred to above and below) all used rather simple materials (line drawings, colour, taste, or odour), while the films with accident scenes were rather complex and were, moreover, followed by a lengthy questionnaire. This complexity may have elicited other cognitive processes that could have obscured or prevented any labelling effect (cf. the problems of replicating "verbal overshadowing" below).

Judgments of taste and odour

A few studies investigated the effects of labels on taste judgments. Although these studies did not directly address the question of how labelling influences memory,

it appears fairly safe to conclude that the observed judgment distortions also generalize to corresponding memory distortions (cf. Chapter 23). Besides, there is at least one explicit demonstration of labelling effects on taste memory (Melcher & Schooler, 1996).

In a study on food preferences (Woolfolk et al., 1983), students tasted two allegedly different brands of cola from different cups marked either "L" or "S." Earlier studies had found the letter "L" (for peculiar reasons) to be more disliked than the letter "S." Consequently, most participants judged the cola in cup "S" to taste better than the one in cup "L," although both cups contained the identical beverage. In a further study by the same authors, students first stated whether they preferred Pepsi or Coke. Then they received both a Coke and a Pepsi bottle and were asked to verify their stated preference by tasting both brands. The trick, however, was that each bottle contained exactly the other brand than was stated on the bottle's label. The most interesting question accordingly was whether participants would base their judgment on their taste or on the label? The results revealed the latter to be true. Most participants continued to prefer their favorite brand as given by the label of the bottle and not as given by its true contents.

In another study demonstrating labelling effects (Wardle & Solomons, 1994), participants first tasted cheese-spread sandwiches and strawberry yoghurt that were either labeled as "low-fat" or as "full-fat," and then rated the food's tastiness. The results showed that food labeled as "low-fat" was rated lower in liking than food labeled as "full-fat," although both foods had in fact been identical.

In a further attempt to investigate the influence of labelling on taste judgments, Pohl et al. (2003) asked students to taste a white wine and to estimate its sweetness. In advance, however, half of the participants were informed that the wine was "sweet" and the other half that it was "dry". Subsequent estimates of the sugar amount revealed clear effects of the given label: The allegedly "sweet" wine was judged to contain more residual sugar than the allegedly "dry" wine.

Labelling effects were also reported for odour perception. Herz and von Clef (2001) presented a set of five different odours repeatedly in two sessions (one week apart). Each odour was accompanied by one of two labels (in Sessions 1 and 2, respectively) suggesting either a pleasant or an unpleasant hedonic impression. It is noteworthy that, in contrast to other labelling studies, the hedonic-value suggestion was manipulated as a *within-subjects* factor, thus allowing a strong test of the labelling effect. The five odours (violet leaf, patchouli, pine oil, menthol, and a mixture of isovaleric and butyric acid) were selected as materials because they had been perceived as rather ambiguous in their hedonic value in pilot studies. The labels had also been collected in a pilot study asking participants to describe the odours. As a result, each odour was combined with a positive or a negative label in Session 1 or 2 (e.g., pine oil was either labeled as "Christmas tree," i.e., positive, or as "spray disinfectant," i.e., negative). These labels were given immediately before each odour was sniffed. The results revealed clear effects of the given labels. If an odour was suggested to be pleasant (positive), it received much higher hedonic ratings than

when the same odour was suggested to be unpleasant (negative). Recall that these ratings were given by the same participants (and not different ones), thus showing indeed strong labelling effects.

Verbal overshadowing

While, for the examples presented so far, the label was always provided by the experimenter, labelling effects could also occur if the labels were self-generated by the participants. In these cases, however, the label does typically not consist of a single word, but rather entails a more elaborated verbalization. In an intensely studied paradigm, self-generated verbal descriptions of difficult-to-describe stimuli (like faces or voices) were found to deteriorate subsequent memory performance ("verbal overshadowing"). Following the now classical study of Schooler and Engstler-Schooler (1990), described below, the processes leading to verbal overshadowing have received much attention (see Schooler, Fiore, & Brandimonte, 1997, for a review). Negative effects of verbalization on memory have meanwhile been shown for shapes, colours, faces, voices, music, taste, spatial maps, and problem solving. Two excellent collections of empirical and theoretical papers representing the respective state of the art were edited as special issues of *Applied Cognitive Psychology* (Memon & Meissner, 2002) and the *European Journal of Cognitive Psychology* (Lloyd-Jones, Brandimonte, & Bäuml, 2008). Recently, the original study by Schooler and Engstler-Schooler (1990) has been replicated in an international replication project, as will be described further below.

Verbal overshadowing on visual objects

Introducing the concept of "verbal overshadowing," Schooler and Engstler-Schooler (1990) studied the negative effects of verbalization on visual memories (faces and colours). In one experiment, participants watched a video depicting a salient person (a bank robber). After a 20-min delay, participants either gave a verbal description of the robber's face (experimental) or did an unrelated filler task (control). In a subsequent recognition test, participants had to identify the robber's face from a lineup of eight faces. The results showed that recognition accuracy was substantially impaired in the verbalization group compared to the control group.

In another experiment, Schooler and Engstler-Schooler (1990) showed one of three different colour swatches (red, blue, or green) to their participants. Immediately after that, participants either wrote descriptions of the seen colour (verbal), visualized it (visual), or did an unrelated task (control) for 30 sec. In a subsequent recognition test, the target colour was shown together with five distractors containing similar colour hues. The whole procedure was then repeated for the other two colour swatches. The results showed that, compared to the control condition, recognition for the original colour was impaired when it had been verbally described, whereas visualizing the colour had no such effect. Interpreting the results of this and other experiments, the authors assumed that "detrimental memory recoding

results from the mismatch between the original visual stimulus and the subsequent verbalization." (p. 52) In this case, verbalization of the given hue (or face) could have been rather difficult, so that the resulting labels apparently deviated from the original stimulus.

In order to test the mediating roles of perceptual and verbal skills, Melcher and Schooler (2004) used a relatively unknown knowledge domain, namely mushrooms. Participants in their studies either received a perceptual training (by categorizing photos of mushrooms), a conceptual training (by learning characteristics of mushrooms), or no training. Their hypothesis was that verbal overshadowing depended on a situation where perceptual skills exceeded verbal ones (cf. Melcher & Schooler's, 1996, results on wine expertise described below), so that verbalization would be more detrimental. Their studies followed the classical design: Pictures of mushrooms had to be verbally described (experimental) or not (control). Then, in a recognition test, participants had to recognize the original mushroom (on a different photo) from a number of foils. Results showed a clear verbal-overshadowing effect after perceptual training, an almost significant (but not predicted) effect after conceptual training, and no effect after no training. The authors suggested that the conceptual training was perhaps not purely conceptual, but also included some perceptual training, thus leading to some decrement. In a second experiment, they accordingly changed the contents of training and now observed verbal overshadowing only in the perceptually trained group, whereas the other two groups even showed an enhancement effect of verbalization. The authors thus concluded that "the results of these two studies demonstrate that the effects of verbal overshadowing depend on both the extent and the type of experience with a stimulus domain." (p. 626)

Verbal overshadowing on other senses

Melcher and Schooler (1996) tested the effect of verbalization on taste memory. Their participants tasted two red wines, then verbally described the wines' taste (experimental) or not (control), and were then asked to recognize the wines from a set of four. A verbal-overshadowing effect, that is, a drop in recognition performance after verbalization (compared to the control group), was observed, but only for the first wine tasted and only for participants who were categorized as "intermediate" in wine expertise. There was neither an effect for the second wine, nor for novices and wine experts. The authors concluded that it was the difference between perceptual and verbal experience that made the "intermediates" vulnerable to the verbal-overshadowing effect (cf. Melcher & Schooler, 2004, described above). At the same time, the results suggest that the verbal-overshadowing effect is not very robust.

Other experiments tested verbal overshadowing in the auditory domain, namely, memory for voices. In an experiment by Perfect, Hunt, and Harris (2002), participants heard a voice speaking one sentence, then worked on a 10-min filler task, and then produced a verbal description of the voice (experimental) or not (control). In the immediately following test phase, participants were asked to identify the

previously heard voice from a sample of six voices. The results showed that the percentage of correct identifications was much lower in the experimental than in the control condition, thus showing verbal overshadowing.

Using a highly similar experimental design, Vanags, Carroll, and Perfect (2005) reported two studies on voice recognition. Experiment 1 was planned to replicate the Perfect et al. findings. However, the results showed no verbal overshadowing. The authors saw one reason for their replication failure in the similarity between original and test stimuli, which was larger here than in the earlier study. As a matter of fact, when the Perfect et al. study was re-analyzed, verbal overshadowing was also absent if the two recordings of the critical voice (at encoding and test) were highly similar. The effect only emerged for recordings of the same voice that were sufficiently different (e.g., microphone recording as original stimulus, but telephone recording as test stimulus). Accordingly, the authors decreased the similarity of encoding and test stimuli in their Experiment 2 and found a very strong overshadowing effect: Participants that had given a verbal description of the original voice recognized it much less often than those that had not given such a description. But again, these studies also showed that the appearance of verbal overshadowing may hinge on subtle procedural conditions.

Theoretical accounts of verbal overshadowing

Schooler and Engstler-Schooler (1990) argued that visual memories can be represented in multiple codes, for example, a visual code for the perceptual information and a verbal code for the labelling information (cf. Allan Paivio's dual coding theory; Paivio, 1969, 1971). Generally, most studies have found that an additional (verbal) code (like in elaboration) will be helpful in remembering visual information rather than impeding it. So what makes the difference between positive and negative effects of a verbal label on visual memories? Some researchers have suggested that verbal processing diminishes the resources left to visual processing, thus impairing encoding of the latter. Others claimed that the verbal code interferes with the use of the visual code, thus impairing memory performance.

Schooler and Engstler-Schooler (1990) concluded from their studies that "verbalization does not eradicate the original representation of the memory [. . .] but rather produces some form of interference." (p. 61) The authors therefore preferred the term "verbal overshadowing" to denote the effects of verbal labels on the memory for visual information. In an attempt to generalize their findings, the authors further assumed that verbal overshadowing may occur with any stimulus that defies complete linguistic description. This may apply not only to visual stimuli, but equally well to taste, touch, smell, sound, and other stimuli, as has indeed been found (see above).

Several later studies, however, failed to replicate the overshadowing effect, leading Schooler et al. (1997) to admit that the effect is somewhat fragile. Accordingly, the meta-analysis of Meissner and Brigham (2001) revealed only a small, but significant detrimental effect of verbalization, but a much larger one of the particular type of instruction and procedure used to elicit verbalization.

Schooler (2002) proposed that "verbalization produces a 'transfer inappropriate processing shift' whereby the cognitive operations engaged in during verbalization dampen the activation of brain regions associated with critical non-verbal operations." (p. 989) The assumed shift could result from limited-capacity competitions (a) between right and left hemisphere processes and (b) between automatic and controlled processes. The right hemisphere is more associated with non-verbal configural processes, while the left one is more associated with language-based processes. Thus, the act of verbalization could shift the center of control from right to left hemisphere regions. Similarly, automatic ("reflexive") processes are associated with other brain regions than controlled ("reflective") processes are. Verbalization then could lead to a shift from more reflexive to more reflective processes.

In a further meta-analysis, Meissner, Sporer, and Susa (2008) identified several moderators of verbal overshadowing (like single vs. multiple face recognition, method of eliciting verbal descriptions, and retention interval). On a theoretical level, they summarized their findings as showing "that retrieval-based processing, transfer inappropriate processing, and levels of processing frameworks can account for a variety of conditions that lead to verbal overshadowing versus verbal facilitation." (p. 422) Accordingly, they admitted that multiple mechanisms may be responsible. Meissner et al. specifically investigated the relationship between the accuracy of the generated verbal descriptions of a face and the accuracy of later recognizing that face (which should be positive). They found a significant, but only small positive correlation, which thus explains only little of the variance in verbal overshadowing. In other words, a number of further factors must also play a role.

Lloyd-Jones et al. (2008) proposed that verbal processes might influence all stages, namely encoding, storage, and retrieval of visual information. Accordingly, they came to the somewhat pessimistic conclusion that "different forms of verbal interference and facilitation are likely to be due to different mechanisms in different contexts. The state of the art is that we are just beginning to understand the rich complexity of the problem." (p. 393)

Replication project

In an attempt to test the robustness of the originally reported overshadowing effect (Schooler & Engstler-Schooler, 1990), the editors of *Perspectives on Psychological Science* initiated a large-scale replication project, involving in the end 31 international labs (Alogna et al., 2014). The studies involved two different procedures following the study phase (and preceding the test phase): In Design 1, verbalization (experimental) or not (control) preceded the filler task, whereas in Design 2, the order was reversed. The results showed a small effect of verbalization in Design 1, but a larger one in Design 2. Both are, however, smaller than in the original study. Nevertheless, Alogna et al. concluded that "these findings reveal a robust verbal overshadowing effect that is strongly influenced by the relative timing of the tasks." (p. 556)

Conclusions

Generally, labelling could be considered a type of suggestion (cf. Chapter 25). In typical labelling experiments, persons either receive some sort of ambiguous information or information that is difficult (if not impossible) to verbalize or they are simply left in a state of uncertainty. The label that is then provided contains some specific information that allows reducing the subjective ambiguity or uncertainty. However, by doing so the label acts like a suggestion and drives reinterpretation or reconstruction into a specific direction. In other words, the label gives meaning to an otherwise more or less obscure situation and thus reduces or even eliminates ambiguity or uncertainty. Such effects of labelling have been observed in a variety of domains and contexts and with a variety of experimental procedures. This would lead one to conclude that labelling is a widespread and robust phenomenon. However, several of the studies reviewed above delineated the possible limits of this effect. In more complex situations, labelling might be overridden by other information that is available and then can be used to reduce uncertainty and thus improve memory. Most of the experiments demonstrating labelling effects, however, used rather simple materials (like line drawings or colours), so that encoding or reconstruction processes more heavily relied on the scanty but relatively "helpful" information provided by the label.

Summary

- A label is something that is affixed to a stimulus. A labelling effect occurs if judging or recalling that stimulus is systematically influenced by and in the direction of the label.
- Labelling effects have been found for the recall of visual forms and colour, and for the judgment of speed, taste, and odour.
- Labelling effects may depend on the complexity of the used material. The more complex the material, the more fragile the effect.
- Under the name of "verbal overshadowing," verbal self-descriptions of visual stimuli (faces, colours, and other stimuli) were found to deteriorate memory for these stimuli.
- The label may apparently exert its influence during encoding as well as during recall.
- In general, labelling could be understood as a suggestive process.

Further reading

An enlightening exercise in classical experimenting is to read the original labelling study by Carmichael et al. (1932). As examples of less well-researched labelling domains, the studies by Pohl et al. (2003) on taste and by Herz and von Clef (2001) on odour judgment are noteworthy. With respect to "verbal overshadowing," Schooler et al. (1997) presented an early review of the field,

and Meissner and Brigham (2001) a meta-analysis. Schooler (2002) discussed in detail several theoretical approaches. Meissner et al. (2008) presented a very thorough meta-analysis. Most recently, the replication project yielded valuable results (Alogna et al., 2014).

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., Birt, A. R., . . ., Zwaan, R. A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science, 9*, 556–578.

Belli, R. F. (1988). Color blend retrievals: Compromise memories or deliberate compromise responses? *Memory & Cognition*, *16*, 314–326.

Bornstein, M. H. (1976). Name codes and color memory. *American Journal of Psychology*, *89*, 269–279.

Braun, K. A., & Loftus, E. F. (1998). Advertising's misinformation effect. *Applied Cognitive Psychology*, *12*, 569–591.

Carmichael, L., Hogan, H. P., & Walters, A. A. (1932). An experimental study of the effect of language on the reproduction of visually perceived form. *Journal of Experimental Psychology*, *15*, 73–86.

Daniel, T. C. (1972). Nature of the effect of verbal labels on recognition memory for form. *Journal of Experimental Psychology*, *96*, 152–157.

Feist, M. I., & Gentner, D. (2007). Spatial language influences memory for spatial scenes. *Memory & Cognition, 35*(2), 283–296.

Hanawalt, N. G., & Demarest, I. H. (1939). The effect of verbal suggestion in the recall period upon the reproduction of visually perceived forms. *Journal of Experimental Psychology*, *25*, 159–174.

Herz, R. S., & von Clef, J. (2001). The influence of verbal labeling on the perception of odors: Evidence for olfactory illusions? *Perception*, *30*, 381–391.

Higgins, E. T., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, *13*, 141–154.

Katz, R. C., Cacciapaglia, H., & Cabral, K. (2000). Labeling bias and attitudes toward behavior modification revisited. *Journal of Behavior Therapy and Experimental Psychiatry*, *31*, 67–72.

Link, B. G., & Phelan, J. C. (2001). Conceptualizing stigma. *Annual Review of Sociology*, *27*, 363–385.

Lloyd-Jones, T. J., Brandimonte, M. A., & Bäuml, K.-H. (Eds.). (2008). Verbalising visual memories [Special issue]. *European Journal of Cognitive Psychology, 20*(3).

Loftus, E. F. (1977). Shifting human color memory. *Memory & Cognition*, *5*, 696–699.

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.

Lupyan, G. (2008). From chair to "chair": A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General, 137*(2), 348-369.

McAllister, H. A., Bregman, N. J., & Lipscomb, T. J. (1988). Speed estimates by eyewitnesses and earwitnesses: How vulnerable to postevent information? *Journal of General Psychology*, *73*, 25–36.

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology*, *15*, 603–616.

Meissner, C. A., Sporer, S. L., & Susa, K. J. (2008). A theoretical review and meta-analysis of the description-identification relationship in memory for faces. *European Journal of Cognitive Psychology, 20*(3), 414–455.

Melcher, J. M., & Schooler, J. W. (1996). The misremembrance of wines past: Verbal and perceptual expertise differentially mediate verbal overshadowing of taste memory. *Journal of Memory and Language*, *35*, 231–245.

Melcher, J. M., & Schooler, J. W. (2004). Perceptual and conceptual training mediate the verbal overshadowing effect in an unfamiliar domain. *Memory & Cognition, 32*(4), 618–631.

Memon, A., & Meissner, C. A. (Eds.). (2002). Investigations of the effects of verbalization on memory [Special issue]. *Applied Cognitive Psychology*, *16*(8).

Nagae, S. (1980). Nature of discriminating and categorizing functions of verbal labels on recognition memory for shape. *Journal of Experimental Psychology: Human Learning and Memory, 6*(4), 421–429.

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review, 76*, 241–263.

Paivio, A. (1971). *Imagery and verbal processes.* New York: Holt, Rinehart, & Winston.

Perfect, T. J., Hunt, L. J., & Harris, C. M. (2002). Verbal overshadowing in voice recognition. *Applied Cognitive Psychology, 16*, 973–980.

Pohl, R. F., & Gawlik, B. (1995). Hindsight bias and misinformation effect: Separating blended recollections from other recollection types. *Memory*, *3*, 21–55.

Pohl, R. F., Schwarz, S., Sczesny, S., & Stahlberg, D. (2003). Hindsight bias in gustatory judgments. *Experimental Psychology*, *50*, 107–115.

Prentice, W. C. H. (1954). Visual recognition of verbally labeled figures. *American Journal of Psychology*, *67*, 315–320.

Read, J. D., Barnsley, R. H., Akers, K., & Whishaw, I. Q. (1978). Variations in severity of verbs and eyewitness' testimony: An alternate interpretation. *Perceptual and Motor Skills*, *46*, 795–800.

Schooler, J. W. (2002). Verbalization produces a transfer inappropriate processing shift. *Applied Cognitive Psychology*, *16*, 989–997.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*, 36–71.

Schooler, J. W., Fiore, S. M., & Brandimonte, M. A. (1997). At a loss from words: Verbal overshadowing of perceptual memories. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 37, pp. 293–334). San Diego, CA: Academic Press.

Sloutsky, V. M., Lo, Y. F., & Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels, similarity, and the development of inductive inference. *Child Development*, *72*, 1695–1709.

Thomas, D. R., & DeCapito, A. (1966). Role of stimulus labeling in stimulus generalization. *Journal of Experimental Psychology*, *71*, 913–915.

Vanags, T., Carroll, M., & Perfect, T. J. (2005). Verbal overshadowing: A sound theory in voice recognition? *Applied Cognitive Psychology, 19*, 1127–1144.

Wardle, J., & Solomons, W. (1994). Naughty but nice: A laboratory study of health information and food preferences in a community sample. *Health Psychology*, *13*, 180–183.

Welder, A. N., & Graham, S. A. (2001). The influences of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. *Child Development*, *72*, 1653–1673.

Woolfolk, A., Woolfolk, R., & Wilson, T. (1977). A rose by any other name: Labeling bias and attitudes toward behavior modification. *Journal of Consulting and Clinical Psychology*, *45*, 184–191.

Woolfolk, M. E., Castellan, W., & Brooks, C. I. (1983). Pepsi versus Coke: Labels, not tastes, prevail. *Psychological Reports*, *52*, 185–186.

21

ASSOCIATIVE MEMORY ILLUSIONS

Henry L. Roediger, III and David A. Gallo

Distortions of memory arise from many causes. Several types of memory illusions reviewed in this volume are created from external sources. In recollecting some target event from the past, people will often confuse events that happened before or after the target event with the event itself. (These confusions are examples of proactive and retroactive interference, respectively.) On the other hand, the illusion described in this chapter involves the remembering of events that never actually occurred. This erroneous information is internally created by processes that would otherwise lead to good memory for actual events. As such, these errors are part and parcel of the natural memory process, and they are extremely difficult to avoid. Although most of the research reviewed here involves a tightly controlled laboratory paradigm using word-lists, we believe (and will cite evidence to support) the claim that similar processes occur whenever people try to comprehend the world around them – reading a newspaper or novel, watching television, or even perceiving scenes with little verbal encoding at all (Roediger & McDermott, 2000b).

The associative tradition

From Aristotle to computational models, scholars have always assumed that the mind is fundamentally associative in nature (see Roediger, McDermott, & Robinson, 1998, see also SARA in Chapter 23 of this volume). In many theories, associations are viewed as a powerful positive force to support remembering – the stronger the associative bond between two elements, the more probable is retrieval of the second element when given the first as a cue. The idea that associative connections might have a dark side – that they may lead to errors of memory – has hardly ever been considered. However, the point of this chapter is that memory distortions can indeed be induced by associative means.

As far as we know, this idea was first suggested, quite offhandedly, in a paper by Kirkpatrick (1894). He was interested in whether items presented as visual objects were better retained than those presented as words, but his side observations are of interest for present purposes and worth quoting (p. 608):

> About a week previously in experimenting upon mental imagery I had pronounced to the students ten common words . . . it appears that when such words as "spool," "thimble" and "knife" were pronounced many students at once thought of "thread," "needle," and "fork," which are so frequently associated with them. The result was that many gave those words as belonging to the list. This is an excellent illustration of how things suggested to a person by an experience may be honestly reported by him as part of the experience.

The process described by Kirkpatrick is the topic of this chapter, how items associated to presented items are often actually remembered as having been overtly presented (rather than inferred covertly). Underwood (1965) did relevant research, but his effect was quite small, and the technique presented in the next section produces much more robust findings. Indeed, false recognition can sometimes be more likely than true recognition when elicited by this newer technique. A simplified version of such an experiment that can be used as a classroom demonstration is described in Text box 21.1 and discussed in the next section.

TEXT BOX 21.1 CLASSROOM DEMONSTRATION

This demonstration can be used to create false memories in only a few minutes. For best results, participants should not be told that the demonstration is on false memories until after the experiment. We will suggest two variations on this theme after this demonstration so that you may compare two other interesting conditions.

Method

Material

The material consists of 4 lists with 15 words each that are all associated to a critical, but not included target word.

List 1: *bed, rest, awake, tired, dream, wake, snooze, blanket, doze, slumber, snore, nap, peace, yawn, drowsy.*
List 2: *door, glass, pane, shade, ledge, sill, house, open, curtain, frame, view, breeze, sash, screen, shutter.*
List 3: *nurse, sick, lawyer, medicine, health, hospital, dentist, physician, ill, patient, office, stethoscope, surgeon, clinic, cure.*

List 4: *sour, candy, sugar, bitter, good, taste, tooth, nice, honey, soda, chocolate, heart, cake, tart, pie.*

The critical target words are *sleep*, *window*, *doctor*, and *sweet*, respectively.

Procedure

The experimenter tells the participants that this will be a memory demonstration, and that they should have a piece of scrap paper and a pen ready for the memory test. The experimenter then tells them that s/he will read lists of words, and that they should try to remember these words. The participants should not be allowed to write the words down as they are being read. The experimenter then reads the first list, at a steady rate of one word every 1 to 2 seconds. After the final word, the participants are asked to write down as many words as they can remember, in any order, without guessing. Participants usually take less than a minute to recall each list. This procedure is then repeated for the next three lists.

Analysis

After the final list is recalled, the experimenter counts separately for each list the number of participants (by having them raise their hands or by tallying the recall sheets) who recalled the critical word. As these critical associates were never presented, their recall represents false memories.

Variations

The demonstration above is quite simple. Here are two variations on the same idea. Using the same condition described above as the control condition, test another group of participants but explain the phenomenon to them *before* they are presented with the lists. That is, tell them that they are to be given a list of words that is intended to make them think of another word and to recall that word in the list even if they are not supposed to. You can also present a couple of lists like the ones above and tell subjects how they are constructed. You can find many lists to use in a paper by Stadler, Roediger, and McDermott (1999). Keep the other instructions the same as in the basic condition, with the warning against guessing. Several experiments have been done using such an instruction, and the general finding is that subjects can reduce the level of false recall in the experiment, but they cannot eliminate the effect. It is usually about half as great as in the basic condition with the general warning against guessing. The reduction in experiments measuring false recognition is even smaller (see Gallo, Roediger, & McDermott, 2001).

Sample experiment: The DRM paradigm

Roediger and McDermott (1995) adapted a paradigm first used by Deese (1959) for a somewhat different purpose. The paradigm – the one described in Text box 21.1 – produces a very strong associative memory illusion and (owing to a suggestion by Endel Tulving) is now called the DRM paradigm (for Deese-Roediger-McDermott). The paradigm and variants of it are frequently used as a straightforward technique for gaining measures of both veridical and false memories using both recall and recognition techniques. We describe here a somewhat simplified version of Experiment 2 in Roediger and McDermott (1995).

Typical method

A set of 24 associative lists were developed, each list being the 15 strongest associates to a nonstudied word, as found in word association norms. These norms are based on a free association task, in which participants were presented a stimulus word (e.g., *rough*) and told to generate the first word that comes to mind. To create each of our lists, we took the 15 words that had been elicited most often by the stimulus word (e.g., the words *smooth, bumpy, road, tough, sandpaper, jagged, ready, coarse, uneven, riders, rugged, sand, boards, ground,* and *gravel*; see Stadler et al., 1999, for a set of 36 lists with normative false recall and recognition data). These study lists were presented to new participants, and their memory was subsequently tested. Critically, the stimulus word (*rough* in this case) was never studied by these participants. Our interest centered on the possible false recall or false recognition of this critical word. If a participant were like a computer or tape recorder, recording and retrieving the words perfectly, one would not expect such systematic memory errors.

Thirty undergraduate participants heard 16 of the 24 word lists, one word at a time. Participants recalled 8 of the lists immediately after their presentation (with 2 minutes provided for recall), and they were instructed not to guess but only to recall the words they were reasonably sure had been in the list. They performed arithmetic problems for two minutes after each of the other 8 lists. Shortly after all 16 of the lists had been presented in this way, participants took a yes/no recognition test that covered all 24 lists. Because only 16 lists had been studied, items from the other 8 lists served as lures or distractors to which participants should respond "no" (it was not on the list). The recognition test was composed of 96 words, with 48 having been studied and 48 new words. Importantly, 16 of these new words were the critical missing lures (words like *rough*) that were strongly associated to the studied words.

After each word that they recognized as having been in the list (the ones judged "yes"), participants made a second judgment. They were asked to judge whether they remembered the moment of occurrence of the word in the list, say by being able to remember the word before or after it, what they were thinking when they heard the word, or some other specific detail. This is called a *remember* judgment and is thought to reflect the human ability to mentally travel back in time and re-experience events cognitively. If they were sure the word had been in the list, perhaps because it was highly familiar, but could not remember its specific moment

of occurrence, they were told to make a *know* judgment. The remember/know procedure was developed by Endel Tulving in 1985, and has since been used by many researchers in order to measure the phenomenal basis of recognition judgments (see Gardiner & Richardson-Klavehn, 2000, for an overview).

Typical results

Let us consider the immediate free recall results first. Recall of list items followed the typical serial position curve, with marked primacy and recency effects reflecting good recall at the beginning (primacy) and end (recency) of the list. Consider next false recall of the critical nonpresented word such as *rough* (in our sample list used above). Despite the fact that recall occurred immediately after each list and participants were told not to guess, they still recalled the critical nonpresented item 55% of the time. In this experiment, recall of the critical nonpresented item was actually higher than recall of the items that were presented in the middle of the list. In other studies, the probability of recall of critical items often approximates recall of items in the middle of the list, with the particular outcome depending on such factors as presentation rate of the lists (1.5 sec in this study) and whether the lists are presented auditorily or visually. The important point is that false recall was very high.

The recognition test also revealed a powerful associative memory illusion. The basic data are presented in Figure 21.1. Shown in the two panels are data from the

FIGURE 21.1 The DRM false-recognition effect (Roediger & McDermott, 1995, Exp. 2). False recognition of critical lures approximated the hit rate for list items. False alarms to list words from nonstudied lists were 0.11, and those to critical words from nonstudied lists were 0.16.

8 lists that were studied and recalled (the right panel) and from the 8 lists that were studied but not previously recalled. Within each panel, the left bar shows veridical or true recognition (the hit rate) of items actually studied, whereas the right bar shows false recognition of the critical lures like *rough* (the critical-lure false alarm rate). The false alarm rates to the items from the 8 nonstudied lists that appeared on the test are given in the figure caption. Finally, each bar is divided into a dark portion (items called old and judged to be *remembered*) and a white portion (items called old and judged to be *known).*

Figure 21.1 shows the very large false recognition effect that is typical of the DRM paradigm. For example, for lists that were studied and had been recalled, participants recognized 79% of the list words as old and said they remembered 57% of the words. Nearly three-quarters of the words called old were judged to be remembered (i.e., 57/79 = 72%). Surprisingly, the data for the critical lures (that, again, were not actually presented) were practically the same! Participants recognized 81% as old and even judged 58% *remembered*; and, just like for studied words, 72% of the words judged old were remembered (58/81 = 72%). So, in the DRM paradigm, the level of false recognition and false remembering is about the same as veridical recognition and remembering of list words. The situation is much the same for the lists that were studied but not recalled, with false recognition of critical lures being as high as (or even higher than) veridical recognition of list words. Note, however, that remember judgments were lower on the recognition test (for both kinds of items) when the lists had not been recalled. In some ways, the data in the left panel show effects of recognition that are "purer" in that they were not contaminated by prior false recall. Nonetheless, striking levels of false recognition and "remember" judgments were obtained.

Discussion

Perhaps because the effect is so robust, a skeptical reaction after learning about the effect is disbelief: Participants are obviously not trying to remember at all. Instead, this reasoning continues, participants are faced with too many words to remember, and so make educated guesses as to which words were presented. In particular, they realize that the lists consist of associated items, so they infer that critical items (which are associated to the study lists) were also presented. Miller and Wolford (1999) formalized this sort of decision process in terms of a liberal criterion shift to any test word that is perceived as related to the study list (i.e., the critical items). This model was primarily directed at false recognition, although a generate/recognize component was included to account for false recall. In either case, it is assumed that participants try to capitalize on the related nature of the lists (via some sort of liberal criterion to related items), in the hopes of facilitating their memory for studied words.

Gallo et al. (2001) directly tested this account by informing participants about the illusion and telling them to avoid false recognition of nonstudied but related words. The critical condition was when participants were warned after studying the lists

but before taking the recognition test, thereby precluding a liberal guessing strategy for related items. The results were straightforward: Warning participants between study and test had negligible effects on false recognition (relative to a no-warning control condition), even though other conditions revealed that warned participants were trying to avoid false recognition. This pattern also was obtained in false recall (e.g., Neuschatz, Payne, Lampinen, & Toglia, 2001). Gallo et al. (2001) reasoned that warned participants did not adopt a liberal strategy to related items because, after all, they were trying to avoid false alarms to these items. Thus, robust false memory effects following warnings were not due to such strategic decision processes alone, but instead were due to processes that are inherent in the memory system. This conclusion is bolstered by the finding that participants will often claim that these false memories are subjectively detailed and compelling (as reviewed below).

In sum, the DRM paradigm is one of the most potent memory illusions ever studied. As noted above, similar associatively based errors have been obtained using a wide variety of materials, including pictures, sentences, and stories, although these errors are usually not as frequent as those observed in the DRM paradigm (see Roediger & McDermott, 2000a, 2000b, for an overview). In general, any set of materials that strongly implies the presence of some object or event that is not actually presented lends itself to producing false recall and false recognition of the missing but implied event (cf. Chapter 22 on the misinformation effect). Why, then, are the false memories produced by the DRM paradigm so robust?

There are several answers to this question, but we will concentrate on the most critical one: the number of associated events that are studied. The DRM paradigm, in contrast to most other memory illusions, relies on the presentation of multiple associates to the critical nonstudied word, thereby taking full advantage of the power of associations. Robinson and Roediger (1997) directly examined the effect of number of associated words on false recall and false recognition. In one experiment, they presented participants with lists of 15 words prior to recall and recognition tests, but the number of words associated to a critical missing word was varied to be 3, 6, 9, 12, or 15. Increasing numbers of associated words steadily increased false recall of the critical nonpresented word, from 3% (with 3 words) to 30% (with 15 words). Thus, even though the total number of words studied was the same in all conditions, the number of studied associates to the critical word had a considerable influence on the strength of this memory illusion. We discuss the theoretical implications of this finding in the next section.

Theories and data

In this section we consider the processes that may be involved and how they might interact to give rise to the associative false-recognition effect. This discussion is divided into two main sections: processes that cause the effect and opposing processes that reduce the effect. Our goal is not to exhaustively review all of the DRM findings – that would be well beyond the scope of this chapter. Indeed, Gallo (2006) wrote a book-length review of the first 10 years of DRM research, and 10 years of

additional research have ensued since then (see Huff, Bodnar, & Fawcett, 2015, for a more recent review). Rather, our goal in the present chapter is to highlight the main theoretical issues and discuss those DRM findings that we feel critically inform these issues. In many instances, more than one group of researchers reported relevant findings, but for brevity we cite only one or two findings to illustrate the point.

Processes that cause the effect

The dominant theories of the DRM effect fall into two classes: *association-based* and *similarity-based*. These classes differ in the types of information or representation that is proposed to cause false remembering (and in terms of the processes that allegedly give rise to these representations). Nevertheless, these theories are not mutually exclusive, and evidence suggests that both types of mechanism make a unique contribution to the effect. We discuss each in turn, followed by a brief consideration of attribution processes that may contribute to the subjectively detailed nature of associative false memories.

Association-based theories

According to the association-based theories, a preexisting representation of the critical nonpresented word becomes activated when its associates are presented. Thus, presenting *bed*, *rest*, *awake*, etc. activates the mental representation of the word *sleep*. Under this theory, false remembering occurs when the participant mistakes associative activation with actual presentation, which can be conceptualized as a reality monitoring error (Johnson, Hashtroudi, & Lindsay, 1993). This theory is similar to Underwood's (1965) classic idea of the implicit associative response (IAR), and is consistent with Deese's (1959) finding that the degree of association between the list words and the critical nonpresented word (dubbed Backward Associative Strength, or BAS) was highly predictive of false recall. That is, the more items in the list have the critical item as an associate (which is what BAS measures), the more likely the list is to produce false recollection.

Deese's (1959) finding was replicated and extended by Roediger, Watson, McDermott, and Gallo (2001) who reported that BAS predicted most of the variance in false recall (among several candidate variables) using multiple regression analysis. Roediger et al. (2001) interpreted this relationship as evidence for associative activation. The notion is that associates activate the lexical representation of the critical word, and this activation supports the thought of the item on a recall test. They also found that BAS was related to false recognition, suggesting that activation might be a common cause of false recall and false recognition, although the differences in recognition tend to be somewhat smaller than those found in recall (see Gallo & Roediger, 2002). The aforementioned list-length effect (e.g., Robinson & Roediger, 1997) is also consistent with an associative activation mechanism: Increasing the number of associates studied increases associative activation, and hence increases false recall and recognition.

Two obvious questions concern the form of this activation (conscious or nonconscious?) and when it occurs (study or test?). The fact that false recall and false recognition occur even with very rapid study presentation rates (under 40 ms per item, or less than a second per list) suggests that conscious thoughts of the critical item during study are not necessary to elicit false memory (see McDermott & Watson, 2001, for recall evidence, and Cotel, Gallo, & Seamon, 2008, for recognition evidence). This is consistent with semantic priming models, which suggest that associative activation at study can automatically spread from one word node to another (see Roediger, Balota, & Watson, 2001). However, just because conscious thoughts of the critical item may not be necessary to elicit false remembering does not imply that they do not occur at the relatively slower presentation rates (e.g., 1–2 sec per item) that are typically used in the paradigm. At more standard rates, overt rehearsal protocols indicate that participants often think of the critical item during study, and the frequency of these thoughts predicts subsequent false recall (e.g., Goodwin, Meissner, & Ericsson, 2001).

Additional evidence that associative activation occurs at study has been obtained using implicit tests. After presenting participants with several DRM lists, McDermott (1997) found priming for the critical items on several implicit memory tests, and these effects have since been replicated and extended to other priming tasks (e.g., Meade, Hutchison, & Rand, 2010). McDermott (1997) argued that such priming was due to lexical activation of the critical item at study.

Similarity-based theories

Within the second class of theories, the core proposal is that DRM false remembering is caused by similarity between the critical item and the studied items, as opposed to associative activation of the critical item. These theories have primarily been used to explain false recognition. For instance, the fuzzy trace theory of memory representations (e.g., Brainerd, Wright, Reyna, & Mojardin, 2001) postulates that studying a list of associates results in the formation of two types of memory traces. Verbatim traces represent detailed, item-specific information, whereas gist traces represent the more general thematic characteristics of the lists. At test, words that are consistent with the gist of the list (such as the critical item) will be highly familiar, and hence falsely remembered. A different similarity-based account was developed by Arndt and Hirshman (1998), as an extension of exemplar-based models of memory. Under their proposal, a separate "gist" representation need not be encoded. Instead, each studied item is encoded as a set of sensory and semantic features. At retrieval, the similarity between the features of the critical item and the encoded features will make this item familiar, and lead to false remembering.

Despite these differences, both of these similarity-based theories explain DRM false recognition via familiarity caused by semantic similarity, and neither theory appeals to activation of the critical item through associative links (see Brainerd, Yang, Reyna, Howe, & Mills, 2008). This last point poses important constraints on these theories. Without positing some sort of item-specific activation of the critical

FIGURE 21.2 The effects of retention interval on true and false recall (Toglia et al., 1999, Exp. 2). Participants studied five DRM lists, and were given a final free recall test at one of three retention intervals (between-subjects). Data are collapsed across blocked and mixed study presentation, although similar patterns were obtained at each level.

item, it is difficult to understand how these theories would explain the generation of this item on a recall test or on perceptually driven implicit memory tests (such as word stem completion). Theories based entirely on semantic similarity also have difficulty explaining why categorized lists (e.g., different pieces of furniture) are less effective than DRM lists at eliciting false memory effects (Pierce, Gallo, Weiss, & Schacter, 2005). Categorized lists are strongly thematic, but tend to have weaker associations than DRM lists.

Perhaps the strongest evidence that similarity-based processes might be involved in addition to associative activation are the effects of retention interval. It has been found that true recall decreases more over a delay than false recall (e.g., Toglia, Neuschatz, & Goodwin, 1999). Illustrative data from Toglia et al. (1999) are presented in Figure 21.2. True recall declined rapidly over a three-week retention interval, whereas false recall persisted at high levels. Fuzzy trace theory can account for such results because it holds that gist traces are more resistant to forgetting than verbatim traces. As a result, memory for list items (which is supported more by verbatim traces) decreases at a more rapid rate than memory for critical items (which is supported more by gist traces) as retention interval is increased.

Associative-based theories cannot account for such effects without additional assumptions. In the strongest form of these theories, the critical item would be activated multiple times at study and rehearsed like a list item. To the extent that the critical item is encoded like a studied item, the two should

have similar forgetting functions (especially when initial levels of true and false remembering were matched). In return, it is unclear how a similarity-based mechanism could account for the powerful relationship between associative strength and false remembering. For example, many of the words in the *whiskey* list seem to converge on the meaning of that word (e.g., *drink, drunk, beer, liquor*, etc.) just as words in the *window* list converge on its meaning (e.g., *door, glass, pane, shade*, etc.). Nevertheless, these lists greatly differ in associative strength (0.022 vs. 0.184), and in turn, they elicit dramatically different levels of false recall (3% vs. 65%; for additional discussion see Gallo & Roediger, 2002; Roediger et al., 2001). In sum, it appears that both associative activation and semantic similarity play a role.

Fluency-based attributions

Although they can explain false recall and false recognition, neither the associative-based account nor the similarity-based account can explain the perceptually detailed nature of DRM false memories very well. Roediger and McDermott (1995) found that false recognition of critical items was accompanied with high levels of confidence and frequent *remember* judgments. Both of these findings can be explained by thoughts of the critical item at study, but even this account cannot explain more detailed recollections. For instance, when lists are presented by multiple sources (auditory vs. visual, or different voices), participants are often willing to assign a source to critical items that are falsely recognized (Roediger, McDermott, Pisoni, & Gallo, 2004) or recalled (Hicks & Marsh, 1999). Similarly, using the Memory Characteristics Questionnaire (MCQ), participants often claim to recollect specific details about a critical item's presentation at study, such as perceptual features, list position, and personal reactions to the word (e.g., Mather, Henkel, & Johnson, 1997).

The cause of such illusory subjective phenomena is still debated in the literature (see Arndt, 2012), but one possible explanation is a *fluency-based attribution* process. Gallo and Roediger (2003) proposed that, at test, participants imagine having been presented with the critical item at study, perhaps in an effort to determine whether it was presented. This imagination is then mistaken for actual presentation because it is processed more fluently, or more easily, than would have otherwise been expected (cf. Chapter 10 on availability, Chapter 13 on the validity effect, Chapter 14 on the mere exposure effect, and Chapter 22 on the misinformation effect). If the attribution process occurs automatically, or nonconsciously, then the phenomenological experience would be one of remembering (cf. Jacoby, Kelley, & Dywan, 1989).

Both the associative-based and similarity-based theories predict that processing of the critical word will be enhanced by presentation of the related list, so that a fluency-based attribution process is consistent with either theory. That said, there are some clues that associative activation is uniquely involved (for some examples, see Franks, Butler, & Bishop, 2016; Gallo & Roediger, 2002, 2003).

Processes that reduce the effect

So far we have discussed processes that drive the DRM effect. No theoretical account would be complete, though, without considering editing processes that oppose these forces and reduce false remembering. Such processes have been conceptualized as *reality monitoring* under association-based theories (e.g., activation/monitoring theory), and item-specific or verbatim-based editing in similarity-based theories (e.g., fuzzy trace theory).

Some of the strongest evidence that such additional processes are involved comes from presentation manipulations that should not affect associative activation or semantic similarity, but nevertheless influence false remembering. These include presentation format (e.g., switching presentation from words to pictures, which has been found to reduce false recognition; Schacter, Israel, & Racine, 1999) and presentation modality (e.g., switching presentation from auditory to visual, which reduces the DRM effect; Smith & Hunt, 1998). Other evidence comes from presentation manipulations that should increase similarity or associative processes, but actually decrease false remembering. These include increasing the number of presentations of the study lists before a recognition test (e.g., Benjamin, 2001), and slowing presentation rate (which has been found to reduce false recall, but not necessarily false recognition; Gallo & Roediger, 2002).

To illustrate, consider a presentation-rate study by McDermott and Watson (2001). In those conditions that are relevant here, participants studied DRM lists at a range of visual presentation durations (20, 250, 1000, 3000, and 5000 milliseconds, between-subjects), and took an immediate free recall test after each list. As expected, true recall increased with more study time (0.17, 0.31, 0.42, 0.50, and 0.51). The pattern for false recall was more striking, with an initial increase and an eventual decrease (0.14, 0.31, 0.22, 0.14, and 0.14). The initial increase suggests that, within this range of extremely rapid presentation rates, slowing the duration afforded more meaningful processing and thus enhanced those activation-based or similarity-based processes that drive false recall. In contrast, the eventual decrease suggests that slowing presentation rates also increases item-specific processing of the list items. Apparently, the accrual of this item-specific information eventually reached a point where it began to facilitate monitoring processes that opposed false recall.

Subsequent research indicates that there are different kinds of monitoring or editing processes that influence DRM false remembering, as well as other kinds of false memories more generally (Gallo & Lampinen, 2016). One monitoring process – dubbed the distinctiveness heuristic by Schacter et al. (1999) – relies on the idea of retrieval expectations. According to this idea, making the studied items more memorable or distinctive allows participants to expect richer or more detailed memories at retrieval, effectively setting a more conservative decision criterion that helps them to reject false memories that fail to meet these expectations. Another kind of monitoring process occurs when participants realize, during the presentation of the study list, that the critical item is missing (Carneiro & Fernandez, 2013). If they remember this realization at test, then they can avoid falsely remembering

the critical item regardless of the distinctiveness of the studied information. This type of monitoring is most likely to occur if participants are warned to avoid the DRM illusion prior to the study phase, although of course the standard procedure does not give such warnings.

Neural mechanisms of the effect

We have discussed how both activation/similarity and editing processes may play a role in the DRM illusion. Further support for the distinction between these two opposing processes comes from neuropsychological data. Amnesic patients with varied etiologies (e.g., Korsakoff's or anoxia) tend to show decreased DRM false recognition relative to age-matched controls (e.g., Schacter, Verfaellie, Anes, & Racine, 1998). This decrease implies that damage to medial temporal regions (which were the primary, but not the sole areas that were damaged) reduces the likelihood of remembering the associative relations or gist that can cause false remembering. Related effects have been found in participants in the early stages of Alzheimer's disease, which also affects medial temporal regions (e.g., Gallo et al., 2006).

In contrast to those effects, patients with frontal lobe lesions showed enhanced DRM false recognition relative to age-matched controls (e.g., Budson et al., 2002). The frontal lobes have traditionally been implicated in monitoring processes, suggesting that the elevated levels of false recognition in this population were due to a breakdown in false-memory editing. Considered as a whole, the data from these different populations nicely illustrate the opposing influences of activation/similarity and editing processes.

Data from neuroimaging techniques, such as fMRI or EEG, have provided further insights into the neural mechanisms of the DRM effect (for example, see Cabeza, Rao, Wagner, Mayer, & Schacter, 2001; McDermott, Watson, & Ojemann, 2005). Some of the fMRI findings have been very illuminating, and there is more work yet to be done.

Conclusion

Associative memory illusions arise when information from the external world activates internal representations that may later be confused with the actual external events that sparked the association. As we have emphasized, we believe that this process is a general one with wide implications, because such associative activation is a pervasive fact of cognition. To use Jerome Bruner's famous phrase, people frequently go "beyond the information given" in drawing inferences, making suppositions, and creating possible future scenarios. Although these mental activities make us clever, they can also lead to errors when we confuse what we thought with what actually happened. The DRM paradigm provides a tractable laboratory task that helps open these processes to careful experimental study, and it also provides a rich arena for testing theories of internally generated false memories.

Summary

- People can falsely remember nonpresented events that are associated to events that occurred.
- Research has identified two sets of factors that are critical for the creation of these types of false memories: activation processes and monitoring processes.
- Activation processes, such as the mental generation of associative information, cause people to believe that the nonpresented event has actually occurred.
- Monitoring processes refer to the strategic editing of these retrieval products, in an effort to reduce false remembering.
- The frequent occurrence of these systematic errors provides important insights into the cognitive mechanisms of memory.

Further reading

For other relevant DRM reviews, see Gallo (2010) and Huff et al. (2015). For some other perspectives, see Schacter, Norman, and Koutstaal (1998) and Mitchell and Johnson (2009). Finally, while intriguing new findings are coming out every day (Prohaska, DelValle, Toglia, & Pittman, 2015), it also is good to keep a historical perspective. On that note, see Bruce and Winograd (1998).

References

Arndt, J. (2012). False recollection: Empirical findings and their theoretical implications. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 56, pp. 81–124). San Diego: Academic Press.

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language*, *39*, 371–391.

Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 941–947.

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

Brainerd, C. J., Yang, Y., Reyna, V. F., Howe, M. L., & Mills, B. A. (2008). Semantic processing in "associative" false memory. *Psychonomic Bulletin & Review, 15,* 1035–1053.

Bruce, D., & Winograd, E. (1998). Remembering Deese's 1959 articles: The Zeitgeist, the sociology of science, and false memories. *Psychonomic Bulletin & Review*, *5*, 615–624.

Budson, A. E., Sullivan, A. L., Mayer, E., Daffner, K. R., Black, P. M., & Schacter, D. L. (2002). Suppression of false recognition in Alzheimer's disease and in patients with frontal lobe lesions. *Brain*, *125*, 2750–2765.

Cabeza, R., Rao, S. M., Wagner, A. D., Mayer, A. R., & Schacter, D. L. (2001). Can medial temporal lobe regions distinguish true from false? An event-related functional MRI study of veridical and illusory recognition memory. *Proceedings of the National Academy of Sciences, 98,* 4805–4810.

Carneiro, P. & Fernandez, A. (2013). Retrieval dynamics in false recall: Revelations from identifiability manipulations. *Psychonomic Bulletin & Review, 20*(3), 488–495.

Cotel, S. C., Gallo, D. A., & Seamon, J. G. (2008). Evidence that nonconscious processes are sufficient to produce false memories. *Consciousness & Cognition*, *17*, 210–218.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17–22.

Franks, B. A., Butler, K. M., & Bishop, J. (2016). The effects of study order and backward associative strength on illusory recollection: A source-strength effect does not always occur. *Memory, 24*, 154–164.

Gallo, D. A. (2006). *Associative illusions of memory: False memory research in DRM and related tasks*. New York: Psychology Press.

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition, 38*, 833–848.

Gallo, D. A., & Lampinen, J. M. (2016). Three pillars of false memory prevention: Orientation, evaluation, and corroboration. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 387–403). New York: Oxford University Press.

Gallo, D. A., & Roediger, H. L., III. (2002). Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. *Journal of Memory and Language*, *47*, 469–497.

Gallo, D. A., & Roediger, H. L., III. (2003). The effects of associations and aging on illusory recollection. *Memory & Cognition*, *31*, 1036–1044.

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.

Gallo, D. A., Shahid, K. R., Olson, M. A., Solomon, T. M., Schacter, D. L., & Budson, A. E. (2006). Overdependence on degraded gist memory in Alzheimer's disease. *Neuropsychology, 20*, 625–632.

Gardiner, J. M., & Richardson-Klavehn, A. (2000). Remembering and knowing. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 229–244). New York: Oxford University Press.

Goodwin, K. A., Meissner, C. A., & Ericsson, K. A. (2001). Toward a model of false recall: Experimental manipulations of encoding context and the collection of verbal reports. *Memory & Cognition*, *29*, 806–819.

Hicks, J. L., & Marsh, R. L. (1999). Attempts to reduce the incidence of false recall with source monitoring. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *25*, 1195–1209.

Huff, M. J., Bodnar, G. E., & Fawcett, J. M. (2015). Effects of distinctive encoding on correct and false memory: A meta-analytic review of costs and benefits and their orign in the DRM paradigm. *Psychonomic Bulletin & Review, 22,* 349–365.

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 honour of Endel Tulving* (pp. 391–422). Hillsdale, NJ: Erlbaum.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*, 3–28.

Kirkpatrick, E. A. (1894). An experimental study of memory. *Psychological Review*, *1*, 602–609.

Mather, M., Henkel, L. A., & Johnson, M. (1997). Evaluating characteristics of false memories: Remember/know judgments and memory characteristics questionnaire compared. *Memory & Cognition*, *25*, 826–837.

McDermott, K. B. (1997). Priming on perceptual implicit memory tests can be achieved through presentation of associates. *Psychonomic Bulletin & Review*, *4*, 582–586.

McDermott, K. B., & Watson, J. M. (2001). The rise and fall of false recall: The impact of presentation duration. *Journal of Memory and Language*, *45*, 160–176.

McDermott, K. B., Watson, J. M., & Ojemann, J. G. (2005). Presurgical language mapping. *Current Directions in Psychological Science, 14,* 291–295.

Meade, M. L., Hutchison, K. A., & Rand, K. M. (2010). Effects of delay and number of related list items on implicit activation for DRM critical items in a speeded naming task. *Journal of Memory and Langauge, 62*, 302–310.

Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychological Review*, *106*, 398–405.

Mitchell, K. J., & Johnson, M. K. (2009). Source monitoring 15 years later: What have we learned from fMRI about the neural mechanisms of source memory? *Psychological Bulletin, 135*, 638–677.

Neuschatz, J. S., Payne, D. G., Lampinen, J. M., & Toglia, M. P. (2001). Assessing the effectiveness of warnings and the phenomenological characteristics of false memories. *Memory*, *9*, 53–71.

Pierce, B. H., Gallo, D. A., Weiss, J. A., & Schacter, D. L. (2005). The modality effect in false recognition: Evidence for test-based monitoring. *Memory & Cognition*, *33*, 1407–1413.

Prohaska, V., DelValle, D., Toglia, M. P., & Pittman, E. (2015). Reported serial positions of true and false memories in the Deese/Roediger/McDermott paradigm. *Memory*, online publication.

Robinson, K., & Roediger, H. L., III (1997). Associative processes in false recall and false recognition. *Psychological Science*, *8*, 231–237.

Roediger, H. L., III, Balota, D. A., & Watson, J. M. (2001). Spreading activation and the arousal of false memories. In H. L. Roediger, J. S. Nairne, I. Neath, & A. M. Suprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 95–115). Washington, DC: American Psychological Association Press.

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*, *21*, 803–814.

Roediger, H. L., III, & McDermott, K. B. (2000a). Distortions of memory. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 149–162). New York: Oxford University Press.

Roediger, H. L., III, & McDermott, K. B. (2000b). Tricks of memory. *Current Directions in Psychological Science*, *9*, 123–127.

Roediger, H. L., III, McDermott, K. B., Pisoni, D. B., & Gallo, D. A. (2004). Illusory recollection of voices. *Memory, 12*, 586–602.

Roediger, H. L., III, McDermott, K. B., & Robinson, K. J. (1998). The role of associative processes in creating false memories. In M. A. Conway, S. E. Gathercole, & C. Cornoldi (Eds.), *Theories of memory II* (pp. 187–245). Hove, England: Psychological Press.

Roediger, H. L., III, Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*, 385–407.

Schacter, D. L., Israel, L., & Racine, C. (1999). Suppressing false recognition in younger and older adults: The distinctiveness heuristic. *Journal of Memory and Language*, *40*, 1–24.

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

Schacter, D. L., Verfaellie, M., Anes, M. D., & Racine, C. (1998). When true recognition suppresses false recognition: Evidence from amnesic patients. *Journal of Cognitive Neuroscience*, *10*, 668–679.

Smith, R. E., & Hunt, R. R. (1998). Presentation modality affects false memory. *Psychonomic Bulletin & Review, 5*, 710–715.

Stadler, M. A., Roediger, H. L., III, & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition*, *27*, 494–500.

Toglia, M. P., Neuschatz, J. S., & Goodwin, K. A. (1999). Recall accuracy and illusory memories: When more is less. *Memory*, *7*, 233–256.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*, 1–12.

Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology*, *70*, 122–129.

22

MISINFORMATION EFFECT

Jacqueline E. Pickrell, Dawn-Leah McDonald, Daniel M. Bernstein, and Elizabeth F. Loftus

Memory is a complicated thing, a relative to truth, but not its twin.
Barbara Kingsolver, *Animal Dreams*

Our memory is an amazing entity with the ability to provide us with the continuous sense of self described by William James. Memory operates in an extremely efficient manner, allowing us to recall more pleasant than unpleasant memories and enough detail to reconstruct past experiences with reasonable accuracy. Memory is, however, subject to distortion, the study of which has historically been rooted in interference theory. Interference theory states that the reason we forget is that other memories in long-term memory impair our ability to retrieve the target memory. Memory involves a reconstructive process: It is vulnerable to interference from other experiences, especially from experiences that occur after a to-be-remembered event.

We describe contemporary studies of memory distortion, focusing on the "misinformation effect." This effect refers to the tendency for post-event misleading information to reduce memory accuracy for the original event. We begin our discussion with a brief overview of forgetting as a function of event encoding and event retrieval errors. We discuss various laboratory techniques that demonstrate the misinformation effect. The misinformation effect is a robust phenomenon whose underlying mechanism remains unclear. For example, we do not know who is susceptible to misinformation. We also do not know what kind of post-event information is necessary to distort a memory, nor how much can we distort that memory. Finally, we do not know what the misinformation effect says about memory's permanence.

Beginning with naturally occurring memory distortions, through use of the suggestion of misleading details, we discuss empirical demonstrations of the misinformation

effect including the most complex misinformation effect of all: the creation of memories for entirely false events. We briefly describe data from an experiment that examined consequences of creating a false memory of being attacked by a small dog as a young child, and provide a simple methodology for demonstrating the misinformation effect in a classroom. Finally, we explore the difficulty in distinguishing true from false memories and explore recent research examining the neural mechanisms of the misinformation effect.

Memory primer: Encoding, storage, and retrieval

Memory consists primarily of information encoded and stored in such a way as to facilitate retrieval of that information. The quality of what is encoded and the way it is encoded directly influence the subsequent retrieval of that information. Numerous factors affect the encoding process, beginning with attention. In order for the encoding process to succeed, we must attend to information. Additionally, the depth to which we process the encoded details influences the encoding process. Elaborating on information that is observed, particularly by linking the event or detail to previously learned information rather than experiencing the information with no conscious effort to remember it, will produce better encoding. Moreover, unusual, unexpected, and unfamiliar information tends to stand out in our mind, resulting in good encoding. It is safe to say that the accessibility of information that is poorly encoded or not encoded in memory at all will not improve with the passage of time.

Researchers typically view remembering as a reconstructive process. The memory for an event is not stored in its entirety, as an exact replica of the event. Rather, the event is broken and organized into personally meaningful pieces. These mental representations capture the gist or essential meaning of the event. Retrieval involves a mental re-enactment of the original experience. This reconstructive process involves distortions, in the form of omissions, elaborations, or misperceptions. Some theorists have argued that the success of retrieving previously encoded details depends, to a large degree, upon the extent to which the retrieval context matches the original encoding context (Tulving & Thompson, 1973).

Related to the organization of memory is research done on schematic memory. The definition of schema is an organized pattern of thought about some aspect of the world, such as events, situations, objects, or a class of people (Bartlett, 1932). Our schematic knowledge may play a role in each of the primary memory processes: encoding, storage, and retrieval. The encoding process, for example, is influenced by what we attend to and what we use to guide our understanding. Extracting an event's gist directs the storage method, in that highly associated information is stored together and distributed across the cortex. Schematic knowledge also can be very useful in the reconstructive process of memory. When retrieval occurs, our schemata aid the process, in that we may rely on what most likely happened and construct memories that allow us to provide a desired level of detail without having to reencode the information each time we experience it. Unfortunately, a schema

can distort memory because as we store schematic knowledge we do not have to attend to everything in our environment. When we rely on our schematic knowledge to report a memory, we usually remember things as being quite typical. Since we use our knowledge to identify an event or situation and to form a representation of that event, the errors committed when we "fill in the gaps" may negatively influence the accuracy of our memory.

Misinformation

By the middle of the 20th century, many scientists were searching in vain for the engram, a hypothetical memory trace that left an identifiable and indelible physical trace in the brain. The engram was believed to be both permanent and localizable, two views that dominated early cognitive neuroscience. The permanence theory gained support in the 1950s when neurosurgeon Wilder Penfield began operating on patients with epilepsy. During surgery, he kept patients awake but anesthetized them so they could respond when he stimulated certain areas of the brain. He claimed that by using this technique he accessed memories. The media embraced this concept and communicated the hype to the public as well as to the scientific community. The permanence theory was given a big boost. In reality, Penfield operated on 1,100 patients, and only 40 responded by producing anything like a memorial response when the temporal lobe was stimulated. Only five of those 40 reported a complete episodic memory, and some of these could be ruled out as real memories by other evidence. When Elizabeth Loftus and her colleagues (Loftus, 1975; Loftus, Miller, & Burns, 1978; Loftus & Palmer, 1974) claimed that people's memories are malleable after exposure to misleading post-event information, the social and theoretical implications of this finding caused a flurry of interest in the misinformation effect.

In their original studies, Loftus and colleagues demonstrated how question wording and the introduction of misleading post-event information influence what is remembered (see Chapter 20). In one study, Loftus (1975) showed participants a film of an automobile accident, and then asked half of them, "How fast was the white sports car going when it passed the barn while traveling along the country road?" No barn was shown in the film; however, many of the participants who had been asked about the barn claimed to have seen it in the film. In another study, participants answered one of the following questions about a car accident depicted in a film that they had seen: (1) How fast were the cars going when they hit each other? or (2) How fast were the cars going when they smashed into each other? (Loftus & Palmer, 1974). The researchers found that the latter question produced higher speed estimates than did the former question. Simply changing the word "hit" to "smashed into" affected the subject-witness's memory for the original event. In another study, participants watched a slide sequence involving a car/pedestrian accident. In the slide sequence, a car arrives at an intersection, turns right and hits a pedestrian. Half the participants saw a yield sign, while the remaining participants saw a stop sign. Later, some participants were asked a question containing a misleading suggestion

about either a stop sign or a yield sign (whichever sign they had not seen in the slide sequence). When tested for their memory of the original slides they had seen, many of the misled participants mistakenly claimed that they had seen the sign that had been suggested rather than the sign that they had actually seen (Loftus et al., 1978). The traffic-sign study is also used as basis for a classroom demonstration given in Text box 22.1. In sum, these studies demonstrate that misleading post-event information affects what people erroneously report about the past.

TEXT BOX 22.1 MISINFORMATION STUDY

For a classroom demonstration of the misinformation effect, we suggest using a modified version of Loftus et al.'s (1978) study. To demonstrate the effect in the most straightforward manner, we will use one independent variable, the type of information, with three levels (consistent, inconsistent, neutral). Dependent variable is the frequency of choosing the correct slide in a forced-choice recognition test. Based on the replicability of the misinformation effect, and the conversion of odds ratios to effect sizes (Blank & Launay, 2014), we assume a large effect size (Cohen's $d = .8$). Thus, with $\alpha = .05$ and $1 - \beta = .95$, plan on a sample size of $N = 42$ per group. To test for statistically significant differences between conditions, a χ^2-test can be used (or a binomial test, if expected frequencies are below 10).

Method

Materials and procedure

A series of 8 color slides are shown in a sequential manner to depict an auto-pedestrian accident. The modified set may be accessed at the following URL: [https://www.routledge.com/products/9781138903425]. The set includes the following 8 slides (plus 8 additional distractor slides):

- a black car approaching on a neighborhood street
- the black car passing a red car driving in the opposite direction
- the black car approaching an intersection
- the black car stopped at a stop sign with a pedestrian on the right
- a female pedestrian, walking her dogs, crossing the street in front of the car
- the pedestrian lying on the road in front of the black car
- the driver of the black car getting out of the car
- the driver of the black car speaking to the injured pedestrian

Eight questions, including one critical question, are administered immediately after participants view the slides. These questions serve as a memory test

for the information presented in the slide sequence, and are as follows (note that the three versions of the critical question are in italics):

- Did you see the black car approaching from a distance?
- Did you see a bicycle?
- Did you see the bridge?
- *Was the male pedestrian on the right or left of the black car while it was stopped at the intersection with the stop sign?* (consistent)
- *Was the male pedestrian on the right or left of the black car while it was stopped at the intersection with the yield sign?* (inconsistent)
- *Was the male pedestrian on the right or left of the black car while it was stopped at the intersection?* (neutral)
- Did you see the taxi cab across the street?
- Was the pedestrian who was hit a man or a woman?
- Was the pedestrian walking one or two dogs?
- Did the driver get out of the car and speak to the injured pedestrian?

All participants view the slide sequence showing a stop sign at the intersection. Participants are then randomly assigned to receive consistent, inconsistent, or neutral information with respect to the original traffic sign viewed in the slide sequence. This information is delivered through administration of the eight questions. Participants receiving information asking about the stop sign are receiving information consistent with the slide series (consistent). Another third of the participants receive information where they have it suggested that they saw a yield sign (inconsistent). The final control-condition participants receive no information (neutral).

A yes-no recognition test is administered after a brief (we suggest 15–20 minutes) filler task. Eight pairs of test slides are shown to participants who report which slide of each pair they had seen before. The critical slide pair, the scene showing either the original stop sign or the misleading yield sign, is randomly placed within the 8 pairs of slides.

Results

What proportion of participants correctly chooses the stop sign from the original slide-show sequence? With the recognition test administered immediately after the exposure to the misinformation, Loftus et al. (1978) found that participants who received the information that was inconsistent with what they had actually seen were much less accurate in their responses compared to the participants in the other two conditions. In addition, when the retention interval is increased to two days (see Figure 22.1) and we compare the response of those in the neutral condition with the other two groups, receiving the consistent information improved performance on the recognition task. Along with

FIGURE 22.1 The effect of the type of information on the proportion of correct answers given on the recognition test two days after viewing the slide show and completing the questionnaire.

the fact that the type of information has an effect on memory for the original information, Loftus et al. demonstrated that delay of the questionnaire containing the misleading information also hindered memory performance. Apparently, time allows the original memory trace to weaken, thus making it easier for an erroneous memory report to be given.

Theoretical accounts of the misinformation effect

About the same time as the first misinformation studies were reported, there was also a rising concern regarding the reliability of eyewitness testimony. The misinformation effect had important practical implications. Soon thereafter, a theoretical issue arose. The very nature of memory was now in question. When participants had seen a stop sign, but received a suggestion about a yield sign, and then claimed that they saw a yield sign, what happened to the stop sign in their memory? Was it still there, but perhaps harder to retrieve? Was it erased or altered in the process of absorbing the suggestive information? McCloskey and Zaragoza (1985) attacked this memory-impairment hypothesis by claiming that "misleading post-event information has no effect on memory for the original event." (p. 2) The misinformation results were not in question, but the interpretation was. McCloskey and Zaragoza's key arguments involved the procedure used in these studies. They argued that the acceptance of misinformation occurred because:

- Participants never encoded the original information and remembered the post-event information.
- Participants encoded both the original and the post-event information and deliberated about what to report, eventually reporting the information they surmised the questioner is looking for.
- The original information was just forgotten.

McCloskey and Zaragoza proposed a modified procedure in an attempt to account for participants who fall into any of the categories above. They conducted six experiments using both this modified procedure and the original. The *modified* procedure was identical to the original procedure except the test was different. The original procedure had participants choose from either the original event/item or the misleading event/item. In the modified procedure, participants chose between the original event/item and a *novel* event/item. The misleading information was not an option. McCloskey and Zaragoza hypothesized that if the misleading information altered memory for the original event, when that information was not an option at test, memory for the original information would be selected less often than selected by those in the control condition.

Collapsing over many experiments McCloskey and Zaragoza successfully replicated the misinformation effect using the original procedure: Participants in the control condition correctly reported the original information 72% of the time, whereas the misled participants correctly reported the original information only 37% of the time. However, the results were different in the modified procedure. Here, control participants correctly reported the original information 75% of the time, while misled participants correctly reported the original information 72% of the time. The researchers' conclusions were direct (McCloskey & Zaragoza, 1985, p. 7):

> Misleading post-event information does not impair participants' ability to remember what they originally saw. In other words, misleading information neither erases the original information nor renders it inaccessible.

McCloskey and Zaragoza explained these results with factors other than memory-impairment, specifically, either task demands or strategic effects. Zaragoza, McCloskey, and Jamis (1987) tested the modified and original procedures with recall rather than recognition and supported their belief that differences between the misled and control conditions in the original procedure are due to factors other than memory-impairment. McCloskey and Zaragoza asserted that the test should be such that process-of-elimination strategies can be controlled.

Subsequent experiments designed to reduce or eliminate strategic effects have demonstrated effects of memory-impairment (see Ayers & Reder, 1998, for a review). Moreover, the modified test may be insensitive to memory-impairment. When overall memory performance is low, the effect may need the additional strength of the exposure to the misleading information once again at test. More recent studies using false images and videos (Wright, Wade, & Watson, 2013) supported the impact of repetition of the post-event information on memory strength, and in a series of experiments, Chan and LaPaglia (2013) demonstrated that even after successful retrieval, original memories were susceptible to misinformation if reactivated prior to relearning. They proposed that this susceptibility may be due to disruption of reconsolidation.

It seems that each of the explanations offered accounts for some of the findings but as yet we have not developed a theory to explain all the results. Metcalfe (1990,

in Ayers & Reder, 1998, p. 19) proposed her CHARM theory, a memory model that accounts for the misinformation data by means of memory trace alteration. This is a single-trace model that explains the integrated and blended memory data but falls short of explaining the small effects sometimes seen using the modified procedure.

Ayers and Reder (1998) proposed an activation-based model that might explain the various misinformation effect findings in terms of a source of activation confusion (SAC) model of memory. This multi-trace model predicts that in our classic misinformation example, a participant might be aware of the high activation of the concept "yield sign" but be unaware of the reason that it was activated. "Yield sign" would be more highly activated at test than the original "stop sign" because it had been activated more recently. This model is consistent with the ideas expressed by Kelley and Jacoby (1996), that under some conditions, the source of activation is unclear. If the source is either unavailable or unknown, it may be misattributed, resulting in memory errors. This model is, however, at direct odds with an integration/blending theory. If the memory trace is altered or overwritten, there can be no source misattribution because there is only one source. In summary, our colleagues Hyman and Pentland (1996) may have said it best: "Although the misinformation effect is easily replicated, the explanation of such memory errors is hotly contested." (p. 101)

Today, there seems to be a consensus among cognitive psychologists that no single process is responsible for all of the misinformation effects (Frenda, Nichols, & Loftus, 2011). In other words, a misinformation response can arise for a number of different reasons. Sometimes the original and the misinformation items can coexist. Sometimes, the misinformation appears to weaken or impair the original item. But whatever the process that leads to a misinformation response, the empirical findings, when taken as a whole, support the reconstructive nature of memory and illustrate how this reconstruction leaves memory susceptible to errors.

Manipulations of the misinformation effect

Many factors influence the effectiveness of misinformation. First, the passage of time renders the original memory less accessible, thereby allowing misinformation to "creep in" undetected. Second, the subtler the discrepancies are between the original information and the post-event information, the stronger the misinformation effect. Third, the more ignorant one is of the potentially misleading effects of post-event information, the more susceptible s/he will be to the misinformation effect.

Who is susceptible to the misinformation effect? One somewhat unusual misinformation study provides insight into just *who* is most likely to accept misinformation (Loftus, Levidow, & Duensing, 1992). At a science museum in San Francisco, approximately 2,000 visitors participated in a typical misinformation study. Participants viewed a short film and then some of the participants received misinformation while others did not. All of the participants then answered questions about the film. While most misinformation studies have been conducted in University laboratories with college students, this particular study included people

ranging in age from 5 to 75 years, providing a unique opportunity to gather information on the effect of age and misinformation. Consistent with other studies involving children (Ceci, Ross, & Toglia, 1987), the youngest participants showed large misinformation effects. Additionally, the elderly showed a large misinformation effect, too (Loftus et al., 1992).

The preceding discussion has focused primarily on the fact that memory details are sensitive to misinformation. For example, subtle word choice embedded within post-event information can influence memory (for more examples, see Chapter 20). Such memory distortion can be seen even in "flashbulb memories," memories named for their highly emotional, meaningful, and subjectively permanent nature. One could imagine how the following, subtly worded question could lead people to remember seeing details about the September 11 terrorist attacks that they never saw: "Did you see the explosion after seeing the plane crash into the Pentagon during the September 11 terrorist attacks?" There was no footage of the plane crashing into the Pentagon. However, the question suggests that such footage not only exists, but that the individual might have seen it. This suggestion, coupled with the knowledge that a plane did, in fact, crash into the Pentagon, might lead people to think mistakenly that they saw the plane crash (Ost, Vrij, Costall, & Bull, 2002).

In addition to changing memory for details, misinformation can also plant entire events into a person's memory. In one study, 25% of participants either partially or wholly accepted the false suggestion that they had been lost in a shopping mall at the age of five (Loftus & Pickrell, 1995). Likewise, Hyman, Husband, and Billings (1995) convinced many of their participants that, as children, they had knocked over a punch bowl at a wedding and spilled punch on the bride's parents. Both studies utilized a procedure in which the researchers acquired three true memories from the participants' parents. The researchers then provided participants with the true memories and the false memory. Participants were asked to try to remember the events and to describe them in detail. Not only did nearly one quarter of the participants come to believe that the false event had occurred, but they also elaborated on the false event (e.g., "I do remember her [an elderly lady] asking me if I was lost, . . . and asking my name and then saying something about taking me to security"; Loftus & Pickrell, 1995, p. 724).

There are, of course, other ways to increase one's confidence in various childhood events. Mock personality profiles and dream interpretations are procedures that utilize the power of suggestion to increase one's subjective confidence in events that never occurred. Participants might learn that their personality profiles reveal that, as young children, they had been attacked by a dog. Or a dream "expert" might interpret a dream as meaning that, as a young child, one had to be rescued by a lifeguard. Such misinformation can increase participants' confidence in the critical events (Mazzoni, Loftus, Seitz, & Lynn, 1999).

Another form of suggestion used to inflate confidence in childhood events involves imagining an event in detail that never occurred. For example, Garry, Manning, Loftus, and Sherman (1996) asked participants about a variety of childhood events, including whether they had broken a window with their hand. Next, some

participants imagined in detail running through the house as a child and tripping and falling and breaking a window with their hand, cutting themselves and bleeding. This type of imagination exercise significantly increased participants' confidence that they experienced this event in their childhood. Additional research on "imagination inflation" established the utility of this procedure (see Drivdahl & Zaragoza, 2001).

Another technique used to create false memories involves altered photographs. Wade, Garry, Read, and Lindsay (2002) asked participants about a variety of childhood events, including riding in a hot air balloon. The researchers then obtained childhood photographs of their participants and inserted these pictures into a photograph depicting a hot air balloon ride. After participants saw themselves riding in a hot air balloon, they began to "remember" the ride, even though the experience had been completely fabricated. Recent techniques continue to capitalize on the development of technology. Not only have we gotten very good at altering photographs, but an increasing literature demonstrates the suggestive power of exposing people to altered video images of their past behavior (Nash & Wade, 2009; Nash, Wade, & Brewer, 2009).

Consequences

We know that there are consequences of true experiences that we no longer remember. One of us (JP) knows firsthand of an individual (her daughter) who retained her fear of dogs long after she had forgotten having been attacked by a large dog when she was two years old. What if the daughter had had a false belief about being attacked? Would this also lead to a similar kind of lingering fear? More generally, are there long-term consequences associated with creating false beliefs or memories? If a person comes to believe that they were attacked by a dog while a child, might they be more inclined as an adult to own a cat instead of a dog?

New research illustrates just how the malleability of memory in general, and specifically, the generation of false memories may be used in behavior modification therapy. Using an imagination paradigm, Clifasefi, Bernstein, Mantonakis, and Loftus (2013) successfully planted a memory of participants becoming ill after consuming a specific alcohol before the age of 16 years. The authors then noted a decrease in participants' self-reported preference ratings for that specific alcohol compared to participants' pretest ratings. The potential to influence not only negative habits we would like to eliminate, but the potential to influence nutritional selections may prove to be quite beneficial as culturally we move toward a desire for healthier living.

Braun (1999) examined the effects of misinformation, in the form of an advertisement, on one's subsequent memory for the taste of orange juice. Braun asked participants to taste orange juice and to describe its flavor. Some participants then evaluated the effectiveness of a false advertisement that described the orange juice that they had tasted 15 minutes earlier as being "sweet, pulpy and pure" (p. 323). The misinformation in this advertisement significantly altered participants'

memories a few minutes later for the original orange juice that they had tasted. Specifically, the false advertisement made participants think that the orange juice that they had tasted earlier was better than it in fact had been (see Chapter 20). The results of Braun's work, and Clifasefi and colleagues' work, suggest that it is possible to alter a true memory and to plant a false memory, with either positive or negative consequences.

Possible mechanisms

What is not clear from the work we have described are the boundary conditions of the misinformation effect and the underlying mechanism responsible for memory distortion. Some have investigated the limits to what types of memories can be created through suggestion or imagination. Mazzoni, Loftus, and Kirsch (2001) argued that an event must be seen as plausible in the rememberer's culture and that it must be seen as personally plausible before one comes to incorporate the experience into his/her own autobiographical memory.

One possible mechanism that might explain memory distortion after different forms of misinformation is that of familiarity (Garry et al., 1996). Larry Jacoby and colleagues (e.g., Jacoby, Kelley, & Dywan, 1989; see also Whittlesea & Williams, 2001) have argued that many false memories arise through the misattribution of familiarity. According to this notion, when participants fluently process an event or experience, they experience a feeling of familiarity. They then search for reasons that might explain this processing fluency. If they are unable to detect an obvious source, they may attribute the fluency to past experience (see Chapters 10, 13, and 14 in this volume).

Familiarity attribution may help explain why people accept misinformation and why they increase their confidence for childhood events after imagining these events or after being told that the events likely occurred. In such cases, people will process the imagined or suggested event more fluently than they would otherwise have processed it. They will, in turn, evaluate their present processing experience. Instead of correctly focusing upon the misinformation, the imagination exercise, or the suggestion as the source of familiarity, they mistakenly attribute the familiarity to their childhood.

Distinguishing true from false memories

Behavioural evidence

Unfortunately, it is very difficult to tell whether an individual memory is real or a product of imagination or some other process. In fact, research over the past 30 years suggests that it is virtually impossible to determine whether a particular memory is real (see Bernstein & Loftus, 2009). In a quest to distinguish true from false memories, researchers compared participants' reports for true and false memories (Loftus & Pickrell, 1995). The participants used more words when describing their

true memories, and more highly rated the clarity of their true memories than they rated the clarity of their false memories. Confidence ratings revealed that people were less confident in their false memories than they were in their true memories.

Roediger and McDermott (1995) created false memories for words not presented in lists (see Chapter 21). Not only were false memories as common as true memories in the study, but participants expressed as much confidence in their false memories as they did in their true memories. Perhaps even more upsetting to those who hope to distinguish true from false memories, participants claimed to "remember" (or mentally relive) the experience of having heard words before that they had not heard. Thus, false memories were easily created, and they were virtually indistinguishable from the true memories.

Porter, Yuille, and Lehman (1999) investigated whether phenomenological and content features could discriminate between real and false memories in an effort to systematically assess the credibility of childhood memories. Porter et al. employed criteria based on the Memory Assessment Procedure (MAP). This procedure consists of seven items designed to record phenomenological (subjective) features, and five items designed to record the specific features of memories. Content analysis revealed that participants rated true memories as more vivid/clear and more detailed. Participants also expressed more confidence in true memories when compared to the implanted memories. No difference was measured in stress levels. Additionally, 40% of participants recalled the real memory from a participant perspective. That is to say, they "re-experienced the event from their own eyes" (p. 28). The remaining 60% of participants viewed the real memory from the observer perspective, meaning they could view themselves as if watching a movie. The percentages were exactly reversed when participants recalled the implanted memory: 60% saw it from the participant perspective and 40% from the observer perspective. Although this was not a reliable difference, it does suggest that real and false memories may possibly differ in terms of their phenomenological and content features.

Heaps and Nash (2001) examined differences between true and false memories. A variation of the Loftus and Pickrell paradigm was employed to plant false memories in participants. They used information from relatives to suggest to people that they had undergone certain experiences. The researchers replicated the basic finding of greater detail remembered for the true memories when compared to the false memories. On first pass, the true memories appeared different from the false memories, because they were rated as being more important, more emotionally intense, more typical, and as having clearer imagery. These distinctions were eliminated when rehearsal frequency was used as a covariate in the statistical analyses. This suggests that increased rehearsal shifts the false memory closer to the recollective experience of true memories. A final observation concerned the consequences of the target event. False memories contained less information about any consequences of the event.

The move into the legal arena often lags behind the scientific advances. It will be long after we are able to distinguish false memories from true ones in the lab that we may apply such standards to tests in the courtroom (Schacter & Loftus,

2013). Heaps and Nash (2001) illustrated the problem: "the possibility [exists] that repeated remembering of false memories over greater periods of time [longer than three weeks] may make recollective experience more complete and more like that found in true autobiographical memories" (p. 17). Few things are more rehearsed than legal testimony, traumatic memories, and events made important by police and legal proceedings. Currently, external corroboration remains the only reliable way to determine the veracity of a memory.

Neuroscientific evidence

Although the behavioural influences of the misinformation effect have been studied extensively, it has only been in the past decade that the neural mechanisms of this effect have been explored. Researchers have used neuroimaging techniques such as functional magnetic resonance imaging (fMRI) and event-related potentials (ERPs) to differentiate false memories from true memories at retrieval (e.g., Baym & Gonsalves, 2010; Curran, Schacter, Johnson, & Spinks, 2001; Fabiani, Stadler, & Wessels, 2000; Kim & Cabeza, 2007). Other work has shown that stimulation of the left anterior temporal lobe with repetitive transcranial magnetic stimulation (rTMS) may reduce certain types of false memories (Gallate, Chi, Ellwood, & Snyder, 2008). After decades of studying the behavioural consequences of the misinformation effect, Loftus (2005) deplored the paucity of physiological measures of this effect and applauded the recent (at that time) work of Okado and Stark (2005).

Okado and Stark (2005) were the first to use neuroimaging techniques to investigate the misinformation effect. They extended the research at that time by using detailed vignettes as opposed to word lists in the study phase (also known as the original event phase). As an example, one of the vignettes depicted a woman showing her friend the *South Park* DVD that she had purchased, and his response. Participants viewed the vignettes while in the fMRI scanner. After a short delay, participants viewed the same vignettes, but unbeknownst to them, with changes to several critical items (misinformation phase). Using the previous example, instead of showing the *South Park* DVD, the woman showed her friend an *X Files* DVD. After two days, participants took a recognition memory test. Differences in neural activity in the left hippocampus tail and the perirhinal cortex during the original event phase and misinformation phase served as predictors for later recounting of true or false memories.

Extending the Okado and Stark (2005) paradigm, Stark, Okado, and Loftus (2010) used both visual and auditory stimuli. The original event phase utilized vignettes that participants viewed in a slide show. The misinformation phase, which occurred a day later, differed from the previous study by using a recorded narrative, where participants heard the information in the vignettes, with critical details changed. Shortly after the misinformation phase, participants entered the fMRI scanner. They read 24 sentences that briefly described events from the vignettes, and answered whether they had seen each in the original event. Outside the scanner, participants viewed the questions and their responses from the previous phase and

were asked the source of their memory (e.g., "Remembered it was in both the slides and the narrative" or "Could not remember and guessed", p. 486). Sensory reactivation was observed in both the visual cortex and auditory cortex. When stimuli were presented visually, correct recollection was accompanied by greater activity in the visual cortex. False recollection of auditory stimuli, however, was accompanied by greater activity in the auditory cortex.

Baym and Gonsalves (2010) adapted the Okado and Stark (2005) paradigm to assess how misinformation and visual imagery at encoding affect subsequent memory retrieval. Participants experienced the original event phase and the misinformation phase inside the fMRI scanner. During the original event phase, participants viewed a written description of each event, followed by photographic descriptions. During the misinformation phase, participants viewed the same initial written description, but without photographs. Participants were asked to imagine the actual event while reading the description. A day later, outside the scanner, participants took a multiple-choice recognition test, followed by a 'conflict test,' where they were asked whether there were any differences between the photograph and the written description. Greater neural activity in the ventral visual areas at encoding predicted subsequent true memories. Both true and false memories were related to greater activity in the right frontal cortex and bilateral MTL, which was not the case for forgetting. Other research has been successful in the prediction of true memories, false memories, and subsequent forgetting, outside the context of studying misinformation (e.g., Abe et al., 2013; Gonsalves et al., 2004; Karanian & Slotnick, 2014).

More recent work by Meek, Phillips, Boswell, and Vendemia (2012) has broadened the misinformation paradigm by adding deception. This research could have important implications for eyewitness testimony. Could individuals lie about an event when they do not know all the details? In this study, participants observed a series of slides that showed a crime being committed, then read a narrative that contained misinformation. Prior to the test phase, participants wore electroencephalograph (EEG) caps that recorded ERPs. Participants were then separated into one of two groups where they had to either tell the truth or lie about their recollections. The ERPs showed lateralized differences between misinformation and information that was consistent with the observed crime. There were also lateralized differences between deception and truth-telling. Such work provides exciting avenues for studying the neurophysiology of false memories.

Conclusion

What is clear from the myriad studies of memory distortion – and, specifically, those exploring the parameters of the misinformation effect – is that misinformation can lead to memory changes, ranging from minute details of events to memory for entire false events. While the misinformation effect is a robust phenomenon, it can affect people in many ways. It can add to memory, it can just change what people report, and it sometimes appears to lead to impairments in previously stored memories.

The current comparison of true and false memories shows us that in general, the phenomenological experiences of both types of memory are indistinguishable (for details see Chapter 21). While some studies point to participants being clearer and more confident in their true memories (Loftus & Pickrell, 1995; Porter et al., 1999), others report less discernible evidence (Heaps & Nash, 2001; Roediger & McDermott, 1995). Along with technological advances in brain imaging techniques comes the potential to distinguish true from false memories at a physiological level. The ability to determine the "true" nature of a memory and exactly which areas of the brain are activated during the remembering process holds promise for both the courtroom and the therapist's office.

Heaps and Nash (2001) showed that false memories contained less information about consequences of the false event when compared to the consequences revealed when the memory was true. Future research on the consequences of memory distortion may also open a window on the issue of distinguishing true from false memories. The need to understand false memory reports and the impact on everyday actions becomes more important when we realize we are all susceptible to misinformation in myriad ways. And given that we can manipulate positive events resulting in positive consequences as well as negative, there may be therapeutic implications whereby changes in memory and associations may be beneficial in changing future behaviors.

Neuroimaging techniques can be valuable tools to help researchers distinguish true from false memories. However, it is important to use caution when drawing conclusions from these results. The data from fMRI studies reflect an average of each individual participant's results as well as across participants (Frenda et al., 2011). While providing important information about how false memories evolve, averages do not enlighten the memory researcher about the veracity of a particular memory (see Bernstein & Loftus, 2009). Certainly more work is needed on the neurophysiology of false memories elicited within the misinformation effect (see Schacter & Loftus, 2013, for possible links between the misinformation effect and what is called reconsolidation, in which reactivated memories become temporarily vulnerable to change).

Summary

- The misinformation effect occurs when a person receives post-event information (e.g., new information after the original event) that interferes with the person's ability to accurately recall the original event.
- The misinformation effect is a very robust phenomenon.
- The underlying cognitive mechanisms remain unclear. As explanation, cognitive theories claim that either the original memory trace is altered or that the original memory trace remains intact but inaccessible.
- The effect has been found with different materials and experimental procedures and ranges from changing details of an event to planting false memories of an entire event.

- The consequences of memory distortions have important implications for several applied problems.
- As yet, we have no reliable means of distinguishing true from false memories.

Further reading

For an excellent overview of research on the misinformation effect, see Ayers and Reder (1998). In addition, practical legal applications are explored in depth in Loftus and Ketcham's (1991) *Witness for the Defense*. For recent empirical articles, *Applied Cognitive Psychology* (published by John Wiley & Sons) and the *Journal of Applied Research in Memory and Cognition* (published by Elsevier) cover theoretical, empirical, as well as applied aspects of the misinformation effect. Additionally, Frenda et al. (2011) is well suited for a general audience.

References

Abe, N., Fujii, T., Suzuki, M., Ueno, A., Shigemune, Y., Mugikura, S., . . . Mori, E. (2013). Encoding- and retrieval-related brain activity underlying false recognition. *Neuroscience Research, 76*, 240–250.

Ayers, M. S., & Reder, L. M. (1998). A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychonomic Bulletin & Review*, *5*, 1–21.

Bartlett, F. C. (1932/1995). *Remembering: A study in experimental and social psychology.* Cambridge: Cambridge University Press.

Baym, C. L., & Gonsalves, B. D. (2010). Comparison of neural activity that leads to true memories, false memories, and forgetting: An fMRI study of the misinformation effect. *Cognitive, Affective, and Behavioral Neuroscience, 10*(3), 339–348.

Bernstein, D. M., & Loftus, E. F. (2009). How to tell if a particular memory is true or false. *Perspectives on Psychological Science*, *4*, 370–374.

Blank, H., & Launay, C. (2014). How to protect eyewitness memory against the misinformation effect: A meta-analysis of post-warning studies. *Journal of Applied Research in Memory and Cognition, 3*, 77–88.

Braun, K. A. (1999). Postexperience advertising effects on consumer memory. *Journal of Marketing Research*, *25*, 319–334.

Ceci, S. J., Ross, D. F., & Toglia, M. P. (1987). Suggestibility of children's memory: Psycholegal implications. *Journal of Experimental Psychology: General, 116*, 38–49.

Chan, J. C. K., & LaPaglia, J. A. (2013). Impairing existing declarative memory in humans by disrupting reconsolidation. *Proceedings of the National Academy of Sciences, 110*(23), 9309–9313.

Clifasefi, S. L., Bernstein, D. M., Mantonakis, A., & Loftus, E. F. (2013). Queasy does it: False alcohol memories lead to diminished alcohol preferences. *Acta Psychologica, 143*, 14–19.

Curran, T., Schacter, D. L., Johnson, M. K., & Spinks, R. (2001). Brain potentials reflect behavioral differences in true and false recognition. *Journal of Cognitive Neuroscience, 13*(2), 201–216.

Drivdahl, S. B., & Zaragoza, M. S. (2001). The role of perceptual elaboration and individual differences in the creation of false memories for suggested events. *Applied Cognitive Psychology*, *15*, 265–281.

Fabiani, M., Stadler, M. A., & Wessels, P. M. (2000). True but not false memories produce a sensory signature in human lateralized brain potentials. *Journal of Cognitive Neuroscience*, *12*, 941–949.

Frenda, S. J., Nichols, R. M., & Loftus, E. F. (2011). Current issues and advances in misinformation research. *Current Directions in Psychological Science, 20*(1), 20–23.

Gallate, J., Chi, R., Ellwood, S., & Snyder, A. (2008). Reducing false memories by magnetic pulse stimulation. *Neuroscience Letters, 44*, 151–154.

Garry, M., Manning, C. G., Loftus, E. F., & Sherman, S. J. (1996). Imagination inflation: Imagining a childhood event inflates confidence that it occurred. *Psychonomic Bulletin & Review*, *3*, 208–214.

Gonsalves, B., Reber, P. J., Gitelman, D. R., Parrish, T. B., Mesulam, M. M., & Paller, K. A. (2004). Neural evidence that vivid imagining can lead to false remembering. *Psychological Science, 15*(10), 655–660.

Heaps, C. M., & Nash, M. (2001). Comparing recollective experience in true and false autobiographical memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 920–930.

Hyman, I. E., Jr., Husband, T. H., & Billings, F. J. (1995). False memories of childhood experiences. *Applied Cognitive Psychology*, *9*, 181–197.

Hyman, I. E., & Pentland, J. (1996). The role of mental imagery in the creation of false childhood memories. *Journal of Memory and Language*, *35*, 101–117.

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. In H. L. Roediger, III, & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 391–422). Hillsdale, NJ: Lawrence Erlbaum Associates.

Karanian, J. M., & Slotnick, S. D. (2014). False memory for context activates the parahippocampal cortex. *Cognitive Neuroscience, 5*(3–4), 186–192.

Kelley, C. M., & Jacoby, L. L. (1996). Memory attributions: Remembering, knowing, and feeling of knowing. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 287–307). Hillsdale, NJ: Erlbaum.

Kim, H., & Cabeza, R. (2007). Differential contributions of prefrontal, medial temporal, and sensory-perceptual regions to true and false memory formation. *Cerebral Cortex, 17*, 2143–2150.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, *7*, 560–572.

Loftus, E. F. (2005). Searching for the neurobiology of the misinformation effect. *Learning and Memory, 12*, 1–2.

Loftus, E. F., & Ketcham, K. (1991). *Witness for the defense: The accused, the eyewitness, and the expert who puts memory on trial*. New York: St. Martin's Press.

Loftus, E. F., Levidow, B., & Duensing, S. (1992). Who remembers best? Individual differences in memory for events that occurred in a science museum. *Applied Cognitive Psychology*, *6*, 93–107.

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.

Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals, 25,* 720–725.

Mazzoni, G. A. L., Loftus, E. F., & Kirsch, I. (2001). Changing beliefs about implausible autobiographical events: A little plausibility goes a long way. *Journal of Experimental Psychology: Applied, 7*, 51–59.

Mazzoni, G. A. L., Loftus, E. F., Seitz, A., & Lynn, S. (1999). Changing beliefs and memories through dream interpretation. *Applied Cognitive Psychology*, *13*, 125–144.

McCloskey, M., & Zaragoza, M. (1985). Misleading postevent information and memory for events: Arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General*, *114*, 1–16.

Meek, S. W., Phillips, M. C., Boswell, C. P., & Vendemia, J. M. C. (2012). Deception and the misinformation effect: An event-related potential study. *International Journal of Psychophysiology, 87,* 81–87.

Metcalfe, J. (1990). Composite holographic associative recall model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General, 119*, 145–160.

Nash, R. A., & Wade, K. A. (2009). Innocent but proven guilty: Eliciting internalized false confessions using doctored-video evidence. *Applied Cognitive Psychology, 23*, 624–637.

Nash, R. A., Wade, K. A., & Brewer, R. J. (2009). Why do doctored images distort memory? *Consciousness and Cognition, 18,* 773–780.

Okado, Y., & Stark, C. E. L. (2005). Neural activity during encoding predicts false memories created by misinformation. *Learning & Memory, 12*, 3–11.

Ost, J., Vrij, A., Costall, A., & Bull, R. (2002). Crashing memories and reality monitoring: Distinguishing between perceptions, imaginations and 'false memories'. *Applied Cognitive Psychology*, *16*, 125–134.

Porter, S., Yuille, J. C., & Lehman, D. R. (1999). The nature of real, implanted, and fabricated memories for emotional childhood events: Implications for the recovered memory debate. *Law and Human Behavior, 23*, 517–537.

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*, *21*, 803–814.

Schacter, D. L., & Loftus, E. F. (2013). Memory and law: What can cognitive neuroscience contribute? *Nature Neuroscience*, *16*, 119–123.

Stark, C. E. L., Okado, Y., & Loftus, E. F. (2010). Imaging the reconstruction of true and false memories using sensory reactivation and the misinformation paradigms. *Learning and Memory, 17*, 485–488.

Tulving, E., & Thompson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*, 352–373.

Wade, K. A., Garry, M., Read, J. D., & Lindsay, D. S. (2002). A picture is worth a thousand lies: Using false photographs to create false childhood memories. *Psychonomic Bulletin & Review*, *9*, 597–603.

Whittlesea, B. W. A., & Williams, L. D. (2001). The discrepancy-attribution hypothesis: II. Expectation, uncertainty, surprise, and feelings of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 14–33.

Wright, D. S., Wade, K. A., & Watson, D. G. (2013). Delay and déjà vu: Timing and repetition increase the power of false evidence. *Psychonomic Bulletin & Review, 20*(4), 812–818.

Zaragoza, M. S., McCloskey, M., & Jamis, M. (1987). Misleading postevent information and recall of the original event: Further evidence against the memory impairment hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 36–44.

23

HINDSIGHT BIAS

Rüdiger F. Pohl and Edgar Erdfelder

Whenever in hindsight, we tend to exaggerate what we knew in foresight. For example, after being told that "absinthe" is not a precious stone but rather a liquor, we are likely to express inflated confidence that we knew the solution all along (Fischhoff, 1975). This effect has been termed "hindsight bias" or "knew-it-along effect" (Fischhoff, 1975) and has been observed in numerous studies up to date. Hindsight bias was the focus of two meta-analyses (Christensen-Szalanski & Willham, 1991; Guilbault, Bryant, Brockway, & Posavac, 2004) and several overviews (e.g., Bernstein, Aßfalg, Kumar, & Ackerman, 2016; Hawkins & Hastie, 1990; Pezzo, 2011; Pohl, 2004; Roese & Vohs, 2012). Common to all studies is that participants initially are in a state of uncertainty (which is necessary in order to observe the effect); usually accomplished by using rather difficult knowledge questions or uncertain events. The results of these studies showed that after knowing the solution or outcome, respectively, participants are quite often simply unable to access their uncontaminated foresight knowledge state (Erdfelder, Brandt, & Bröder, 2007; Pohl & Hell, 1996) so that they are left to reconstruct their earlier given estimate in a biased manner. Moreover, participants are generally unaware of the biasing process itself, that is, of how the solution or outcome might have influenced reconstruction.

In this chapter, we will first describe the phenomenon of hindsight bias in more detail and how it is typically assessed. Next, we present a typical hindsight-bias experiment that can be used as a classroom demonstration, before we finally move to a discussion of theoretical explanations and applied perspectives.

The phenomenon of hindsight bias

As the examples in Text box 23.1 demonstrate, the designs, materials, and measures used in hindsight studies are quite diverse (see Pohl, 2007, for an overview). In this section, we first describe the typical experimental designs, measures of hindsight

TEXT BOX 23.1 EXAMPLES OF HINDSIGHT BIAS

1 After a political election, people's recollections of their pre-election estimates of the election outcome were on average closer to the actual outcome than the original estimates had been (e.g., Blank, Fischer, & Erdfelder, 2003).
2 Being asked after the end of a basketball match and in comparison to other people's pre-game predictions, spectators were too convinced that they would have correctly predicted the winning team (Pezzo, 2003).
3 In a cross-cultural Internet study on hindsight bias (Pohl, Bender, & Lachman, 2002), 227 participants received 20 numerical almanac-questions, with half of them accompanied by the solution (experimental items) and the other half not (control items). Participants were instructed to ignore the solutions (if given) and to generate their estimates independently. However, the mean distance between estimates and solutions was significantly smaller for experimental than for control items, namely 1.20 vs. 1.65 (given in *z*-score units).

bias, and the definition of hindsight bias. We then summarize pertinent empirical findings including individual differences.

Experimental designs

Most importantly, two different general experimental procedures were employed. In the *memory* design (Example 1 in Text box 23.1), people first give original judgments (OJs), then receive the correct judgment (CJ) to some items (experimental condition), but not to others (control condition), and are finally asked to recall all their original judgments (ROJs). In the *hypothetical* design (Examples 2 and 3), people receive the CJ to some items (experimental condition) right away and are then asked to give their OJ as if they did not know the CJ (hence the term "hypothetical"): "What would you have estimated before being informed about the CJ?" The remaining items are provided without the CJ (control condition) and participants are simply asked to generate an OJ. The hypothetical design reminds one of the similar *anchoring* design (see Chapter 12 in this volume and Pohl, Eisenhauer, & Hardt, 2003).

Measures

In the memory design (on which we focus from here on), hindsight bias measures are based on the original judgment (*OJ*), the correct judgment (*CJ*), and the recalled original judgment (*ROJ*). In typical applications, all these values will be numerical variables. For example, a person could answer the almanac question

"In which year was Mozart born?" with "1710" (as *OJ*) and later, after having received the solution (i.e., 1756 as *CJ*), recall "1740" (as *ROJ*). There are a variety of possible HB measures (see Pohl, 2007, for an overview). We distinguish here more traditional from model-based measures, which both contain measures of accuracy and distortion.

Traditional measures

Two measures of hindsight bias have been used frequently. One of these is Hell, Gigerenzer, Gauggel, Mall, and Müller's (1988) percentage measure, defined as

$$\text{HB\%} := \frac{ROJ - OJ}{CJ - OJ} 100 \tag{1}$$

for all $OJ \neq CJ$. For our example, we obtain a positive HB% = ([1740 – 1710] / [1756 – 1710]) · 100 = 65%. The measure has some appeal, as it becomes HB% = 0 when there is no hindsight bias and HB% = 100 when there is perfect hindsight bias. Note, however, that HB% is not a strict percentage measure as it can become larger than 100 (when $ROJ - OJ$ happens to be larger than $CJ - OJ$) and even less than zero (when *ROJ* and *CJ* differ from *OJ* in opposite directions). An even more serious concern are outliers, especially when the denominator $CJ - OJ$ is small. Thus, when using HB%, robust statistics such as the median should be used to avoid results that are biased by extreme values.

A better alternative is to use Pohl's (1992) difference measure (see also Fischer & Budescu, 1995)

$$\Delta\text{HB} := |OJ - CJ| - |ROJ - CJ| \tag{2}$$

which would give us ΔHB = |1710 – 1756| – |1740 – 1756| = 30 as the "difference of differences from the correct judgment". Again, positive values indicate hindsight bias in the sense that, compared to *OJ*, *ROJ* is closer to *CJ*. In general, the larger ΔHB, the larger the drift towards the CJ information evident in the posterior judgments. To ensure comparability of measures for different numerical scales typically associated with different questions, it is advisable to z-transform the *OJ*, *CJ,* and *ROJ* variables separately for each question (across all participants) prior to calculating the difference measure, resulting in Pohl's (1992) Δz measure

$$\Delta z := |z_{OJ} - z_{CJ}| - |z_{ROJ} - z_{CJ}|, \tag{3}$$

where z_x denotes the z-transform of x using the mean and the standard deviation calculated from all values of *OJ*, *CJ*, and *ROJ* involved. The Δz measure produces less extreme values than the HB% measure does, and it is also less sensitive to scale differences between different items, making it the measure of choice whenever an overall measure of hindsight bias is needed (cf. Pohl, 2007).

Overall measures of hindsight bias like the ones discussed above should not be applied indiscriminately to all items, but rather to cases of $ROJ \neq OJ$ only. Cases of perfect recall (with $ROJ = OJ$) should be analyzed separately. Pohl (2007) discussed several research findings where the overall index could be misleading otherwise. The reason is that several different processes could be involved, some of which affect the quality of memory (and thus the probability of perfect recall), whereas others influence the reconstruction of a forgotten judgment. It seems therefore advisable to have at least two separate measures, one for the quality of memory (percentage of perfect recall) and one for the shift of wrongly recalled judgments (e.g., HB% or ΔHB).

Model-based measures

Measures of overall hindsight bias do not tell us much about the cognitive processes that underlie a specific result. In other words, they provide a description but no explanation of the observed hindsight bias. To illustrate, a perfect recollection ($ROJ = OJ$) could stem from good memory or just a lucky hit during reconstruction (given that *OJ*s are often salient numbers that are easy to "hit"). Similarly, recalling the correct judgment as one's earlier given estimate ($ROJ = CJ$) could either result from a maximum shift towards *CJ* (i.e., to erroneously recall that $OJ = CJ$) or from a source confusion of numerical values (i.e., to remember the *OJ* but to confuse it with the *CJ*). Finally, as demonstrated in the preceding paragraph, *ROJ*s that are shifted towards the *CJ*s (or even beyond) could either stem from an unbiased reconstruction (uninfluenced by the *CJ*, just as in the no-CJ control condition) or from a biased reconstruction (influenced by the *CJ*).

To disentangle all these different cases, Erdfelder and Buchner (1998) introduced the HB13 model, a multinomial processing-tree model. The model is based on the frequencies of all possible rank orders of the numerical values *CJ*, *OJ*, and *ROJ*, separately for experimental and control items, and explains their distribution in terms of 13 cognitive parameters. The four central parameters of the model are r_C, r_E, b, and c (see Figure 23.1 for an abridged version). Parameters r_C and r_E represent the probabilities of a perfect recollection for control and experimental items, respectively. The difference $r_C - r_E$ reflects "recollection bias" (i.e., how much the presence of CJ impaired memory for OJ). Parameter b denotes "reconstruction bias" (i.e., the probability of a reconstruction biased by CJ) and parameter c "source confusion" (i.e., the probability of confusing CJ with OJ).

In order to estimate the HB13-model parameters, we first derive model equations by multiplying the probabilities that lead to a particular rank-order event. If more than one branch of a tree leads to the same rank-order event, the branch probabilities are summed (see Erdfelder & Buchner, 1998, for the details). Based on the whole set of model equations for experimental and control items, parameters of the model are then estimated using the maximum likelihood method. To test how good the model fits the data, the likelihood ratio chi-square statistic G^2 is used. Meanwhile, handy software exists for running the cumbersome model fitting and also

FIGURE 23.1 An abridged version of the HB13 multinomial model for control items (top tree) and experimental items (bottom tree) (adapted from the corrected version of Figure 1 from "Hindsight bias from 3 to 95 years of age" by Bernstein, Erdfelder, Meltzoff, Peria, and Loftus, 2011, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Volume 37, p. 381. Copyright © 2011 by the American Psychological Association. Adapted with permission of the publisher.).

for testing parameter differences, namely "multiTree" (Moshagen, 2010). The software can be downloaded free of charge for Windows, MacOS, and Linux platforms (from http://psycho3.uni-mannheim.de/Home/Research/Software/multiTree/). The software comes with demonstration files for model equation input (HB13.eqn) and sample data (HB13.mdt) that are easily adapted to other data sets.

The HB13 model has been applied to numerous studies to date. The model can be used to analyze the data either aggregated across participants or separately for each participant (given that enough items are included to allow meaningful estimation). Recently, the HB13 model was extended to hierarchical processing-tree models, in order to capture individual parameters and potential covariates that may impact these parameters (see, e.g., Coolin, Erdfelder, Bernstein, Thornton, & Thornton, 2015).

Definition

Generally, hindsight bias may be said to exist whenever the recalled judgment lies closer to the correct judgment (CJ) than the original one did, that is, whenever shift indices such as Δz are significantly larger than zero. However, an "improvement" in the judgments' quality may also simply result from thinking twice about the same question. Another source of "improvement" is given by possible regression effects (Pohl, 1995), because if the OJs are distributed around the CJ, then the chances of recollecting a judgment that deviates in the direction of the CJ or even beyond are

on average larger than of recollecting one that is in the opposite direction. Therefore, one needs to control presentation of CJs experimentally, in order to attribute hindsight bias to knowing the CJ and not to repeated thinking or regression effects. This can be done by presenting some of the questions together with the CJ (experimental items) and others without (control items). The definition of hindsight bias then needs to be extended accordingly to include that the shift index in the experimental condition should be positive and significantly larger than in the control condition.

Empirical evidence

Several overviews (Bernstein et al., 2016; Hawkins & Hastie, 1990; Pezzo, 2011; Pohl, 2007; Roese & Vohs, 2012) and meta-analyses (Christensen-Szalanski & Willham, 1991; Guilbault et al., 2004) summarized the main findings. An impressive result is that hindsight bias appears to be a robust and widespread phenomenon. It has been observed with many different experimental procedures in many different content domains. The two main designs were already introduced above. The materials could be sorted into three categories:

1 Assertions

   Participants judge the truth of assertions (like "Dallas is the capital of Texas"), rate their confidence, and later recall their judgments and confidence. In some cases, the assertion consists of a 2-alternative forced-choice question (like "Is absinthe a liquor or a precious stone?").

2 Events or episodes

   Participants receive an event or episode with two or more potential outcomes (e.g., the outcome of sporting events or political elections), rate the likelihood of each outcome, and later recall their judgments.

3 Quantitative estimates

   Participants generate numerical estimates in response to knowledge questions (like "How many books did Agatha Christie write?" or "What percentage of the earth's surface is covered by water?") and later try to recall these estimates.

Christensen-Szalanski and Willham (1991) identified in their meta-analysis (across 126 studies using probability judgments) three moderators, all of which relate to characteristics of the material: (1) Hindsight bias was smaller for familiar materials and larger for unfamiliar ones. (2) Hindsight bias was smaller when participants were told that an outcome occurred, rather than it did not occur (cf. the true–false asymmetry; Fischhoff, 1977; Hertwig et al., 1997). (3) Hindsight bias was smaller with events and episodes than with almanac questions (i.e., assertions and quantitative estimates).

The meta-analysis by Guilbault et al. (2004) (covering 95 studies) yielded four relevant moderators: (1) Hindsight bias was smaller for estimates on rating scales than

for probability estimates. (2) Hindsight bias was smaller for real-world events or case histories (i.e., outcomes of episodes) than for almanac questions (i.e., assertions and quantitative estimates), thus replicating the finding from Christensen-Szalanski and Willham (1991). (3) Hindsight bias was smaller after a positive or negative outcome than after a neutral outcome. (4) Manipulations that were intended to increase hindsight bias were more successful than manipulations intended to decrease hindsight bias.

Other typical results are as follows: (a) Hindsight bias was smaller in the memory than in the hypothetical design. (b) Hindsight bias increased with the retention interval between OJ and CJ, and decreased with the retention interval between CJ and ROJ. (c) Hindsight bias also decreased with a deeper encoding of one's OJs, and increased with a deeper encoding of the CJs. (d) Experts showed less hindsight bias than lay persons. All four findings may, however, mainly be due to the fact that correct recollections (i.e., $ROJ = OJ$) were often included in the reported shift measures, thus biasing these measures towards zero. By implication, shift measures tend to be smaller in conditions where perfect OJ recollections are more likely. When perfect recollections were taken out of the overall shift measure, the reported differences reduced considerably (see Pohl, 2007, for more details).

Extensions

Interestingly, hindsight bias has not only been observed for estimates of probabilities or numerical quantities in the verbal domain, but also for other materials. For example, several studies found hindsight bias with visual (e.g., Bernstein & Harley, 2007), auditory (e.g., Bernstein, Wilson, Pernat, & Meilleur, 2012), or even gustatory materials (e.g., Pohl, Schwarz, Sczesny, & Stahlberg, 2003). Additional examples are given in the section on "Applied perspectives" in this chapter.

Individual differences

Whereas manipulations of cognitive processes have been the main focus of hindsight-bias research, a few authors have addressed the question of individual differences (see Musch & Wagner, 2007, for a summary). For example, Musch (2003) found that hindsight bias (in the hypothetical design) was positively related to field dependence, to the tendency for a favourable self-presentation, to the participants' conscientiousness, and to their need for predictability and control (see also Campbell & Tesser, 1983). The same relations were, however, not substantial in the memory design. The motive of self-presentation has possibly gained the most attention. Musch and Wagner (2007) reported significant relations, especially regarding the "impression management" part of social desirability. In addition, Campbell and Tesser (1983) reported a positive correlation between hindsight bias and the amount of ego involvement. Musch and Wagner (2007, p. 64) summarized the findings and reported that "the variables that seem to be most strongly associated with the magnitude of hindsight bias are field dependence, intelligence, and self-presentational concerns," although not all studies that tested these relations found an effect.

Another variable that has been repeatedly investigated is expertise. For example, Christensen-Szalanski and Willham (1991) reported in their meta-analysis that hindsight bias was negatively related to the participants' familiarity with the material. In contrast, others failed to replicate this effect (Guilbault et al., 2004) and found either no effect (Pohl, 1992) or even positive effects of expertise (Musch & Wagner, 2007). Thus, the role of expertise remains unclear and presumably depends on further factors.

One further area that has recently attracted some research is age-related changes in hindsight bias. Several cross-sectional studies compared younger and older children (e.g., Pohl, Bayen, & Martin, 2010) and younger and older adults (e.g., Bayen, Erdfelder, Bearden, & Lozito, 2006), and one study even compared participants from 3 to 95 years of age (Bernstein et al., 2011). A common finding of these studies is that, across the lifespan, hindsight bias follows a U-shaped function, with younger children and older adults exhibiting the largest hindsight bias (Bayen, Pohl, Erdfelder, & Auer, 2007). Some studies (e.g., Coolin, Bernstein, Thornton, & Thornton, 2014; Groß & Bayen, 2015) tried to explain these differences with corresponding differences in episodic memory and inhibitory capacities, which are both known to follow inverted U-shaped functions across the lifespan. Coolin et al. (2014), for example, found that both measures indeed mediated age-related increases in the occurrence of hindsight bias, but not in its magnitude. In an extension of the HB13 model that allows capturing both individual differences and influences of covariates, Coolin et al. (2015), moreover, reanalyzed the older adults' data of Coolin et al. (2014) and found that better inhibitory control (that varied with age) was associated with less reconstruction bias. In a similar vein, Groß and Bayen (2015, p. 186) found that "older adults showed stronger hindsight bias and more fixations to the task-irrelevant CJ, indicating an age-related deficit in access control."

Hindsight-bias experiment

As a classroom demonstration of hindsight bias, we suggest an adapted version of the almanac-question paradigm employed in several studies (e.g., Hell et al., 1988; Pohl, 1998; Pohl et al., 2003; Pohl & Hell, 1996, Exp. 1). Text box 23.2 provides all necessary details to set up a simplified classroom experiment.

TEXT BOX 23.2 A CLASSROOM DEMONSTRATION OF HINDSIGHT BIAS

Method

Participants

Setting α and β error probabilities to .05, the experiment needs 19 participants to optimally test for a large effect size (Cohen's $d_z = .80$), or 45 participants to test for a medium effect size ($d_z = .50$) using a one-tailed matched-pairs *t*-test.

Material

Difficult numerical almanac questions from different domains should be used as materials (a sample set can be found in the Appendix). An example is "How many islands make up Hawaii?" Between 40 and 90 items are typically used.

Design

This experiment uses the memory design with only one within-subjects factor. Participants receive the CJ to half the questions (experimental condition), but not to the other half (control condition). Dependent measures are the percentage of correct recollections and the shift index Δz (as given above) after excluding correct recollections. We also suggest estimating recollection and reconstruction bias using the HB13 model.

Procedure

In Session 1, participants receive questions without CJs and are asked to estimate the answers (OJs) as exactly as possible. In order to avoid frustration on the side of the participants, instructions should stress that the questions were deliberately selected to be very difficult and that it is not expected that anyone knows the exact answer. At the end of the session, participants are kindly requested not to look up any of the questions and not to talk to other persons about their judgments.

One week later, in Session 2, participants receive the same questionnaire again, but now accompanied by the solutions (CJs) to half of the questions. Which half is given with and which one without CJs should be counterbalanced across participants, so that each question serves equally often as experimental and as control item. Participants are instructed to carefully read through the questions and – if supplied – the solutions, but to ignore them, when trying to recall as exactly as possible their OJs from last week. At the end, participants are debriefed about the purpose of the study and are given the remaining CJs if they wish.

Analysis

In order to take care of extreme judgments, a convenient procedure is to delete all judgments that are outside the median plus or minus three times the interquartile range for each question. The remaining data (*OJ*s, *CJ*s, and *ROJ*s) are then transformed into *z*-scores separately for each question.

Every *ROJ* that exactly matches the *OJ* is counted as a correct recollection. The absolute frequencies of correct recollections in the two conditions (experimental vs. control) can be compared with a paired *t*-test (with $df = N - 1$). For

the remaining pairs of *OJ*s and *ROJ*s, the shift indices Δz are computed and then averaged across all items in the same condition for each participant. The resulting two mean shift indices (experimental vs. control) can then be compared by running a paired *t*-test (with $df = N - 1$).

In addition, the HB13 model can be applied to the data. The necessary model equations are given in Erdfelder and Buchner (1998) and also provided in the standard multiTree sample files. The frequencies of the different rank orders of *OJ*, *CJ*, and *ROJ* can be obtained from the data. The analysis can then be run by using the software multiTree (Moshagen, 2010) that is freely available and that allows tests of model fit and parameter estimates.

Results

Correct recollections

After one week, approximately 25% of the OJs were correctly recalled (e.g., Pohl, 1992, 1998; Pohl et al., 2003). Using fewer items and thus probably boosting OJ memory, Pohl and Hell (1996, Exp. 2) found 36% correct recall after one week. Similarly, Hell et al. (1988), who requested reasons from their participants for half of their judgments, found a relatively high percentage of 35.1.

With respect to the comparison of experimental and control condition, differences in the probability of perfect recollections are typically absent or only small when these conditions are manipulated within-subjects (see Erdfelder et al., 2007, for an overview). Only a few studies reported statistically significant differences (e.g., Erdfelder & Buchner, 1998, Exp. 3; Pohl, 1998, Exp. 3; Pohl & Gawlik, 1995), showing that knowing the CJs can impair memory for one's OJs (i.e., recollection bias). Erdfelder et al. (2007) further explored this issue and identified conditions that influence recollection bias differentially depending on the relative proportions of experimental versus control items employed.

Hindsight bias

Comparing the mean shift indices for experimental and control items yielded significant differences in all of the cited studies (see Figure 23.2; the results of Hell et al., 1988, were not included, because the authors used a different measure). When the CJ was given (experimental items), ROJs were on average much closer to the CJs than the OJs, but they remained on average virtually unchanged relative to OJs when the CJ was not given (control items). Presenting the CJs obviously led to a strong hindsight bias. More precisely, ROJs in the experimental condition "moved" about one fifth of a standard deviation towards the CJs. This shift varied from $\Delta z = 0.11$ to 0.27 when looking at the individual studies (see Figure 23.2). This could, of course, be due to random error or to other variables that were manipulated in those studies but were omitted here for the sake of simplicity.

FIGURE 23.2 Mean shift indices Δz for experimental and control items in several studies in the memory design (ordered according to the values for experimental items; positive shift values indicate that the *ROJ*s were on average closer to the *CJ*s than the *OJ*s had been).

Model-based measures

To illustrate model-based measures of hindsight bias, we refer to Erdfelder et al. (2007), comprising 9 separate conditions (each with experimental and control items). Firstly, the HB13 model fitted all but one data set so that almost all parameter estimates could be interpreted and tested. Secondly, recollection parameters r_C and r_E varied from .17 to .52, depending on the specific condition. More importantly, recollection bias (i.e., $r_C > r_E$) was observed in only 2 of the 9 conditions, with decrements in performance of 13 and 14%. But note that these experimental conditions were especially designed to boost recollection bias. Thirdly, all conditions showed a substantial reconstruction bias, as indicated by significant estimates of parameter b (ranging from .15 to .41). Fourthly, the probability of source confusions is typically rather low (i.e., $c < .10$; cf. Erdfelder & Buchner, 1998), but it may play a larger role in children or older adults (see, e.g., Bayen et al., 2006; Bernstein et al., 2011; Pohl et al., 2010).

Discussion

Firstly, about one fourth of all estimates were correctly recalled. This percentage was similar for experimental and control items, except in a few cases in which the percentage was slightly, but statistically significantly, lower in the experimental condition, thus exhibiting recollection bias (Erdfelder et al., 2007). As the second stable result, all studies provided clear evidence of hindsight bias. If the CJs were given prior to or simultaneously with the recall attempt, judgments were recalled as being closer to the CJs than they had actually been, as indicated by a positive

shift index Δz. In contrast, judgments in the control condition were not recalled with a systematic bias. The stability of this result conforms to the finding of Christensen-Szalanski and Willham (1991), who found among 122 experiments on hindsight bias only 6 without the typical effect. It appears that participants are simply not able to avoid utilizing the CJ in reconstructing their earlier knowledge state (Fischhoff, 1975; Hawkins & Hastie, 1990; Hoffrage, Hertwig, & Gigerenzer, 2000). Model-based results confirmed this impression and further identified reconstruction bias as the main source of hindsight bias, whereas recollection bias and source confusions were less influential (cf. Erdfelder et al., 2007; Erdfelder & Buchner, 1998).

Theoretical accounts

Despite the impressive evidence for the existence of hindsight bias, the underlying mechanisms that are responsible for hindsight bias are still under debate (cf. Bernstein et al., 2016). Most approaches focus on specific cognitive processes that are caused by encoding the CJs, namely an impairment of memory for the OJs (recollection bias) and, if the OJ cannot be accessed in the recall attempt, a contaminated rejudgment process (reconstruction bias). In addition, some approaches point out the role of metacognitions and motivational processes. Before presenting all these approaches in more detail, though, we focus on a framework that distinguishes different types or components of hindsight bias.

Components view

Blank, Nestler, and von Collani (2008) advocated a more fine-grained view on hindsight bias. They suggested distinguishing three types or components of hindsight bias: memory distortion, impression of necessity (or inevitability), and impression of foreseeability. Nestler, Blank, and Egloff (2010) further showed that specific manipulations may affect only one or the other of the components, that is, that the three components can in principle be dissociated. This approach has been helpful to better understand seemingly contradictory findings that actually focussed on separate components of hindsight bias. Roese and Vohs (2012) suggested to view the three components in a hierarchical framework, with memory distortions representing the lowest and basic level (influenced by recollection and knowledge updating), inevitability as the medium level (influenced by sense-making), and foreseeability as the highest level (influenced by metacognitions and motivational processes).

Cognitive processes

The ongoing debate on how to explain hindsight bias through cognitive processes has mainly focussed on two mechanisms: (1) Does encoding the CJs impair memory for one's OJs? (2) Does encoding the CJs bias the necessary rejudgment

process, once the OJ cannot be retrieved? These two mechanisms are known as recollection and reconstruction bias, respectively, and form the theoretical basis of the HB13 model described above (Erdfelder & Buchner, 1998).

Recollection and reconstruction bias

The recollection-reconstruction theory of hindsight judgments assumes a two-stage ROJ process (cf. Hawkins & Hastie, 1990; Stahlberg & Maass, 1998). In the first stage, participants attempt to retrieve their OJ. For control items (without CJ), successful recollection should depend on how well the OJ was encoded and consolidated as well as how easily it can be retrieved later on. For experimental items (with CJ given), recollection depends in addition on whether CJ knowledge interferes with OJ consolidation or retrieval, perhaps resulting in poorer recollection for experimental compared to control items (i.e., recollection bias). If recollection fails, participants enter a second stage in which they try to repeat the original judgment process, based on available context information. For experimental items, however, this process may suffer from over-reliance on outcome knowledge and thus result in a biased reconstruction of the OJ (i.e., reconstruction bias). Several specific mechanisms have been proposed to account for this bias (see Christensen-Szalanski & Willham, 1991; Guilbault et al., 2004; Hawkins & Hastie, 1990):

- Tversky and Kahneman (1974), for example, proposed a cognitive heuristic for the anchoring effect, named anchoring and insufficient adjustment (cf. Chapter 12), which may also apply here with the CJ serving as an "anchor" in the reconstruction process. The reconstruction starts from this anchor and then adjusts into the direction of one's knowledge base. This adjustment process, however, may stop too early, for example, at the point where the first plausible value is reached, thus leading to a biased reconstruction.
- Fischhoff (1975) assumed an immediate and irreversible assimilation of the CJ into one's knowledge base. Note, however, that this process should boost both reconstruction and recollection bias because OJ retrievability will certainly be hampered by CJ assimilation.
- Focussing on learning and efficient use of cue knowledge, the CJ may also be used to update hitherto uncertain or unknown cue knowledge (see Hertwig, Fanselow, & Hoffrage, 2003; Hoffrage et al., 2000), thus again irreversibly altering one's knowledge.
- In terms of selective activation during encoding (Pohl et al., 2003), the CJ may activate – by way of spreading activation – knowledge that is most similar to it (cf. the confirmation bias in Chapter 5 and the explanation of anchoring effects in Chapter 12), thus leading to a retrieved set that is most likely biased rather than representative of the given knowledge base.
- In addition, the CJ may be used as retrieval cue when in fact it should be ignored, leading to a biased sampling of available information from memory (cf. Pohl et al., 2003; Schwarz & Stahlberg, 2003).

Previous research has shown that reconstruction bias (based on one or more of these mechanisms) is the major determinant of hindsight bias, although recollection bias may also play a role under some conditions (Erdfelder et al., 2007) or for special populations such as older adults (Bayen et al., 2006; Coolin et al., 2015). Which of the specific processes is responsible for reconstruction bias may depend on the type of task (like the focussed hindsight-bias component) and the experimental procedure (like memory versus hypothetical design).

Computational models

Pohl et al. (2003) introduced a cognitive process model, named SARA (*Selective Activation and Reconstructive Anchoring*), that incorporates two of the possible mechanisms discussed above that could lead to hindsight bias: Encoding the CJ may change the association strengths within the knowledge base (selective activation), and at retrieval, the CJ may bias the memory search towards CJ-related entries (biased sampling). SARA has been successfully implemented as a computer program that allows simulating empirical data as well as predicting new findings. Hoffrage et al. (2000, see also Hertwig et al., 2003) introduced another computational model, termed RAFT (*Reconstruction After Feedback with Take-the-best*), that incorporates another mechanism for reconstruction bias, namely, knowledge updating (see Blank & Nestler, 2007, for a comparison and discussion of both models).

Metacognitions

Several studies showed that metacognitions might significantly moderate and even eliminate or reverse hindsight bias. Sanna and Schwarz (2007) presented an integrative framework of metacognitive influences on hindsight bias. Among the studied mechanisms are overconfidence, processing fluency, and surprise.

Overconfidence in one's knowledge typically exists for rather difficult items (see Chapter 16). Accordingly, Hoch and Loewenstein (1989) found hindsight bias only for items of high and medium difficulty, but not for easy ones. Similarly, Schwarz and Stahlberg (2003, Exp. 2) manipulated participants' assumptions of how close their original judgments allegedly were to the CJs. The authors found that hindsight bias was significantly larger when participants believed that their judgments had been rather good than when they thought that they had been rather poor. With respect to fluency, the fluent processing of the CJ may erroneously suggest that it is familiar and thus increase hindsight bias (Werth & Strack, 2003). Bernstein and Harley (2007) provided further support for misattributions of both perceptual and conceptual fluency as determinants of hindsight bias.

Pohl (1998, Exp. 2 and 3) found that CJs that were labelled as "another person's estimate" and that were additionally considered implausible did not lead to hindsight bias. This result is possibly based on feelings of surprise (which is related to feelings of implausibility), because several studies found that hindsight bias was often reduced, absent, or even reversed when the CJ was considered surprising (Müller &

Stahlberg, 2007; Pezzo, 2003). Instead of "I knew that all along," participants might then experience a feeling of "I would never have known that." Pezzo (2003) suggested a sense-making process as the underlying mechanism (see also Müller & Stahlberg, 2007). He assumed that a CJ might elicit an initial level of surprise that then triggers attempts of sense-making, that is, attempts to integrate the CJ into one's knowledge base. If that succeeds, thereby reducing or eliminating the initially felt surprise, hindsight bias will occur and be rather large. But if sense-making fails, feelings of surprise remain and prevent one from knowledge integration, so that hindsight bias will be absent or even reversed. The causal-model theory (Nestler, Blank, & von Collani, 2008a, 2008b; see also Yopchick & Kim, 2012) describes one potential process of how sense-making can be achieved, namely by establishing causal connections between context information and the outcome. The effects of surprise may, however, differ for the respective components of hindsight bias (Nestler & Egloff, 2009).

Motivational processes

The discussion of individual differences (see above) already included some motivational factors. Roese and Vohs (2012, p. 416) summarized the research as showing that "motivational factors fuel hindsight bias (particularly foreseeability) in two ways: first, by way of a need to see the world as orderly and predictable and, second, by way of a need to protect and enhance one's self-esteem." The first factor is known as "need for closure" (or need for predictability and control), which increases hindsight bias (Campbell & Tesser, 1983; Musch, 2003) and is presumably related to illusions of control (see Chapter 7). The second factor, self-presentational concerns, has received quite some attention, mainly because when informed about the hindsight-bias phenomenon, most laypersons assume as an explanation that people simply try to appear smarter than they are (cf. Hawkins & Hastie, 1990). Thus, self-presentation should be a major factor, but several studies did not find a relation to hindsight bias (see Musch & Wagner, 2007). The reason probably is that the typically used materials (like highly difficult almanac questions) did not invoke much self-concern. The picture, however, changed when more self-relevant materials were used, especially in studying foreseeability of positive and negative outcomes. For example, to manage disappointment after a negative outcome, people might engage in "retroactive pessimism" (Tykocinski & Steinberg, 2005), leading to increased hindsight bias in foreseeability (like "I saw it coming"). In contrast, in terms of "defensive processing" (Pezzo & Pezzo, 2007), that is, to avoid feeling responsible for the negative outcome, hindsight bias is reduced (like "I could never have expected that"; see, e.g., Mark & Mellor, 1991). To integrate these seemingly contradictory results, Pezzo and Pezzo (2007) suggested a "motivated sense-making" model (extending Pezzo's, 2003, model). They assumed that it depends on whether external or internal reasons can be used to make sense of an unexpected negative outcome, and thus how much control a person may feel to have: With external reasons (and less control), people may

adjust their self-esteem and claim more foreseeability (retroactive pessimism), but with internal reasons (and more control), they may protect their self-esteem and claim less foreseeability (defensive processing).

Applied perspectives

Hindsight bias has several practical implications. For example, consider a physician who is asked for a second opinion on a serious diagnosis but knows the first one. Many studies have demonstrated that new and allegedly independent judgments are most likely biased towards the already available ones. In other words, second judgments are less independent than we and the judges themselves like to think that they are. This could have serious consequences, especially if the first judgment is poor, dubious, arbitrary, or simply wrong. A number of studies have investigated such consequences in applied contexts, for example, in medical (Arkes, 2013), legal (Harley, 2007), economic (Biais & Weber, 2009), or everyday decision making (Pieters, Baumgartner, & Bagozzi, 2006). Louie, Rajan, and Sibley (2007) provided a number of real-world examples of hindsight bias and discussed their consequences (see also Pezzo, 2011).

A general problem with hindsight bias might be that feeling wiser after the outcome is known could also lead us to a too optimistic evaluation of our prior knowledge state. But if we consider ourselves more knowledgeable than we really were, we could easily overestimate our abilities for similar situations in the future. For example, having understood – in hindsight – how an accident came about, could lull us into a false sense of security. The opposite problem occurs if, for example, relatives and friends of suicide victims feel guilty because they overestimate their prior chances to predict, and possibly to prevent, the suicide. Hindsight bias may be at least partially responsible for this inflated feeling of guilt. Similar influences have been discussed in connection with depression or the chronification of pain. Louie et al. (2007) gave an overview of such negative, but also positive, consequences of hindsight bias. Roese and Vohs (2012) discussed as the main negative consequences myopia (i.e., the erroneous attribution of a cause to an outcome) and overconfidence (in one's knowledge or abilities). The potential impediment to learning has presumably received the most attention (see, e.g., Biais & Weber, 2009; Bukszar & Connolly, 1988; Hoch & Loewenstein, 1989). As an example, consider Pezzo and Pezzo (2007), who discussed that retroactive pessimism might aid in stabilizing the planning fallacy, because unexpected outcomes (like the failure to finish a task in time) are attributed to less controllable external factors.

Conclusions

Hindsight bias is an extremely robust phenomenon that can easily be demonstrated, leading some authors to question the nature of this effect. Accordingly, hindsight bias may not be viewed as a bothersome consequence of a "faulty" information processing system, but rather as an unavoidable by-product of an evolutionary evolved

function, namely adaptive learning (e.g., Campbell & Tesser, 1983; Hertwig et al., 2003; Hoffrage et al., 2000; Pohl et al., 2002). According to this view, hindsight bias is seen as the consequence of our most valuable ability to update previously held knowledge. This may be seen as a necessary process in order to prevent memory overload and thus to maintain normal cognitive functioning. Besides, updating allows us to keep our knowledge more coherent and to draw better inferences. Of course, there are situations in which one may wish to exactly assess his or her previous knowledge state. But these cases might be relatively rare in real life, so that the disadvantage of a biased memory reconstruction (like that exemplified in hindsight bias) is probably more than outweighed by the benefits of adaptive learning. Going beyond this view, some authors even suggested that hindsight bias itself represents an adaptive function (see Pohl et al., 2002). Being in hindsight, thus, may also be advantageous (cf. Chapter 25). This is an intriguing, but presently rather speculative idea that certainly deserves further research.

Summary

- Hindsight bias occurs when persons in hindsight (e.g., after an outcome is known) overestimate what they had (or would have) known in foresight.
- Hindsight bias is a very robust phenomenon that cannot be reduced intentionally.
- As cognitive explanations, either impairment of memory (recollection bias) or a biased rejudgment process (reconstruction bias) might be responsible. Overall, reconstruction bias has more impact than recollection bias.
- Metacognitions (like surprise) and motivational factors (like impression management) may moderate the strength of hindsight bias.
- Individual differences (like field dependence) play only minor roles.
- Studies on age-related differences suggest a U-shaped lifespan function, with children and older adults showing the largest hindsight bias.
- The phenomenon has important implications for several applied problems.

Further reading

Hawkins and Hastie (1990) gave an excellent early overview about research on hindsight bias. Christensen-Szalanski and Willham (1991) and Guilbault et al. (2004) reported instructive meta-analyses. Two special issues are devoted to hindsight bias, one in *Memory* (Hoffrage & Pohl, 2003) and the other in *Social Cognition* (Blank, Musch, & Pohl, 2007), both covering theoretical, empirical, as well as applied aspects of hindsight bias. More recently, Bernstein et al. (2016), Pezzo (2011), and Roese and Vohs (2012) provided comprehensive overviews.

References

Arkes, H. R. (2013). The consequences of the hindsight bias in medical decision making. *Current Directions in Psychological Science, 22*(5), 356–360.

Bayen, U. J., Erdfelder, E., Bearden, J. N., & Lozito, J. P. (2006). The interplay of memory and judgment processes in effects of aging on hindsight bias. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*(5), 1003–1018.

Bayen, U. J., Pohl, R. F., Erdfelder, E., & Auer, T.-S. (2007). Hindsight bias across the life span. *Social Cognition, 25*(1), 83–97.

Bernstein, D. M., Aßfalg, A., Kumar, R., & Ackerman, R. (2016). Looking backward and forward on hindsight bias. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 289–304). New York: Oxford University Press.

Bernstein, D. M., Erdfelder, E., Meltzoff, A. N., Peria, W., & Loftus, G. R. (2011). Hindsight bias from 3 to 95 years of age. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 378–391.

Bernstein, D. M., & Harley, E. M. (2007). Fluency misattribution and visual hindsight bias. *Memory, 15*(5), 548–560.

Bernstein, D. M., Wilson, A. M., Pernat, N. L. M., & Meilleur, L. R. (2012). Auditory hindsight bias. *Psychonomic Bulletin & Review, 19*(4), 588–593.

Biais, B., & Weber, M. (2009). Hindsight bias, risk perception, and investment performance. *Management Science, 55*(6), 1018–1029.

Blank, H., Fischer, V., & Erdfelder, E. (2003). Hindsight bias in political elections. *Memory, 11*(4–5), 491–504.

Blank, H., Musch, J., & Pohl, R. F. (2007). The hindsight bias [Special issue]. *Social Cognition, 25*(1).

Blank, H., & Nestler, S. (2007). Cognitive process models of hindsight bias. *Social Cognition, 25*(1), 132–146.

Blank, H., Nestler, S., & von Collani, G. (2008). How many hindsight biases are there? *Cognition, 106*(3), 1408–1440.

Bukszar, E. W., & Connolly, T. (1988). Hindsight bias and strategic choice: Some problems in learning from experience. *Academy of Management Journal, 31*(3), 628–641.

Campbell, J. D., & Tesser, A. (1983). Motivational interpretations of hindsight bias: An individual difference analysis. *Journal of Personality, 51*, 605–620.

Christensen-Szalanski, J. J. J., & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes, 48*, 147–168.

Coolin, A., Bernstein, D. M., Thornton, A. E., & Thornton, W. L. (2014). Age differences in hindsight bias: The role of episodic memory and inhibition. *Experimental Aging Research, 40*(3), 357–374.

Coolin, A., Erdfelder, E., Bernstein, D. M., Thornton, A. E., & Thornton, W. L. (2015). Explaining individual differences in cognitive processes underlying hindsight bias. *Psychonomic Bulletin & Review, 22*, 328–348.

Erdfelder, E., Brandt, M., & Bröder, A. (2007). Recollection biases in hindsight judgments. *Social Cognition, 25*(1), 114–131.

Erdfelder, E., & Buchner, A. (1998). Decomposing the hindsight bias: A multinomial processing tree model for separating recollection and reconstruction in hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(2), 387–414.

Fischer, I., & Budescu, D. V. (1995). Desirability and hindsight bias in predicting results of a multi-party election. In J.-P. Caverni, M. Bar-Hillel, & F. H. Barron (Eds.), *Contributions to decision making I* (pp. 193–212). Amsterdam: Elsevier.

Fischhoff, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1*, 288–299.

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance, 3*(2), 349–358.

Groß, J., & Bayen, U. J. (2015). Hindsight bias in younger and older adults: The role of access control. *Aging, Neuropsychology, & Cognition, 22*(2), 183–200.

Guilbault, R. L., Bryant, F. B., Brockway, J. H., & Posavac, E. J. (2004). A meta-analysis of research on hindsight bias. *Basic and Applied Social Psychology, 26*(2–3), 103–117.

Harley, E. M. (2007). Hindsight bias in legal decision making. *Social Cognition, 25*(1), 48–63.

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin, 107*, 311–327.

Hell, W., Gigerenzer, G., Gauggel, S., Mall, M., & Müller, M. (1988). Hindsight bias: An interaction of automatic and motivational factors? *Memory & Cognition, 16*, 533–538.

Hertwig, R., Fanselow, C., & Hoffrage, U. (2003). Hindsight bias: How does forsesight knowledge and its processing affect our reconstruction of the past? *Memory*, *11*, 357–377.

Hertwig, R., Gigerenzer, G., & Hoffrage, U. (1997). The reiteration effect in hindsight bias. *Psychological Review, 104*(1), 194–202.

Hoch, S. J., & Loewenstein, G. F. (1989). Outcome feedback: Hindsight and information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(4), 605–619.

Hoffrage, U., Hertwig, R., & Gigerenzer, G. (2000). Hindsight bias: A by-product of knowledge-updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 566–581.

Hoffrage, U., & Pohl, R. F. (Eds.). (2003). Hindsight bias [Special issue]. *Memory*, *11*(4/5).

Louie, T. A., Rajan, M. N., & Sibley, R. E. (2007). Tackling the Monday-morning quarterback: Applications of hindsight bias in decision-making settings. *Social Cognition, 25*(1), 32–47.

Mark, M. M., & Mellor, S. (1991). Effect of self-relevance of an event on hindsight bias: The forseeability of a layoff. *Journal of Applied Psychology, 76*(4), 569–577.

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42*(1), 42–54.

Müller, P. A., & Stahlberg, D. (2007). The role of surprise in hindsight bias: A metacognitive model of reduced and reversed hindsight bias. *Social Cognition, 25*(1), 165–184.

Musch, J. (2003). Personality differences in hindsight bias. *Memory*, *11*, 473–489.

Musch, J., & Wagner, T. (2007). Did everybody know it all along? A review of individual differences in hindsight bias. *Social Cognition, 25*(1), 64–82.

Nestler, S., Blank, H., & Egloff, B. (2010). Hindsight ≠ hindsight: Experimentally induced dissociations between hindsight components. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1399–1413.

Nestler, S., Blank, H., & von Collani, G. (2008a). Hindsight bias and causal attribution: A causal model theory of creeping determinism. *Social Psychology, 39*(3), 182–188.

Nestler, S., Blank, H., & von Collani, G. (2008b). Hindsight bias doesn't always come easy: Causal models, cognitive effort, and creeping determinism. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(5), 1043–1054.

Nestler, S., & Egloff, B. (2009). Increased or reversed? The effect of surprise on hindsight bias depends on the hindsight component. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1539–1544.

Pezzo, M. V. (2003). Surprise, defense, or making sense: What removes the hindsight bias? *Memory*, *11*, 421–441.

Pezzo, M. V. (2011). Hindsight bias: A primer for motivational researchers. *Social & Personality Psychology Compass, 5*(9), 665–678.

Pezzo, M. V., & Pezzo, S. P. (2007). Making sense of failure: A motivated model of hindsight bias. *Social Cognition, 25*(1), 147–164.

Pieters, R., Baumgartner, H., & Bagozzi, R. (2006). Biased memory for prior decision making: Evidence from a longitudinal field study. *Organizational Behavior and Human Decision Processes, 99*(1), 34–48.

Pohl, R. F. (1992). Der Rückschau-Fehler: Systematische Verfälschung der Erinnerung bei Experten und Novizen [Hindsight bias: Systematic distortions of the memory of experts and laymen]. *Kognitionswissenschaft, 3*, 38–44.

Pohl, R. F. (1995). Disenchanting hindsight bias. In J.-P. Caverni, M. Bar-Hillel, F. H. Barron, & H. Jungermann (Eds.), *Contributions to decision making* (pp. 323–334). Amsterdam: Elsevier.

Pohl, R. F. (1998). The effects of feedback source and plausibility on hindsight bias. *European Journal of Cognitive Psychology, 10*, 191–212.

Pohl, R. F. (2004). Hindsight bias. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (pp. 363–378). Hove, UK: Psychology Press.

Pohl, R. F. (2007). Ways to assess hindsight bias. *Social Cognition, 25*(1), 14–31.

Pohl, R. F., Bayen, U. J., & Martin, C. (2010). A multi-process account of hindsight bias in children. *Developmental Psychology, 46*(5), 1268–1282.

Pohl, R. F., Bender, M., & Lachmann, G. (2002). Hindsight bias around the world. *Experimental Psychology*, *49*, 270–282.

Pohl, R. F., Eisenhauer, M., & Hardt, O. (2003). SARA – A cognitive process model to simulate anchoring effect and hindsight bias. *Memory*, *11*, 337–356.

Pohl, R. F., & Gawlik, B. (1995). Hindsight bias and misinformation effect: Separating blended recollections from other recollection types. *Memory, 3*(1), 21–55.

Pohl, R. F., & Hell, W. (1996). No reduction of hindsight bias with complete information and repeated testing. *Organizational Behavior and Human Decision Processes, 67*, 49–58.

Pohl, R. F., Schwarz, S., Sczesny, S., & Stahlberg, D. (2003). Hindsight bias in gustatory judgments. *Experimental Psychology, 50*(2), 107–115.

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science,* 7(5), 411–426.

Sanna, L. J., & Schwarz, N. (2007). Metacognitive experiences and hindsight bias: It's not just the thought (content) that counts! *Social Cognition, 25*(1), 185–202.

Schwarz, S., & Stahlberg, D. (2003). Strength of the hindsight bias as a consequence of metacognitions. *Memory*, *11*, 395–410.

Stahlberg, D., & Maass, A. (1998). Hindsight bias: Impaired memory or biased reconstruction? In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 8, pp. 105–132). Chichester: Wiley.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Tykocinski, O. E., & Steinberg, N. (2005). Coping with disappointing outcomes: Retroactive pessimism and motivated inhibition of counterfactuals. *Journal of Experimental Social Psychology, 41*, 551–558.

Werth, L., & Strack, F. (2003). An inferential approach to the knew-it-all-along-phenomenon. *Memory, 11*(4–5), 411–419.

Yopchick, J. E., & Kim, N. S. (2012). Hindsight bias and causal reasoning: A minimalist approach. *Cognitive Processing, 13*, 63–72.

Authors' note

The cited studies of the authors have been supported by various grants from the Deutsche Forschungsgemeinschaft.

APPENDIX

Almanac questions and solutions [in brackets]

1 What percentage of the surface of the earth consists of water? [71%]
2 What is the mean life expectancy of a canary in years? [25 years]
3 How often on average does the heart of a mouse beat in one minute? [650 times]
4 How many times larger is the diameter of the planet Jupiter compared to the Earth? [11 times]
5 How many different kind of insects inhabit the Antarctic? [52 kinds]
6 How many keys does a piano have? [88]
7 How long is the life expectancy of a healthy red blood corpuscle (in days)? [120 days]
8 How many teeth has a dog? [42]
9 How many prime numbers does the interval between 1 and 1,000 contain? [168]
10 How many days does it take the sun to fully rotate around its axis? [25.4 days]
11 When did the first manned space flight take place? [1961]
12 How many bones does a human have? [214]
13 How old was Mahatma Gandhi when he was shot? [78 years]
14 What is the length of pregnancy of a rat in days? [25 days]
15 How old is the oldest tree on earth (in years)? [4,600 years]
16 How many crime novels were written by Agatha Christie? [67]
17 In what year did the first Olympiad of the modern era take place? [1896]
18 What is the maximum length of a total solar eclipse (in minutes)? [7 min]
19 What is the radius of the earth (in kilometres)? [6,378 km]
20 In what year did Leonardo Da Vinci paint the "Mona Lisa"? [1503]

21 How high is the Statue of Liberty in New York (in meters)? [93 m]
22 What is the average length of a human adult's kidney (in centimetres)? [12 cm]
23 In what year was a human heart transplanted for the first time? [1967]
24 What is the mean body temperature of a ground hog (in °C)? [31.7 °C]
25 How long is an international sea mile in meters? [1852 m]
26 How old was Martin Luther King when he was shot? [39 years]
27 What is the height of Mount Everest (in meters)? [8,848 m]
28 In what year did the Roman emperor Nero commit suicide? [68 AD]
29 How high is the Cheops pyramid in Egypt (in meters)? [147 m]
30 What is the airline distance between New York and Berlin, Germany? [6,374 km]
31 In what year was the North Atlantic Treaty Organization (NATO) founded? [1949]
32 What is the weight of a regular tennis ball (in grams)? [57 g]
33 How long is the mean pregnancy period of a female elephant in days? [631 days]
34 What is the average winter temperature in the Antarctic (in °C)? [-68 °C]
35 How many star constellations (including the zodiac signs) are officially recognized? [88]
36 When did Albert Einstein first visit the USA? [1921]
37 At what temperature does tin ore begin to melt (in °C)? [232 °C]
38 How long is the Great Wall of China (in kilometres)? [2,450 km]
39 How many islands make up Hawaii? [132]
40 How long is the Panama canal (in kilometres)? [81.6 km]

24

RETRIEVAL-INDUCED FORGETTING

Oliver Kliegl and Karl-Heinz T. Bäuml

Does retrieval of a specific episode in memory affect the remembering of related episodes? For instance, is it advisable, immediately prior to an exam, to retrieve some of the course material to co-activate the other material, or does such selective memory retrieval rather have a detrimental effect on memory of the related information? Intuitively, one may assume that the effect of selective memory retrieval should mostly promote retrieval of related episodes, and indeed, prominent models of memory from the 1960s and 1970s are well in line with such a notion, suggesting that activation of a specific episode in memory facilitates activation of related episodes (Collins & Loftus, 1975). Moreover, empirical evidence from eyewitness memory research suggests that active retrieval of some previously experienced episodes or events can benefit memory of other episodes or events (Geiselman, Fisher, MacKinnon, & Holland, 1985). However, since the early 1970s, numerous studies have also found evidence for the seemingly paradoxical feature of human memory that, in a variety of situations, recalling previously learned information can impede our ability to recall related information. Evidence for such retrieval-induced forgetting (RIF) has arisen mainly from two experimental tasks: the output-interference task (Roediger, 1974) and the retrieval-practice task (Anderson, Bjork, & Bjork, 1994).

Experimental tasks

RIF was first examined employing the output-interference task, in which it is shown that RIF can occur in the course of a recall test. In this task, participants are often asked to study a categorized list and, after a delay that is usually filled with an unrelated distractor task, to recall the list's items. At test, the list's category labels are successively provided as retrieval cues and participants are asked to recall a category's items. The typical finding with this task is that recall performance declines as a function of the category's testing position, indicating that the act of retrieving some items from

the study list decreases chances for the list's other items to be successfully retrieved (Roediger, 1974). Most of the more recent research on RIF, however, has employed the retrieval-practice task, which was first introduced by Anderson et al. (1994). In this task, a subset of previously studied material is repeatedly retrieved, and the effect of this manipulation on later recall of the practiced and unpracticed material is examined. Often participants study items from different semantic categories (e.g., FRUIT–Orange, FRUIT–Banana, FURNITURE–Table); in a subsequent retrieval-practice phase, they are asked to repeatedly retrieve half of the items from half of the categories using a word stem completion task (e.g., FRUIT–Or___). After a delay, typically filled with an unrelated distractor task, recall performance for all initially studied items is tested. The selective retrieval practice phase between study and test creates three types of items, usually labeled Rp+ items, Rp− items, and control items. Rp+ items refer to practiced items from practiced categories (Orange), Rp− items refer to unpracticed items from practiced categories (Banana), and control items refer to items from unpracticed categories (see Figure 24.1). Figure 24.1 also illustrates the

FIGURE 24.1 Retrieval-induced forgetting. (A) Experimental procedure. Participants study a categorized item list. In a subsequent retrieval-practice phase, half of the items from half of the studied categories are repeatedly retrieved. On the final memory test, participants are asked to recall all previously studied items. (B) Typical pattern of results. Practiced (Rp+) items show higher recall rates and unpracticed (Rp–) items from retrieval-practiced categories show lower recall rates than control (C) items from not retrieval-practiced categories.

typical finding with this task. Recall of the Rp+ items is improved on the final test relative to the control items, and, more important, recall of the Rp− items is impaired relative to the control items, which reflects the RIF finding.

RIF is a very general phenomenon and has been observed over a wide range of study materials – like verbal, visuospatial, or autobiographical materials – and also over a variety of experimental settings. For instance, Shaw, Bjork, and Handal (1995) found RIF in eyewitness memory by showing that repeated questions on specific details of a previously observed robbery scene improved subsequent recall of these details but reduced subsequent recall of related details. In another study employing an impression formation task, participants were shown personality characteristics of two men, Bill and John (e.g., JOHN–creative, BILL–romantic, JOHN–skillful), before participants' memory for half of the personality traits of one of the two characters was tested (e.g., JOHN–cr___). When participants were asked to recall all personality traits that they had previously seen in a subsequent surprise recall test, typical RIF arose, that is, recall of the unpracticed traits from the practiced person (e.g., John) was impaired relative to the unpracticed traits from the other, unpracticed person (e.g., Bill; Macrae & MacLeod, 1999). In a recent series of experiments, RIF has even been shown to arise in social groups. In these studies, pairs of individuals studied a list of items, and, subsequently, only a single member of the pair (the "speaker") selectively retrieved some of the items, while the other member (the "listener") simply listened to the speaker's responses, monitoring their accuracy. When the listener was subsequently asked to recall the remaining items, memory for the related items was impaired (for a review, see Hirst & Echterhoff, 2012).

A classroom demonstration of retrieval-induced forgetting

As a classroom demonstration of retrieval-induced forgetting, we suggest a version of the retrieval-practice task employed by Anderson et al. (1994). Text box 24.1 provides all necessary details to set up a simplified classroom experiment.

TEXT BOX 24.1 CLASSROOM EXPERIMENT

Method

Participants

A recent meta-analysis revealed that effect sizes in the retrieval-practice task are usually between small and medium (see Murayama, Miyatsu, Buchli, & Storm, 2014). Estimating the required number of participants to optimally test for a small to medium effect size (d = .4) with the help of the program G*power (Faul, Erdfelder, Lang, & Buchner, 2007) reveals that the experiment needs 40 participants when α is set at 0.05 and β is set at .20.

Material and Design

A list of 48 items, consisting of 6 exemplars from 8 semantic categories, are used as material. Within each category, each item has a unique initial letter (see Appendix). The experiment is conducted in 3 main phases. In the study phase, participants study all of the items and then, in the retrieval-practice phase, half of the items from 4 of the 8 semantic categories are practiced, before the final test phase, in which all items from the study phase have to be recalled. This design creates 3 types of items: The practiced items from the 4 practiced categories define the Rp+ items, the unpracticed items from the 4 practiced categories define the Rp– items, and the items from the 4 unpracticed categories define the control items.

Procedure

Prior to the experiment, participants are informed that they should study each to-be-presented word together with its category label. Following the initial instruction, all 48 items are shown to the participants via presentation software (such as Microsoft PowerPoint or Apple Keynote). Each category-exemplar pair (e.g., FRUIT–Orange) is presented on an individual slide for 5 s per pair. The order of the exemplars consists of a sequence of 6 blocks, consisting of one exemplar from each of the 8 categories, with the restriction that no two exemplars from the same category are presented on adjacent slides. In a subsequent 1 min distractor phase, participants solve simple arithmetic tasks (e.g., [8 + 32] × 9 = ?), before the retrieval-practice phase, in which they are shown another set of slides, with each slide containing the test of a single category exemplar. Overall, three exemplars from four of the eight semantic categories are practiced. All exemplars are tested twice, with an average spacing of 12 trials between the first and the second test. The category label appears centered on the slide with the first two letters of the exemplar printed to the right of it (e.g., FRUIT–Or___). No two category members are tested on adjacent slides. Participants are given 6 s to complete each of the fragments on an empty sheet of paper. After another 2 min of solving arithmetic tasks as a distractor, participants are shown a final set of slides, in which they are tested on all the items that were shown to them during the original study phase. On each slide, participants are provided with the category label plus the unique initial letter of one of the initially learned items (e.g., FRUIT–O___) and are given 6 s to complete the fragment. Within those categories from which items were retrieved in the retrieval-practice phase, the three Rp+ items and the three Rp– items are blocked, and the three Rp– exemplars are recalled prior to the Rp+ exemplars, thus ensuring that output-interference effects at test are avoided (see explanations later in this chapter). Table 24.1 summarizes the procedure. The whole experiment should take approximately 15 min.

TABLE 24.1 Overview of the procedural phases of the classroom demonstration

Study phase	• Participants are presented with the study list, which consists of 6 items from each of 8 semantic categories. • All 48 items of the study list are shown individually together with their category label (e.g., FRUIT–Orange).
Retrieval-practice phase	• Participants are tested twice on half of the exemplars from half of the categories, that is, 3 exemplars from each of 4 of the categories. • For this retrieval-practice task, the category label is presented together with the first two letters of the exemplar, and participants are asked to complete the fragment (e.g., FRUIT–Or___).
Final test phase	• Participants are tested on all 48 items from the original study list. • For this final test, the category label is presented together with the unique first letter of the exemplar and participants are asked to complete the fragment (e.g., FRUIT–O___).

Results

First, you should verify whether participants' repeated retrieval of Rp+ items during the retrieval-practice phase was successful. The percentage of correctly recalled items should be relatively high, typically in the order between 75 and 100 percent. Regarding final-test performance, a paired *t*-test should show that the percentage of correctly recalled Rp+ items is higher than the percentage of correctly recalled control items, which would be line with the results of the Anderson et al. (1994) study and numerous further RIF studies and demonstrate the expected positive effect of retrieval practice on memory of Rp+ items. More importantly, a second paired *t*-test should reveal that the percentage of correctly recalled Rp– items is lower than the percentage of correctly recalled control items, thus reflecting the RIF effect as originally demonstrated by Anderson et al. (1994) in the retrieval-practice task.

Mechanisms of RIF

Several theoretical accounts have been suggested to explain RIF. The two most prominent accounts are the inhibition account (Anderson, 2003) and the blocking account (Raaijmakers & Jakab, 2012). According to the inhibition account, RIF arises because the memory representation of the unpracticed items is impaired during retrieval practice. This account suggests that during retrieval practice of

some of the studied items, related not-to-be-practiced items interfere and compete for conscious recall. To reduce the interference and facilitate selection of the to-be-practiced items, the memory representation of the not-to-be-practiced items becomes suppressed. For instance, when participants are cued with FRUIT–Or___ during the retrieval-practice phase, other fruits, like Banana, may come to mind and compete for conscious recall. To reduce the interference from Banana, the memory representation of the item is suppressed and, as a result, recall of that item in the subsequent final test is impaired.

While the inhibition account thus assumes that RIF arises due to an inhibitory process that acts to overcome interference, the blocking account claims that RIF arises as a direct consequence of interference. According to this account, the associations between the practiced items and their category cues are strengthened during retrieval practice, and such strengthening introduces interference of these items when at test participants are asked to recall the unpracticed items, thus reducing these items' recall relative to the control items. For instance, successful retrieval of Orange in the retrieval-practice phase may lead to a stronger association between Orange and its category label FRUIT. At test, when the participant attempts to retrieve another studied exemplar from the category FRUIT, like Banana, Orange may then interfere and block successful recall of that exemplar, thus creating RIF (for an alternative noninhibitory account of RIF, see Jonker, Seli, & MacLeod, 2013).

In the last two decades, a large number of studies have been conducted designed to contrast the inhibition and blocking accounts of RIF (for a meta-analytic review, see Murayama et al., 2014). In most of this research, some version of the retrieval-practice task was employed. These studies can be divided into studies that (1) manipulated the format of retrieval practice, (2) manipulated how RIF was assessed during the final test phase, (3) and examined individual differences in RIF. The following sections provide a short overview of this research.

Retrieval-practice phase

Retrieval specificity

A central prediction of the inhibition account is that RIF should be retrieval specific, which means that only a practice phase that involves active retrieval of previously studied items should reduce recall of the unpracticed items. Thus, retrieval practice, but not restudy of the practiced items, should induce interference and inhibition of the unpracticed items, and thus impair memory for the unpracticed items. In contrast, the blocking account suggests that the strengthening of the cue-item associations of some of the studied items, be it through retrieval practice or restudy, should always induce blocking during the final memory test and thus impair recall of the unpracticed items.

Early studies that sought to determine whether or not RIF is retrieval specific employed two variants of the retrieval-practice task: restudy and noncompetitive retrieval practice. In both methods, the to-be-practiced "Rp+" items are reexposed

intact with the goal of strengthening the items' associations to their cue without inducing interference and inhibition of the unpracticed items. In the restudy method, some of the originally studied category-item pairs are reexposed (e.g., FRUIT–Orange) and participants are asked to study the word pairs once again. In the noncompetitive retrieval practice method, some of the originally studied items are reexposed and participants are asked to recall the items' category label given the category's word stem as a retrieval cue (e.g., FR___–Orange). The typical finding with these methods is that RIF arises in the (competitive) retrieval practice condition, but no RIF-like impairment is observed following restudy or noncompetitive retrieval practice (e.g., Anderson, Bjork, & Bjork, 2000; Bäuml & Aslan, 2004), which suggests that RIF is retrieval specific.

However, findings from more recent studies seem to indicate that at least some reexposure formats can induce RIF very similar to how retrieval practice does. Raaijmakers and Jakab (2012), for instance, employed a variant of the noncompetitive retrieval-practice method, in which the word stems of the category labels were not presented during reexposure (e.g., ___–Ball) and exemplars with a relatively low frequency within their categories were used. The results showed reliable RIF-like impairment after noncompetitive retrieval practice, indicating that retrieval may not be necessary to induce RIF. These results, however, do not generalize to other testing formats, like item recognition (for more details on RIF in item recognition, see the subsection *Final test phase* later in this chapter). Indeed, when examining retrieval specificity in both recall and item recognition, two recent studies found that in recall, RIF arises after both competitive and noncompetitive retrieval practice, whereas in item recognition, it arises after competitive retrieval practice only (Grundgeiger, 2014; Rupprecht & Bäuml, 2016). These findings support retrieval specificity. In particular, they challenge the blocking account of RIF by demonstrating that, in general, the strengthening of cue-item associations is not sufficient to induce RIF.

Divided attention

In recent years, there have been several attempts to specify the nature of the suppression mechanism which, according to the inhibition account of RIF, underlies RIF. For instance, Anderson (2003) argued that inhibition in memory retrieval is caused by executive-control mechanisms that are not limited to controlling memory. Along these lines, it was suggested that inhibition is a general executive process that also operates to control overt behavior or to ignore irrelevant stimuli. In contrast, other authors proposed that inhibition is mostly an automatic suppression mechanism that acts whenever irrelevant information is coactivated with relevant information, and suppresses the irrelevant information (Conway & Fthenaki, 2003).

Róman, Soriano, Gómez-Ariza, and Bajo (2009) put these two variants of the inhibition account to a test by stressing participants' attentional resources during the retrieval-practice phase with a secondary, concurrent updating task. The idea

was that if the suppression mechanism underlying RIF needs executive control, then overloading attentional resources with a secondary task during retrieval practice should impair the action of the suppression mechanism and thus reduce RIF. On the other hand, if the suppression mechanism was mainly an automatic process, then stressing participants' attention during retrieval practice should not affect RIF. The results showed that, relative to a standard retrieval-practice condition, there was no RIF effect when the secondary task was performed during retrieval practice, but there was reliable facilitation of the Rp+ items. This pattern of results is hard to reconcile with the view that an automatic suppression mechanism underlies RIF. Furthermore, it is inconsistent with the blocking account, because, according to this account, the successful strengthening of Rp+ items in the secondary-task condition should have resulted in impaired memory for the Rp− items. As a whole, the findings suggest that RIF requires attentional control, thus supporting the executive-control version of the inhibition account.

Neural correlates

Several studies have investigated the neural processes underlying RIF by examining neural correlates of the forgetting effect during the retrieval-practice phase. According to the executive-control version of the inhibition account (Anderson, 2003), frontally mediated executive control processes should be recruited to suppress competing items during retrieval practice, which leads to the expectation of enhanced activation in such frontal areas during retrieval practice. No such expectation arises on the basis of the blocking account, which explains RIF through processes that act during the final test phase, and not during the retrieval-practice phase.

Employing electrophysiological measures of brain activity (EEG), Johansson, Aslan, Bäuml, Gäbel, and Mecklinger (2007) compared neural activity in a retrieval-practice condition with neural activity in a restudy baseline condition. Stronger positivity over frontal electrodes was elicited in the retrieval-practice condition relative to the restudy condition, and the stronger positivity was correlated with the amount of RIF arising at the final test. Consistently, a functional magnetic resonance imaging (fMRI) study by Kuhl, Dudukovic, Kahn, and Wagner (2007) showed that the BOLD signal in the prefrontal cortex decreases with increasing retrieval practice trials, indicating a decrease of cognitive control demands with repeated selective retrieval. Activity in ventro-lateral portions of the prefrontal cortex predicted the subsequent forgetting. Most recently, another fMRI study reported even more direct evidence for a critical role of inhibition in RIF, revealing that retrieval practice of some items measurably reactivates competing (Rp−) items and then progressively suppresses those interfering competitors (Wimber, Alink, Charest, Kriegeskorte, & Anderson, 2015). The competitors' cortical traces were even suppressed below the activity observed for the control items, which supports the inhibition account. Overall, the findings from both fMRI and EEG studies suggest that inhibitory processes operate during retrieval practice, inducing RIF at the final memory test.

Final test phase

Cue independence

The inhibition account suggests that RIF should arise regardless of whether, during the final test, participants' memory for a particular item is assessed with the category label that was presented during study, or with a novel test cue that was not present at study or during the retrieval-practice phase. This cue-independence assumption is based on the idea that RIF arises from suppression of the competing memory itself, rather than from strengthening-induced blocking processes. Thus, after study of FRUIT–Orange and FRUIT–Banana, retrieval practice of FRUIT–Orange should weaken the memory representation of Banana, and recall of Banana should be impaired irrespective of which type of cue was provided for recall of the item. In contrast, according to the blocking account, employing novel cues on the final test should generally prevent Rp+ items from interfering with the recall of the Rp− items, and no RIF should arise. Whether or not RIF is observed thus should depend on the nature of the particular retrieval cue.

To test the cue-independence assumption, Anderson and colleagues developed the independent-probe test in which, on the final test, an Rp− item (e.g., Banana) is not tested with its original study cue (FRUIT–B___) but with a novel test cue (YELLOW–B___). The rationale behind this method is that in the presence of the novel retrieval cue, participants may not draw on the original study cue to retrieve the Rp+ item, thus preventing possible blocking effects arising from the Rp+ items. Several studies have indeed found RIF with independent probes, thus supporting the inhibition account (Anderson & Spellman, 1995; see also Murayama et al., 2014). Arguably, however, employing independent probes may not necessarily avoid blocking effects at test, because, when participants receive an independent probe, they may not limit themselves to those probes but may instead covertly recall the original cues that went with the Rp+ items, which may then induce blocking at test. Weller, Anderson, Gómez-Ariza, and Bajo (2013) addressed the issue by asking participants to engage in covert cuing during independent-probe testing. In contrast to the prediction of the blocking account, such instruction decreased RIF, thus supporting cue independence and the inhibition account of RIF.

Item recognition and implicit tests

Further studies supported the proposal that RIF does not reflect blocking by showing that the forgetting effect also arises in further types of memory tests that supposedly do not leave room for blocking, like item recognition. Here the argument is that the presentation of the "old" items from the study phase during item recognition directly cues the episodic representations of those items without activating the shared categories, thus bypassing any form of blocking from the Rp+ items. Indeed, several studies found that memory of Rp− items was still impaired relative to the control items when RIF was assessed through item recognition, a finding regarded as evidence for inhibition (e.g., Hicks & Starns, 2004; Spitzer & Bäuml, 2007).

The argument that in RIF blocking is absent in item recognition has been criticized by some researchers, because, in general, interference effects can be present in item recognition (Raaijmakers & Jakab, 2013). However, there is evidence that not all types of interference effects arise in item recognition and that, for instance, strength-based interference effects, like the list-strength effect – the demonstration that restudy of a subset of studied items can impair memory of the not restudied material – are absent in item recognition (Shiffrin, Ratcliff, & Clark, 1990). Grundgeiger (2014) and Rupprecht and Bäuml (2016) addressed the issue directly by examining the effects of competitive and noncompetitive retrieval practice in recall and item recognition. They found both forms of retrieval practice to induce RIF in recall, but found competitive retrieval practice only to induce RIF in item recognition. Because the effects of noncompetitive retrieval practice have been attributed to blocking (Raaijmakers & Jakab, 2012), these findings support the view that blocking effects may be present in recall but that they are absent in item recognition, thus supporting the proposal of a critical role of inhibition in RIF.

RIF has also been assessed with implicit memory tests. In these types of tests, participants are not directly required to deliberately or consciously recollect the study items and therefore the shared category cue is assumed to be bypassed here. Veling and van Knippenberg (2004), for instance, employed a lexical decision task, in which participants were presented with letter strings and were asked to indicate for each string whether it was a word or a nonword. Critically, some of the letter strings were items from the earlier study phase. The results showed that the word/nonword judgments were impaired (that is, slowed) for Rp− items relative to control items, thus demonstrating RIF. Other implicit tests have also been shown to impair recall of Rp− items, like category generation or category matching, although there is also evidence that not all types of implicit tests may produce a RIF effect (see Murayama et al., 2014).

The findings from cue independence, item recognition, and (some) implicit memory tests point to a critical role of inhibition in RIF. In addition, blocking may also contribute to RIF, at least in tests that leave room for interference effects, like free recall or category-cued recall tests. Further support for this suggestion comes from Murayama et al.'s (2014) recent meta-analysis, which found that the size of the RIF effect is typically larger in free recall and category-cued recall than in independent-probe tests and item recognition, suggesting that in the former type of tests, not only inhibition but also blocking may play some role. Indeed, in both free recall and category-cued recall, there is no control of the items' output order and participants tend to recall the (stronger) practiced items before the (weaker) unpracticed items. This output order bias can induce interference for the unpracticed items and thus lead to an increase in the size of the RIF effect.

Neural correlates

A few studies also searched for neural markers of RIF during the final memory test. On the basis of the inhibition account, one may expect to find neural markers of RIF that reflect the suggested reduced memory-strength signal of the unpracticed

items. The results from two studies point to such neural markers. The fMRI study by Wimber et al. (2008), for instance, found the retrieval of Rp− items to be associated with increased activation in the left anterior region of the ventro-lateral prefrontal cortex. This finding is consistent with the inhibition account, because, in prior neuroscientific work, increased activity in this region was found to reflect the retrieval of weak memory traces. In another study, Spitzer, Hanslmayr, Opitz, Mecklinger, and Bäuml (2009) examined electrophysiological correlates of RIF on recognition memory and found that recognition of Rp− items was associated with reduced power in certain frequency bands. The findings of both studies are in line with the inhibition account, because both effects have previously been suggested to index a reduced memory signal.

Individual differences

On the basis of the view that a frontally mediated executive control process is critically involved in RIF (see above), there is reason to expect that populations that are supposed to suffer from deficient executive control processes show a reduction in the size of the RIF effect. Several studies examined this prediction of the inhibition account, for instance, by examining the relation between RIF and individuals' working-memory capacity and by examining the relation between RIF and participants' age.

Working memory capacity

The efficiency of executive control processes is moderated by individuals' working-memory capacity (WMC), so that individuals with higher WMC are better able to deal with interference and inhibit task-irrelevant information than individuals with lower WMC (Redick, Heitz, & Engle, 2007). Indeed, measures of WMC have been found to predict performance in a number of cognitive tasks supposed to require controlled inhibition, including the Stroop task and the antisaccade task. Because RIF does also seem to depend on executive control processes (Kuhl et al., 2007; Róman et al., 2009), the prediction arises that RIF should be more pronounced in participants with high WMC than in participants with low WMC.

Aslan and Bäuml (2011) addressed the issue. They measured participants' RIF by employing the retrieval-practice task and assessed participants' WMC by means of the operation-span task. The operation-span task is a widely used tool in individual-differences research; it requires participants to simultaneously store and process information and provides reliable and valid measures of individuals' WMC. When relating participants' WMC with participants' retrieval-induced enhancement of the practiced items (recall rate of Rp+ items minus recall rate of C items), these researchers found the enhancement to be largely unaffected by participants' WMC. In contrast, when relating participants' WMC with participants' RIF (recall rate of control items minus recall rate of Rp− items), they found a positive relationship between the two measures, indicating that participants with larger WMC showed

more RIF than participants with lower WMC. This pattern of results is in line with the executive-control-processes view and supports the inhibition account of RIF.

Age

Executive control processes also vary with participants' age, and both young children and older adults often suffer from impaired executive control processes. On the basis of this finding, there is reason to expect that both young children and older adults show less well developed inhibitory function and thus reduced RIF. Aslan and Bäuml (2010, 2012) examined this issue and indeed found RIF to be absent in younger children and older adults, though the absence arose for very young children (i.e., kindergartners) and very old adults (above 75 years) only. Ortega, Gómez-Ariza, Román, and Bajo (2012) replicated the finding that young-old adults (60–75 years) show the RIF effect and additionally showed that, when a secondary task was introduced during retrieval practice, RIF did no longer arise in these adults. This finding suggests that young-old adults may not be "old enough" to show a general reduction in the size of the RIF effect, but that there may be a partial deficit in their inhibitory function that becomes manifest when participants' attentional resources are stressed during retrieval practice.

The two faces of selective memory retrieval

In most studies that examine the RIF effect, the retention interval between study and retrieval practice is very short and lasts a few minutes at best. Because of this, there is typically hardly any change in participants' internal context between study and retrieval practice. On the basis of this, several recent studies examined whether the effects of retrieval practice remain the same when retrieval occurs in a context that differs from the study context, thus making access to the study context during retrieval practice more difficult. Different methods were employed in these studies to impair access to the study context, like the presentation of a forget cue (e.g., Bäuml & Samenieh, 2010, 2012), imagination tasks conducted immediately after study (e.g., Bäuml & Samenieh, 2012; Schlichting, Aslan, Holterman, & Bäuml, 2015), and prolonged retention intervals (e.g., Bäuml & Dobler, 2015; Bäuml & Schlichting, 2014). Also, both the output interference and the retrieval-practice tasks were used to examine the role of study-context access for the effects of selective memory retrieval.

Bäuml and Schlichting (2014), for instance, let participants study a list of unrelated items consisting of predefined target and nontarget items, which was followed by either a short retention interval of 4 min or a prolonged retention interval of 48 hours. While the short retention interval is typically assumed to maintain participants' access to the study context, the prolonged retention interval is supposed to impair access to the study context due to increasing context drift (e.g., Estes, 1955). Subsequently, there was retrieval practice of the predefined nontarget items (retrieval practice) or there was no such retrieval practice (control). Participants

were then asked to recall the predefined target items. When nontarget recall had preceded target recall, the target items served as Rp− items; in the absence of the preceding nontarget recall, the target items served as control items. In the short retention interval condition, the results showed typical RIF, that is, selective retrieval of the nontargets impaired target recall. In contrast, in the prolonged retention interval condition, selective retrieval of the nontargets improved target recall. These results suggest two faces of selective memory retrieval, with a detrimental (RIF) face when the study context is easily accessed during retrieval, and a beneficial face when access is more difficult (see Figure 24.2).

FIGURE 24.2 The two faces of selective memory retrieval. (A) Experimental procedure. Participants study a list of unrelated items consisting of target and nontarget items, and afterwards, participants' mental context is either changed or left unchanged. Subsequently, participants repeatedly retrieve the nontarget items or are engaged in an unrelated distractor task. On the final memory test, participants are asked to recall the target items. (B) Typical pattern of results. When there is no change in context after study, target recall is higher in the absence (C items) than in the presence (Rp− items) of prior nontarget retrieval practice. In contrast, when there is a change in context after study, target recall is lower in the absence (C items) than in the presence (Rp− items) of the retrieval practice.

Bäuml and Samenieh (2012) proposed a two-factor account to explain the two faces of selective memory retrieval. The core assumptions of the account are that (1) in general, selective retrieval triggers two processes, inhibition of competing memories and reactivation of the original study context; and (2) which of the two processes dominates in a certain situation depends on study-context access. When access to the study context is (largely) maintained – like after a short retention interval – interference between items may be high enough to trigger inhibition, whereas not much room is left for context-reactivation processes; as a net result, prior nontarget recall may reduce recall of the target (that is, Rp−) items and RIF may arise. In contrast, when access to the study context is impaired and interference between the single items is low – as may occur after a prolonged retention interval – not much room is left for inhibition and more room is left for context reactivation. Preceding retrieval of the nontargets may reactivate the study context, which may then serve as a retrieval cue for the target (Rp−) items and thus improve target recall.

To date, the two faces of selective retrieval have been shown for both word lists and prose material (Bäuml & Schlichting, 2014) and for both individuals and social groups (Abel & Bäuml, 2015). Also, there is evidence that individual differences do not only modulate the detrimental (RIF) effect but do also modulate the beneficial effect. Schlichting et al. (2015), for instance, examined the relation between individuals' WMC and the beneficial effect of selective retrieval and found a positive relation between the two measures; participants with higher WMC showed a more pronounced effect than participants with lower WMC. Aslan and Bäuml (2014) examined the developmental trajectory of the effect and found seventh graders but not second and fourth graders to show the beneficial effect. Aslan, Schlichting, John, and Bäuml (2015) examined the effect in older age and found the beneficial effect to be present in younger adults but to be absent in older adults. Together with the results from the previous RIF studies, these findings suggest that some groups of people, like young children, older adults, or individuals with low WMC, can not only show impaired retrieval-induced inhibition but can also show impaired retrieval-induced context reactivation. As a result, they show both a reduced detrimental (RIF) effect and a reduced beneficial effect of selective retrieval.

Summary

- Selective retrieval of some memories can induce forgetting of other memories, a phenomenon referred to as retrieval-induced forgetting (RIF).
- RIF is a very robust finding and has been shown for a wide range of study materials and settings.
- The inhibition account of RIF assumes that during selective retrieval of some studied list items the list's not-to-be-retrieved items interfere and are inhibited to reduce the interference.
- The blocking account explains RIF by assuming that the selective retrieval of some items blocks recall of the remaining items at test due to increased interference from the practiced items.

- Findings from numerous RIF studies indicate that inhibitory processes play a critical role for the RIF effect, thus supporting the inhibition account.
- Some other findings suggest that blocking can also contribute to RIF, although only when RIF is assessed via memory tests that leave much room for interference, like free recall or category-cued recall.
- Typically, RIF has been observed when the context during retrieval practice is similar to the context at test; if access to the study context is impaired, however, the RIF effect can be reversed and retrieval practice can improve recall of the other items.

Further reading

For readers looking for a very basic and concise overview about the research on RIF, we recommend Bäuml (2008). Storm et al. (2015) recently provided a very general and more detailed review of RIF, covering the phenomenon as it arises in a variety of situations, such as in eyewitness memory or creative cognition. Anderson (2003) gives a detailed overview of the theoretical properties of the inhibition account of RIF, and also addresses critical factors that can modulate or mask inhibitory processes. Murayama et al. (2014) offer a very comprehensive meta-analysis of RIF, which does also allow evaluation of the inhibition and blocking accounts of RIF.

References

Abel, M., & Bäuml, K.-H. T. (2015). Selective memory retrieval in social groups: When silence is golden and when it is not. *Cognition, 140,* 40–48.

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49,* 415–445.

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review, 7,* 522–530.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1063–1087.

Anderson, M. C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review, 102,* 68–100.

Aslan, A., & Bäuml, K.-H. T. (2010). Retrieval-induced forgetting in young children. *Psychonomic Bulletin & Review, 17,* 704–709.

Aslan, A., & Bäuml, K.-H. T. (2011). Individual differences in working memory capacity predict retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 264–269.

Aslan, A., & Bäuml, K.-H. T. (2012). Retrieval-induced forgetting in old and very old age. *Psychology and Aging, 27,* 1027–1032.

Aslan, A., & Bäuml, K.-H. T. (2014). Later maturation of the beneficial than the detrimental effect of selective memory retrieval. *Psychological Science, 25,* 1025–1030.

Aslan, A., Schlichting, A., John, T., & Bäuml, K.-H. T. (2015). The two faces of selective memory retrieval: Earlier decline of the beneficial than the detrimental effect with older age. *Psychology and Aging*, 30, 824–834.

Bäuml, K.-H. T. (2008). Inhibitory processes. In H. L. Roediger, III (Ed.), *Cognitive psychology of memory* (pp. 195–220). Oxford: Elsevier.

Bäuml, K.-H., & Aslan, A. (2004). Part-list cuing as instructed retrieval inhibition. *Memory & Cognition, 32,* 610–617.

Bäuml, K.-H. T., & Dobler, I. M. (2015). The two faces of selective memory retrieval: Recall specificity of the detrimental but not the beneficial effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 246–253.

Bäuml, K.-H. T., & Samenieh, A. (2010). The two faces of memory retrieval. *Psychological Science, 21*, 793–795.

Bäuml, K.-H. T., & Samenieh, A. (2012). Selective memory retrieval can impair and improve retrieval of other memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 488–494.

Bäuml, K.-H. T., & Schlichting, A. (2014). Memory retrieval as a self-propagating process. *Cognition, 132,* 16–21.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82,* 407–428.

Conway, M. A., & Fthenaki, A. (2003). Disruption of inhibitory control of memory following lesions to the frontal and temporal lobes. *Cortex, 39,* 667–686.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62*, 145–154.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191.

Geiselman, R. E., Fisher, R. P., MacKinnon, D. P., & Holland, H. L. (1985). Eyewitness memory enhancement in the police interview: Cognitive retrieval mnemonics versus hypnosis. *Journal of Applied Psychology, 70,* 401–412.

Grundgeiger, T. (2014). Noncompetitive retrieval practice causes retrieval-induced forgetting in cued recall but not in recognition. *Memory & Cognition, 42*, 400–408.

Hicks, J. L., & Starns, J. J. (2004). Retrieval-induced forgetting occurs in tests of item recognition. *Psychonomic Bulletin & Review, 11,* 125–130.

Hirst, W., & Echterhoff, G. (2012). Remembering in conversations: The social sharing and reshaping of memories. *Psychology, 63,* 55–79.

Johansson, M., Aslan, A., Bäuml, K.-H., Gäbel, A., & Mecklinger, A. (2007). When remembering causes forgetting: Electrophysiological correlates of retrieval-induced forgetting. *Cerebral Cortex, 17,* 1335–1341.

Jonker, T. R., Seli, P., & MacLeod, C. M. (2013). Putting retrieval-induced forgetting in context: An inhibition-free, context-based account. *Psychological Review, 120,* 852–872.

Kliegl, O., & Bäuml, K.-H. T. (2016). Retrieval practice can insulate items against intralist interference: Evidence from the list-length effect, output interference, and retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 202–214.

Kuhl, B. A., Dudukovic, N. M., Kahn, I., & Wagner, A. D. (2007). Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nature Neuroscience, 10,* 908–914.

Macrae, C. N., & MacLeod, M. D. (1999). On recollections lost: When practice makes imperfect. *Journal of Personality and Social Psychology, 77,* 463–473.

Murayama, K., Miyatsu, T., Buchli, D., & Storm, B. C. (2014). Forgetting as a consequence of retrieval: A meta-analytic review of retrieval-induced forgetting. *Psychological Bulletin, 140,* 1383–1409.

Ortega, A., Gómez-Ariza, C. J., Román, P., & Bajo, M. T. (2012). Memory inhibition, aging, and the executive deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 178–86.

Raaijmakers, J. G. W., & Jakab, E. (2012). Retrieval-induced forgetting without competition: Testing the retrieval specificity assumption of the inhibition theory. *Memory & Cognition, 40,* 19–27.

Raaijmakers, J. G. W., & Jakab, E. (2013). Rethinking inhibition theory: On the problematic status of the inhibition theory for forgetting. *Journal of Memory and Language, 68*, 98–122.

Redick, T. S., Heitz, R. P., & Engle, R. W. (2007). Working memory capacity and inhibition. In D. S. Gorfein & C. M. MacLeod (Eds.), *Inhibition in cognition* (pp. 125–142). Washington, DC: American Psychological Association.

Roediger, H. L. (1974). Inhibiting effects of recall. *Memory & Cognition, 2,* 261–269.

Román, P., Soriano, M. F., Gómez-Ariza, C. J., & Bajo, M. T. (2009). Retrieval-induced forgetting and executive control. *Psychological Science, 20,* 1053–1058.

Rupprecht, J., & Bäuml, K.-H. T. (2016). Retrieval-induced forgetting in item recognition: Retrieval specificity revisited. *Journal of Memory and Language, 86*, 97–118.

Schlichting, A., Aslan, A., Holterman, C., & Bäuml, K.-H. T. (2015). Working memory capacity predicts the beneficial effect of selective memory retrieval. *Memory, 23*, 786–794.

Shaw, J. S., Bjork, R. A., & Handal, A. (1995). Retrieval-induced forgetting in an eyewitness-memory paradigm. *Psychonomic Bulletin & Review, 2,* 249–253.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 179–195.

Spitzer, B., & Bäuml, K.-H. (2007). Retrieval-induced forgetting in item recognition: Evidence for a reduction in general memory strength. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 863–875.

Spitzer, B., Hanslmayr, S., Opitz, B., Mecklinger, A., & Bäuml, K.-H. (2009). Oscillatory correlates of retrieval-induced forgetting in recognition memory. *Journal of Cognitive Neuroscience, 21,* 976–990.

Storm, B. C., Angello, G., Buchli, D. R., Koppel, R. H., Little, J. L., & Nestojko, J. F. (2015). A review of retrieval-induced forgetting in the contexts of learning, eye-witness memory, social cognition, autobiographical memory, and creative cognition. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 62; pp. 141–194). San Diego, CA: Academic Press & Elsevier Inc.

Veling, H., & van Knippenberg, A. (2004). Remembering can cause inhibition: Retrieval-induced inhibition as cue independent process. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 315–318.

Weller, P., Anderson, M. C., Gómez-Ariza, C. J., & Bajo, M. T. (2013). On the status of cue-independence as a criterion for memory inhibition: Evidence against the covert blocking hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1232–1245.

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience, 18,* 582–589.

Wimber, M., Bäuml, K.-H., Bergström, Z., Markopoulos, G., Heinze, H. J., & Richardson-Klavehn, A. (2008). Neural markers of inhibition in human memory retrieval. *The Journal of Neuroscience, 28,* 13419–13427.

APPENDIX

Categories and exemplars for the classroom demonstration (material adapted from Kliegl & Bäuml, 2016).

Hobby	*Fruit*	*Animal*	*Musical instrument*
Painting	Plum	Elephant	Flute
Dancing	Cherry	Lion	Trumpet
Running	Orange	Scorpion	Organ
Sailing	Lemon	Tiger	Bass
Basektball	Guava	Roach	Cello
Tennis	Banana	Mouse	Drums
Body part	*Profession*	*Flower*	*Kitchen device*
Finger	Architect	Aster	Knife
Nose	Policeman	Violet	Skillet
Hair	Teacher	Narcissus	Whisk
Elbow	Carpenter	Orchid	Fork
Knee	Gardener	Marguerite	Teapot
Belly	Soldier	Sunflower	Plate

This page intentionally left blank

PART IV

Extensions

This page intentionally left blank

25

SUGGESTIONS AND COGNITIVE ILLUSIONS

Günter Molz and Rüdiger F. Pohl

illudere (Latin) = to play one's game, to pass, to ridicule, to misguide

Research on cognitive illusions has developed a tradition of its own. This book documents the state of the art for an impressive list of cognitive illusions, thus supplying profound insight into the relevant factors causing or modifying these illusions. Most of these illusions have been studied in complete isolation without any reference to the other ones, and moreover without any reference to the large body of research on suggestions. As a matter of fact, the editor of this book noted that the first versions of the chapters' manuscripts contained almost no explicit links to any of the other chapters. Experimental methods, as well as theoretical explanations of cognitive illusions, thus, resembled an archipelago spread out in a vast ocean without any sailor having visited more than one (or at the most two) of these small islands yet. In trying to further explore this largely unknown territory, we believe that (a) cognitive illusions are more connected than previously acknowledged, that (b) they show parallels to suggestions so that (c) both research domains should benefit from each other methodologically as well as theoretically, and that (d) both should be placed within the broader context of general cognitive processes. This approach owes much to the work of Vladimir Gheorghiu (see Text box 25.1).

TEXT BOX 25.1 AN APPRECIATION OF VLADIMIR A. GHEORGHIU (1926–2010)

Our colleague and friend Vladimir initiated this chapter in the first edition of this book (Gheorghiu, Molz, & Pohl, 2004). His main scientific pursuit was to explore suggestion and suggestibility. He discussed and initiated

respective research projects for more than 40 years. Vladimir was deeply convinced that this domain is relevant for a vast variety of topics in basic as well as in applied psychology (Gheorghiu & Kruse, 1991). He was especially interested in the relation between cognitive illusions and sensory suggestibility (Gheorghiu, Koch, Fialkovski, Pieper, & Molz, 2001) and motor relaxation (Gheorghiu & Molz, 2008). His supporting and engaging approach was highly valued by colleagues and students, whom he involved on an equal footing in his research activities. He organized four times the International Symposium on Suggestion and Suggestibility, namely, 1987 in Gießen (Gheorghiu, Netter, Eysenck, & Rosenthal, 1989), 1994 in Rome (De Pascalis, Gheorghiu, Sheehan, & Kirsch, 2000), 2000 in Munich, and 2003 in Krakow. Everyone appreciated also Vladimir's optimism, which he never lost despite all the ups and downs during his conflicted and burdened life (cf. Peter & Netter, 2011).

We are very grateful for having had the pleasure to work with Vladimir and to be inspired by his sophisticated ideas.

Günter Molz and Rüdiger F. Pohl

We start this chapter by summarizing common issues of research linking suggestions and illusions. Next, we discuss a conceptual framework for these two domains. This framework has heuristic value for understanding the functions of cognitive illusions, which we describe in the third and last part of this chapter.

Common issues of research on cognitive illusions and on suggestions

Unfortunately, research on cognitive illusions and on suggestions has proceeded mainly in parallel without too many references to each other. However, there are a number of common issues and relations, which we will point out in this section. For example, suggestion research has incorporated cognitive-illusion methods to induce and measure suggestions, while on the other hand, some cognitive-illusion phenomena were explicitly explained by assuming suggestive processes. This section highlights these points of contact in order to unveil the common mechanisms underlying both suggestion and illusion. Knowing these basic processes might then lead to a more general theory encompassing both cognitive illusions and suggestions.

A brief outline of the history of the research on suggestions

In its beginnings, suggestion research addressed primarily hypnotic contexts in which a certain behavior was forced to occur by means of *direct* and sometimes even authoritative methods (Bernheim, 1884; Breuer & Freud, 1893). This research was

based on direct modes of communication. Persons or patients were told what to do. Their behavior was a result of *immediate reaction patterns*. In order to be able to work with more representative and more ecologically valid suggestions in non-hypnotic contexts, *indirect* methods were developed that activated *mediate reaction patterns*. These were based upon illusions: Stimulus presentations were simulated (Seashore, 1895) or memories were biased by suggestive interrogations (Binet, 1900). The common feature was to hide the researcher's aim to measure the individual propensity to be influenced. For example, Binet (1900) had developed the indirect suggestibility test of progressive lines (or weights). He demonstrated that suggestible persons who had to judge the lengths of successively presented lines (which originally became progressively longer) continued to "see" longer lines even if their length remained constant.

In the following period of suggestion research, the early techniques were further developed and applied in research on suggestibility and persuasibility (e.g., Eysenck & Furneaux, 1945) and in classical experiments in social psychology (e.g., Asch, 1956). Most of these, as well as the contemporary illusion techniques, addressed the manipulation of perception and sensory judgments.

Points of contact

Initiated by the works of Tversky and Kahneman (1974) on heuristics in judgment and decision making, research on cognitive illusions was and is based on several points of contact. Until today, however, the connections between suggestion research and work on cognitive illusions remained rather vague. But this integration is desirable because suggestion research will surely benefit from cognitive psychology's results concerning, for example, knowledge representation and inferential processes. Apart from that, suggestion research may also be helpful for understanding cognitive illusions as is demonstrated by the following examples, though not all of the respective authors used explicitly the term *suggestion*. The first two examples, interrogations and misinformation, probably show the largest overlap between suggestions and illusions, but the remaining five ones also share several aspects. All have been mainly studied in the laboratory, but most have also been observed (or are easily imagined) in everyday situations.

Examples

(1) Among several others, Gudjonsson investigated the psychology of interrogations and confessions. This domain possesses probably the most points of contact with the psychology of suggestion. He demonstrated the decisive role of suggestions in this context – for example, that a false belief could even lead to a false confession for murder (Gudjonsson, 2003). In a more everyday context, adults were asked about their memory of a childhood event, for example, when they were riding with their father in a hot-air balloon (which in reality never happened). In addition, the adults were shown a fabricated photo of that event. A substantial number of them came

to belief that they had actually experienced the event and reported corresponding details (cf. Chapter 22 on the implanting of false memories).

(2) Studies by Loftus and others on the misinformation effect have convincingly shown that suggestions can distort memory, here in the case of eyewitness testimony (see Chapters 20 and 22). In this research paradigm, the misleading information suggests to adapt one's memory of previously received information accordingly. This in turn could lead to having a poorer memory for the original information (compared to a control group), either by having less access to the original information or by readily accepting misinformation as original. Recently, both issues of misinformation and suggestibility have been addressed outside the context of eyewitness research (e.g., Fazio, Dolan, & Marsh, 2015; Gordon, Thomas, & Bulevich, 2015). Fazio, Barber, Rajaram, Ornstein, and Marsh (2013) were even able to distort their participants' long-term semantic memory. Previous knowledge (e.g., "Edison invented the light bulb.") was distorted by introducing misinformation ("Franklin invented the light bulb."). This occurred even for knowledge facts that were previously answered correctly with high confidence.

(3) Research on the hindsight bias (see Chapter 23) focused on the recall of earlier judgments, which were evaluated to be wrong and dysfunctional because of their non-congruency with later provided information. Recall of earlier judgments is typically biased towards the new information. Some researchers interpreted this bias to be adaptive and to support ecologically valid behavior (e.g., Pohl, Bender, & Lachmann, 2002). It allows to better cope with undesirable situations. In order to prevent getting compromised another time (which may have happened for the first time while producing the earlier erroneous judgment), the new information may act as a directive and suggestive cue bringing memory closer in line with current knowledge, which then leads to hindsight bias. Similarly, Conway and Ross (1984) claimed that biasing one's memories may lead to being more satisfied in the actual situation (see Wilson & Ross, 2004).

(4) In a similar paradigm, namely anchoring (see Chapter 12), participants tend to base their numerical judgments on a previously presented "anchor" (a numerical value) that could even be a random number, that is, completely meaningless (Pohl, 2000). But still it exerts its influence on judgment, just as would be expected from any personal or environmental suggestion.

(5) In research on labelling effects (see Chapter 20), a word that was used to label an ambiguous figure was found to bias later reconstruction of that figure in the direction of the given label. Lupyan (2008) found that labeling of typical instances (non-ambiguous figures) of a category (e.g., "chairs") led to poorer recognition rates, when the original figure was compared to a highly similar one. In contrast, labelling of atypical instances (ambiguous figures) did not impair recognition. Lupyan explained his findings with a representational shift account, according to which typical instances of a category activate more of the prototypical instance of that category compared to the activation following atypical instances. As a consequence, recollection is more impaired for typical than for atypical instances. This explanation could also be applied to the well-known "verbal overshadowing" effect

(Schooler & Engstler-Schooler, 1990), which describes the finding that persons' recognition performance, for example, for previously seen faces, can be distorted if these persons had produced a verbal description of the faces after seeing them. In this case, the verbal "label" that overshadowed visual memory consisted of self-generated descriptions.

(6) Another example is given by the associative memory illusion (see Chapter 21). A list of words is presented that contains words that are all highly associated to a target word, which is not on the list. Later, people reliably tend to recall the target word as having been on the list. The converging associations in this case provided a suggestive direction in which to integrate the given word list.

(7) As a final example in our attempt to support the close links between suggestions and cognitive illusions, it may be helpful to look at illusory correlations (see Chapter 6). The tendency to attribute a connection between two simultaneously, but independently occurring events or to ascribe causality to two contiguous, but again independent events (post hoc, ergo propter hoc), appears to be a general human characteristic (Tobacyk, 1991). These illusory correlations and causalities could lead to rigid or even superstitious behavior (e.g., meaningless rituals). In these instances persons tend to confirm their subjective beliefs through auto-suggestive techniques, thereby increasing the subjective sense of stability in a changing environment.

A conceptual framework linking suggestions and illusions

To further substantiate our view of a close link between suggestions and illusions, beyond the few examples given above, we describe in this section a number of reaction patterns and strategies that originate in the above mentioned research on suggestions in the late 19th and early 20th century. We refer to the two profoundly different modes of communicating suggestions as *directly* (Bernheim, 1884; Breuer & Freud, 1893) and *indirectly* (Binet, 1900; Seashore, 1895). These two different modes could well be observed and indeed have been used as explanations also in the domain of cognitive illusions. Then we integrate these strategies and other aspects of suggestions into a larger framework, labelled "suggestionality". Finally, we turn to an alternative view on suggestions and illusions that comes from outside psychology, namely, linguistic pragmatics (Maillat & Oswald, 2009).

General reaction patterns and strategies

In order to fulfill demands of everyday life, people have general reaction patterns that could be categorized into three groups, namely *immediate*, *mediate*, and a combination of both.

(1) *Immediate* patterns consist of unconditioned as well as conditioned reflexes and reactions, which are evoked automatically. Behavior based on these reactions is deterministic, because there are no other options for reacting. These reflexes and reflex-like reactions therefore appear as exclusive.

(2) Contrarily, *mediate* reaction patterns result from intuitive guesses or judgments. There are always at least two options available, one of which must be selected. In order to decide which option to use thereby rejecting the other ones, persons may use one or more of several mechanisms of disambiguation, namely, rational, suggestive, intuitive, or aleatoric strategies (cf. Gheorghiu et al., 2004):

a In *rational* strategies, the underlying decisional process is based primarily upon a reflected comparative assessment of all available options. The outcome thus gains the character of a consciously induced solution. Nevertheless, if the evaluation of all options is too complex, disambiguation mechanisms might be combined with immediate patterns (see below).

b In *suggestive* strategies, one-sided explicit or implicit cues lead to a differential evaluation of the available options. Individuals who follow this strategy behave as if the implemented option were the only one without any alternatives.

c *Intuitive* strategies are similar to suggestive ones. However, evaluation and decision here involve emotional aspects. Bowers (1984) characterized "intuition" as "sensitivity and responsiveness to information that is not consciously represented, but which nevertheless guides inquiry towards productive and sometimes profound insights" (p. 256). These aspects govern the selection of the preferred option, accompanied by certitude to have decided correctly (which is not always true as we all know).

d And finally, selecting an option by chance is an *aleatoric* strategy. These random decisions can easily be overlooked in everyday life. But there are certainly many situations in which none of the other strategies could be applied because the information appears too fragmentary for a rational decision and it does not show any obvious (suggestive or intuitive) clues for which alternative to prefer.

In everyday life, these different mediate reaction patterns are often used in combinations with one another, possibly with a certain individual preference for one or the other of these strategies for judgment, evaluation, and decision.

(3) There are tasks in which a *combination* of both immediate and mediate patterns of reaction are involved and needed. A situation may automatically activate immediate reaction patterns, but will still leave some degrees of freedom for the individual. An example from social psychology is that evoked attitudes, prejudices, or convictions would generally suffice to "choose" an option to act, but we might also feel free to overcome this impulse and to act otherwise. Similarly, developmental findings on imitative behavior provide evidence for these complex reaction patterns: While infants simply cannot avoid imitating, older children and adults mostly have other options available and thus may or may not imitate.

The differentiation between immediate and mediate patterns is to some extent identical to the differentiation between System 1 (unconscious reasoning) and System 2 (conscious reasoning) by Stanovich and West (2000) and between heuristic

TABLE 25.1 Suggested polarities for classifying cognitive functions

Polarity	*Source*
direct – indirect	Gheorghiu (2000)
immediate – mediate	Gheorghiu (2000)
heuristic – analytic	Evans (1984)
System 1 – System 2	Evans (1984)
unconscious – conscious	Stanovich and West (2000)
Type 1 – Type 2	Evans and Stanovich (2013)
nondeliberate – deliberate	Michael et al. (2012)

and analytic processes in the dual-process theory (Evans, 1984, 2008; see also Chapter 8 in this volume). However, the polarities *immediate / heuristic / unconscious* versus *mediate / analytic / conscious* are of different entities. Stanovich and West (2000) referred to systems, whereas the dual-process theory claimed processes. However, both Evans and Stanovich agreed that research has given stronger support for the process-based perspective (Evans & Stanovich, 2013). They suggested to label heuristic processes as *Type 1* and analytic processes as *Type 2*. Another dichotomous denotation was proposed by Michael, Garry, and Kirsch (2012) who differentiated between *deliberate* and *nondeliberate* suggestions. Telling highly suggestible persons to see color though there is only gray is an example for a deliberate suggestion; whereas the well-known Hawthorne effect is a nondeliberate suggestion causing not intended behavior which is due to the awareness of being part of a scientific study (Michael et al., 2012). Regarding this division, *nondeliberate* corresponds to *immediate* and *deliberate* to *mediate*. Table 25.1 lists the terms for the polarities described in this paragraph.

Cognitive illusions could be seen as mixtures of immediate and mediate reaction patterns. They are partly reflex-like in the sense of affective reactions, but also partly deliberate in the sense of a rational calculus, which might be affected by suggestive influences. For example, Bayesian reasoning might be facilitated by the suggestion to use frequencies instead of probabilities (see Chapter 2). In general, however, reflex-like mechanisms are too rigid to dissolve complex ambiguity. The challenge to overcome these narrow patterns also results from the necessity to gain space for interpersonal influence. To maintain subtle interactions between the individual and his or her social environment, fixed and routinely performed actions will certainly not suffice. Rational strategies surely play an important role to dissolve social and other types of ambiguity, but may be too complicated and too time-consuming in a rapidly changing environment. This idea is not new at all and was introduced by Simon (1955). Later he coined this concept as *satisficing* (March & Simon, 1958). In addition, rational strategies may also block themselves by running into a deadlock if no decision can be reached.

As a solution to this state of affairs, suggestive and intuitive strategies may show an efficient way out. On one hand, they allow for more flexibility than immediate reactions (thus overcoming too rigid behavior) and, on the other hand, they are also fast and smooth (thus avoiding strategies which are too complicated and time-consuming). This might explain why suggestive strategies play such an important role in our lives (Gheorghiu & Kruse, 1991).

The domain of suggestionality

Gheorghiu (2000; see also Gheorghiu et al., 2004) proposed a "domain of suggestionality" which was thought of as being analogous to the terms "rationality" or "emotionality" referring to the whole areas of rational or emotional processes, respectively. Likewise, suggestionality was supposed to represent the general frame for classifying and understanding specific suggestional elements. These elements were ordered under four headings, each one representing a different approach to the domain of suggestionality: (a) suggestive situations and phenomena, (b) suggestibility (encompassing general as well as individual characteristics), (c) techniques of suggestions, and (d) theory of suggestionality.

Suggestive situations

Text box 25.2 and Figure 25.1 provide a short characterization of a suggestive situation. In such situations, Person A (or an environmental agent, e.g., a price tag; Molz & Gielnik, 2006) exerts a suggestive technique on Person B, the suggestee. These applications imply a previous state of uncertainty in Person B. In regard to

TEXT BOX 25.2 DESCRIPTION OF A SUGGESTIVE SITUATION

1 In a given context, a person (the suggestee) is guided to react in a certain way by simple or complex, explicit or implicit, coarse or subtle cues.
2 The suggestee may choose among two or more options how to react. In contrast to reflexes or deliberate mechanisms entailing predictable actions, the suggestee follows the given cues or reacts in another way.
3 Readiness for reacting in the suggested way can be impaired or supported by many biological, psychological, and sociological factors and may depend on the individual. Taking these variables into account allows improving the prediction whether the suggestion will succeed or not.
4 The effects of suggestions result in *initiation* or *blocking*, *accentuation* or *diminution*, *conservation* or *transformation* of thoughts, judgments, memories, attitudes, sensations, feelings, and behavior.

FIGURE 25.1 Visualization of a suggestive situation and the involved determinants

single cognitive illusions, various sources affecting the subjective reliability of information can be identified:

- inadequate representations (e.g., framing effect in Chapter 4; illusory correlation in Chapter 6; belief bias in Chapter 9)
- violations of general rules (e.g., conjunction fallacy in Chapter 1; base-rate neglect in Chapter 2; Wason Selection Task in Chapter 8)
- confusions (e.g., Moses illusion in Park & Reder, 2004; halo effects in Chapter 15; associative memory illusion in Chapter 21; misinformation effect in Chapter 22)
- repetitions (e.g., validity effect in Chapter 13; mere exposure effect in Chapter 14)

Note that the suggestive situation is embedded in a functional context that represents the involved goals and potentially adaptive values (see further below).

Suggestibility

Individuals differ in their degree of showing cognitive illusions. Several authors examined the relationships between individual differences and cognitive illusions (e.g., Musch & Wagner, 2007; Pohl, 2000; Read & Winograd, 1998; Stanovich, 1999; Stanovich, Sa, & West, 2004; Stanovich & West, 2000; Stanovich, West, & Toplak, 2013). However, the problem of how to measure the "illusionability" of an individual person was not yet addressed. Therefore, this section is devoted to corresponding proposals from suggestibility research that might be used as starting points for developing similar instruments to assess illusionability.

As mentioned above, individual differences (in addition to other variables) determine the probability with which an individual will follow a suggested direction. Gudjonsson and others (e.g., Gudjonsson, Sigurdsson, Brynjolfsdottir, & Hreinsdottir, 2002) investigated and identified several relevant factors like acquiescence, intelligence, and memory capacity. Gudjonsson (1987) also published the probably most well-known suggestibility scale (Gudjonsson Suggestibility Scales, GSS).

Gheorghiu and co-authors introduced several suggestibility scales for sensory judgments and motor reactions. In traditional research, suggestibility is assessed by pretending that test persons are presented with real stimuli, but in fact these are faked. This implies the possibility that results are strongly influenced by participants' acquiescence. Controlling for acquiescence was the main motive for constructing the Sensory Suggestibility Test (SST; Gheorghiu, Netter, & Tichi, 1989). This was achieved by warning the participants: They were told that stimuli might or might not occur and that they should only react to actually presented stimuli. In fact, a real stimulus was never presented in experimental trials (only in control trials). Given these objectives, a scale of sensory suggestibility was developed for the tactile, auditive, and visual senses.

Gheorghiu and Hübner (1992) addressed an additional problem, namely, whether persons can be sensitized not to be wooed by cognitive illusions. Their Sensory Suggestibility Scale (SSK) can be applied in two variants differing only in their instruction: One instruction pretends that the purpose of the test is to assess one's sensory thresholds (like in the SST), whereas according to the other instruction the purpose of the test is to measure the individual deception. The idea of this latter test instruction is to avoid that persons try to behave according to the assumed test's intention. In fact, the second instruction suggests to fake-bad by giving biased responses. Accordingly, test scores were lower in the second condition ("deception") than in the first one ("threshold"). This can be taken as evidence for the impact of cognitive factors while being tested for sensory suggestibility. Apart from that, both versions have satisfactory psychometric qualities, and variants of them have been successfully applied for assessing the placebo effect (De Pascalis, Chiaradia, & Carotenuto, 2002).

Suggestive techniques

Cialdini (2000) suggested a taxonomy of six different classical techniques of suggestive influences. The first one is the principle of *reciprocation* which implies that whenever we receive we should give in turn. The second principle is based on the concepts of *commitment and consistency*: Once a stand or a choice is taken and is repeatedly confirmed, it is very likely that our or others' future behaviour will be in line with this prior commitment. Cialdini's third and fourth principles are especially of interest in the context of social influence. These techniques are based on *social proof* (do what many others already have done) and *liking* (do what attractive persons / persons you like have already done). The fifth principle has been the subject of psychological research for decades: *authority*. Obedience to authority occurs

where an individual reacts upon suggestions from another individual who is considered to be an authoritative figure (Milgram, 1963). Conformity with authority results from social norms or is enhanced by major groups or groups who venture their opinions in a highly consistent manner (Moscovici, 1976). *Scarcity* is Cialdini's sixth principle, which affects the supply demand ratio. If supply is perceived to be limited, persons are suggested to pursue their needs for demand.

Additionally, fostering of several situational circumstances like overexerted demand characteristics and perceived need for self-serving motivation or coping can also be regarded to be suggestional techniques. Imposing time limitations or masking relevant information is likely to minimize the impact of rational strategies and to maximize the impact of suggestive strategies searching for a satisfying option. Also changing low-stake into high-stake situations (which might overlap with the application of Cialdini's principles of liking or scarcity) increments the possibility that suggestional strategies are applied. Other possible sources for successful suggestions can be seen in self-service motivation and coping. We address these issues further below on functions of cognitive illusions.

Theory of suggestionality

Gheorghiu (2000) demanded that a theory of suggestionality should reflect the importance and meaning of suggestional phenomena. Both importance and meaning become evident if we understand the fact that suggestional phenomena do not only influence behavior in relation to others but also an individual's behavior, thinking, judgment, and memory in isolation. Therefore, in classical psychological research, suggestions (together with imitation and sympathy) can be seen as one of the few major general tendencies in psychology (McDougall, 1918; Taylor, 1954; Young, 1927). Thus, in contrast to many other well-known and often researched subjects in psychological science, suggestional phenomena are ubiquitous, but unfortunately an overarching theory is still missing.

Alternative view: Linguistic pragmatics

Taking a perspective from linguistic pragmatics, Maillat and Oswald (2009, 2013; Maillat, 2013) published a stimulating analysis of cognitive illusions. Their contributions focus on the pragmatics of cognitive illusions within manipulative discourse. Similar to our conception, their claim is that there is still a need for an analytic model to describe and explain manipulative discourse. They view cognitive illusions as well suited to provide a framework for understanding manipulative discourses. Previous accounts from that perspective have focussed on features of the message or of the speaker (like the truth or felicity conditions of the message, the speaker's goals, the covertness of the manipulation, and the social inequality between speaker and addressee), but all these aspects, according to Maillat and Oswald, do not suffice to capture the whole domain of manipulative discourse. Of course, the authors readily acknowledged that all these aspects can be influential, but that for an integrative

theory, one needs to shift perspective from the message and the speaker to the addressee, and especially the addressee's information-processing strategies.

Maillat and Oswald (2009) proposed that cognitive illusions can be explained by relevance theory (Sperber & Wilson, 1995) which provides a framework for understanding the pragmatics of cognitive illusions. According to this approach, the addressee searches for meaning within the manipulative communication (Maillat & Oswald, 2009, p. 361): "Manipulation can be best defined in terms of the constraints it imposes on mental processing." The addressee stops searching after finding meaning which meets his or her expectations of relevance. Thus, manipulative discourse relies on the same processes as any discourse, but especially exploits context-selection procedures (like, e.g., shallow processing, accessibility, or effort minimization), with the effect that the addressee stops searching as soon as a specific but "biased" interpretation context has been selected, thereby missing other potential (and possibly less biased) contexts. In other words (Maillat, 2013, p. 193), "manipulation could be analysed as an attempt to control the hearer's context selection process." As a consequence, the addressee may fall prey to a cognitive illusion. This explanation had already been proposed previously in an interdisciplinary paper by Sperber, Cara, and Girotto (1995), focusing on the Wason Selection Task (see Chapter 8 in this volume).

Maillait and Sperber's approach also bears a striking resemblance to the domain of suggestionality outlined above. There as well, a strong emphasis is placed on the suggestee and his or her processing strategies, so that the two approaches have more in common than one might have expected. The pragmatic approach could even go further than the suggestionality approach, in that it also encompasses situations without speaker or a message, like memory-retrieval driven illusions. In that case, only the information-processing context (or strategy) counts.

Functions of cognitive illusions

Due to the frequent occurrence of illusions, there is no need to answer the question *whether* we are illusioned. More important is to find the reasons *why* such illusions occur, or as Alan Baddeley (debating the role of autobiographical memory) once exclaimed: "But what the hell is it for?" (Baddeley, 1988).

In early research on cognitive illusions, any biases in thinking, judgment, or memory were judged to be malfunctions of the cognitive system (e.g., Tversky & Kahneman, 1974). Later, the focus shifted towards a viewpoint incorporating the possibly adaptive advantages of the mechanisms underlying these biases, so that cognitive illusions were seen as faulty, situation-specific by-products of otherwise highly functional and adaptive processes (e.g., Hoffrage, Hertwig, & Gigerenzer, 2000), or as Jungermann (1983), inspired by the work of Elster (1979), put it: "People are global maximizers with local inconsistencies" (pp. 72–73). But one may even go one step further and claim that the "illusions" themselves are adaptive, for example, to allow influencing others, to gain knowledge in an uncertain domain, or to allow for coping in stressful situations. Hence, the individual

tendency to follow cognitive illusions can be regarded to be to some extent an *ability*, not just a *propensity*. For example, Polczyk, Frey, and Szpitalak (2013) coined the term *influenceability*. Ability implies that following illusion is a skill to adaptively carry out activities and perform smoothly as may be necessary.

Communicative and learning functions

In communication, we may mislead others through manipulative discourse (see above), which thus allows us to exert social influence and to persuade or convince others. This, in turn, could render a relevant group more coherent and efficient, which is also advantageous for the suggestee who has been misled.

In an uncertain world, following suggested information has the benefit of being able to update one's world knowledge, even at the price of being misguided and therefore committing a biased judgment. It might simply feel better to have some knowledge (even if not entirely true) than no knowledge at all. For example, hindsight bias helps in avoiding compromising situations of having poor knowledge (see Chapter 23).

Perceptual functions

Since perceptual errors and biases affect cognitions, visual illusions also belong to the category of cognitive illusions (Gregory, 1997; Molz, 2004). Therefore, we will discuss this class of illusions briefly. Research on optical illusions has a tradition of more than one hundred years. The Müller-Lyer, Poggendorf, or Ponzo-illusions are some prominent examples. Prima facie one might be inclined to evaluate these illusions to be malfunctions of human perception. But a closer analysis shows that they help us to interpret optical stimuli of our three-dimensional environment correctly (e.g., Brentano, 1892).

Coping functions

Another area in which such illusions could possibly be helpful are coping mechanisms. The crucial functions of adaptive coping mechanisms are *orientation*, *support*, and *protection* (Ai et al., 2013; Gheorghiu et al., 2004). These mechanisms are related to suggestive and illusive processes and, typically, they do not work independently but complementary to each other.

Orientation

Orientation is in many situations the prerequisite for fast and efficient acting. But it can be severely hampered by ambiguity or uncertainty, which is all too common. These shortcomings can in principle be reduced or even resolved by means of suggestive directions for one's behavior. Orientation in this sense can be achieved by pseudo-rules, normative rules, rituals, conventions, leading ideas, or self-fulfilling

prophecies. The underlying suggestions are supported by pseudo-rationalizations and pseudo-attributions, or by elaboration of subjective criteria for the truth of personal constructs (Gheorghiu & Kruse, 1991).

Support

Coping functions leading to support ensure that biological, psychological, and social necessities are met. Different illusive techniques may serve the satisfaction of needs concerning food, sexuality, and social relations. This can be achieved in different ways, for example, by encouragement to carry on (by suggesting advantages or stimulating curiosity), by activation of potential capabilities (by influence of feedback or evaluative judgments), or by reduction or elimination of negative implications (by reframing the problem or by increased positivity). These mechanisms are also used in order to reach personal advantages and social power (Cialdini, 2000).

Protection

Conservation and defense of the self is enhanced by protective coping functions (e.g., Greve, Rothermund, & Wentura, 2005). To this end, complex self-illusion techniques may be employed. *Defensive* strategies, for example, protect or enhance self-esteem and reduce dissonance. Several cognitive illusions belong to this kind of coping function. Conway and Ross (1984), for example, discussed the protective function in relation to hindsight bias (see Chapter 23). *Compensational* strategies offer substitute "solutions" within an illusionary environment including religion, esoteric, and magic. In scientific contexts, these coping functions have been discussed in relation to the placebo effect (e.g., Kirsch, 1999). *Ludic* or *playful* strategies are also important and common. Both words, "ludic" and "illusion," derive from the same Latin word "ludere," which means "to play." These strategies are suitable for protection because they may help to create a kind of protective context. They also can give support with new ideas in order to increase problem-solving competence (e.g., McInnes, Howard, Miles, & Crowley, 2009). Human readiness to playfully share and take part in different roles may ease identification, regression, and dissociative procedures. These are relevant in applied contexts like psychotherapy, but also in everyday settings, for example, to reduce boredom.

The adaptive functions of suggestions and illusions in this context can also be seen in combination with other cognitive strategies. The individual not only welcomes and accepts the possibility of an illusion, but also sometimes even reacts frustrated (or disappointed) if the illusion is involuntarily uncovered, forcing him or her to face the naked truth. To prevent this, people often seek refuge in the belief that they have a personal guardian angel and that some dangers (like getting killed in a car accident) are in fact much smaller for themselves than for others (see Chapter 7 on illusions of control). In this way, people may build their own private and secure world, thus underlining the importance of illusions in coping with stressful situations.

Concluding remarks

One important parallel between research on suggestional processes and research on cognitive illusions (cf. Molz, 2004) is that systematic deviations from normatively correct behavior were in both domains typically judged to be malfunctions of the human cognitive system. Jungermann (1983) characterized this point of view as *pessimistic*, which has been the predominant perspective until the 1980s. Tversky's and Kahneman's (1974) work on heuristics and biases can be interpreted as such a perspective searching for systematic errors in human judgments. Beach, Christensen-Szalanski, and Barnes (1987) promoted an orientation towards an epistemic instead of an aleatory (calculated) logic, which we have already mentioned above: Logics of random variables and axioms of probability are not necessarily the appropriate yardsticks for evaluating human performance (see, e.g., Gigerenzer, 1991). This – in Jungermann's terms, *optimistic* – perspective was shared by Gheorghiu (see Gheorghiu & Kruse, 1991; Gheorghiu et al., 2004) not only for cognitive illusions but also for suggestional processes. According to that argumentation, suggestions very often have a positive effect on everyday life and thus should be considered to be basically adaptive. Therefore, one might imagine to coin the term *suggestionability* in the context of situations in which following suggestions is not harmful. In future research, a double-tracked approach seems wise, that is, to further investigate the decisive conditions leading to single cognitive illusions, but at the same time to pursue the goal of developing a more integrative account uncovering the regular set of basic cognitive processes that are responsible for such illusions.

References

Ai, A. L., Lemieux, C., Richardson, R., Tice, N. T., Plummer, C., Huang, B., & Ellison, C. (2013). Character strengths and deep connections following hurricanes Katrina and Rita: Spiritual and secular pathways to resilience among volunteers. *Journal for Scientific Study of Religion*, *52*, 537–536.

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs, 70,* 17–190.

Baddeley, A. (1988). But what the hell is it for? In M. M. Gruneberg & R. N. Sykes (Eds.), *Practical aspects of memory* (Vol. 1, pp. 3–18). New York: Wiley.

Beach, L. R., Christensen-Szalanski, J. J. J., & Barnes, V. E. (1987). Assessing human judgment: Has it been done, can it be done, should it be done? In G. Wright & P. Ayton (Eds.), *Judgmental forecasting* (pp. 49–62). Chichester: Wiley.

Bernheim, H. (1884). *De la suggestion dans l'état hypnotique et dans l'état de veille* [On suggestion in hypnotic state and in sleeping state]. Paris: Doin.

Binet, A. (1900). *La suggestibilité* [Suggestibility]. Paris: Schleicher Frères.

Bowers, K. S. (1984). On being unconsciously influenced and informed. In K. S. Bowers & D. Meichenbaum (Eds.), *The unconscious reconsidered* (pp. 227–272). New York: Wiley.

Brentano, F. (1892). Über ein optisches Paradoxon [On an optical paradox]. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane, 3,* 349–358.

Breuer, J., & Freud, S. (1893). Über den psychischen Mechanismus hysterischer Phänomene [On the psychic mechanism of hysteric phenomena]. *Neurologisches Centralblatt, 1*, 4–10.

Cialdini, R. B. (2000). *Influence: Science and practice*. Boston: Allyn and Bacon.

Conway, M., & Ross, M. (1984). Getting what you want by revising what you had. *Journal of Personality and Social Psychology, 47*, 738–748.

De Pascalis, V., Chiaradia, C., & Carotenuto, E. (2002). The contribution and expectation to placebo analgesia phenomenon in an experimental setting. *Pain, 96*, 393–402.

De Pascalis, V., Gheorghiu, V. A., Sheehan, P. W., & Kirsch, I. (Eds.). (2000). *Suggestion and suggestibility: Theory and research. Hypnosis International Monographs* (Vol. 4). München: M.E.G.-Stiftung.

Elster, J. (1979). *Ulysses and the sirens: Studies on rationality and irrationality.* Cambridge: Cambridge University Press.

Evans, J. St. B. T. (1984). Heuristic and analytic processing in reasoning. *British Journal of Psychology, 75*, 451–468.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology, 59*, 255–278.

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*, 223–241.

Eysenck, H. J., & Furneaux, W. D. (1945). Primary and secondary suggestibility: An experimental and statistical study. *Journal of Experimental Psychology*, *35*, 485–503.

Fazio, L. K., Barber, S. J., Rajaram, S., Ornstein, P. A., & Marsh, E. J. (2013). Creating illusions of knowledge: Learning errors that contradict prior knowledge. *Journal of Experimental Psychology: General, 142*(1), 1–5.

Fazio, L. K., Dolan, P. O., & Marsh, E. J. (2015). Learning misinformation from fictional sources: Understanding the contributions of transportation and item-specific processing. *Memory, 23*, 167–177.

Gheorghiu, V. A. (2000). The domain of suggestionality: Attempt to conceptualize suggestional phenomena 1. Particularities of suggestion. In V. De Pascalis, V. A. Gheorghiu, P. W. Sheehan, & I. Kirsch (Eds.), *Suggestion and suggestibility: Theory and research. Hypnosis International Monographs* (Vol. 4, pp. 1–28). München: M.E.G.-Stiftung.

Gheorghiu, V. A., & Hübner, M. (1992). Die Sensorische Suggestibilitätsskala (SSK) als Erhebungsverfahren für Täuschbarkeit [The Scale of Sensory Suggestibility as assessment of deception]. *Experimentelle und Klinische Hypnose, 8*, 117–129.

Gheorghiu, V. A., Koch, E., Fialkovski, H., Pieper, W., & Molz, G. (2001). Factors influencing the illusion of warmth. *Contemporary Hypnosis*, *18,* 2–31.

Gheorghiu, V. A., & Kruse, P. (1991). The psychology of suggestion: An integrative perspective. In J. Schumaker (Ed.), *Human suggestibility: Advances in theory, research, and application* (pp. 59–75). New York: Routledge.

Gheorghiu, V. A., & Molz, G. (2008). Indirect motor suggestion and relaxation: Development of a scale measuring the ability to relax. *Contemporary Hypnosis, 25*, 78–93.

Gheorghiu, V. A., Molz, G., & Pohl, R. F. (2004). Suggestions and illusions. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement, and memory* (pp. 399–421). Hove, UK: Psychology Press.

Gheorghiu, V. A., Netter, P., Eysenck, H. J., & Rosenthal, R. (Eds.). (1989). *Suggestion and suggestibility: Theory and research.* Berlin: Springer.

Gheorghiu, V. A., Netter, P., & Tichi, H. J. (1989). A test of sensory suggestibility. In A. F. Bennet & K. M. McConkey (Eds.), *Cognition in individual and social contexts* (pp. 507–517). Amsterdam: Elsevier.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology, 2*, 8–115.

Gordon, L. T., Thomas, A. K., & Bulevich, J. B. (2015). Looking for answers in all the wrong places: How testing facilitates learning of misinformation. *Journal of Memory and Language, 83*, 140–151.

Gregory, R. L. (1997). Visual illusions classified. *Trends in Cognitive Science, 1*, 190–194.

Greve, W., Rothermund, K., & Wentura, D. (Eds.). (2005). *The adaptive self: Personal continuity and intentional self-development*. Göttingen: Hogrefe.

Gudjonsson, G. H. (1987). A parallel form of the Gudjonsson suggestibility scale. *British Journal of Clinical Psychology, 26*, 215–221.

Gudjonsson, G. H. (2003). *The psychology of interrogations and confessions: A handbook*. New York: Wiley.

Gudjonsson, G. H., Sigurdsson, J. F., Brynjolfsdottir, B., & Hreinsdottir, H. (2002). The relationship of compliance with anxiety, self-esteem, paranoid thinking and anger. *Psychology, Crime and Law, 8*, 145–153.

Hoffrage, U., Hertwig, R., & Gigerenzer, G. (2000). Hindsight bias: A by-product of knowledge-updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 566–581.

Jungermann, H. (1983). The two camps on rationality. In R. W. Scholz (Ed.), *Decision making under uncertainty* (pp. 67–86). Amsterdam: North-Holland.

Kirsch, I. (Ed.). (1999). *How expectancies shape experience*. Washington: American Psychological Association.

Lupyan, G. (2008). From chair to "chair": A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General, 137*, 348–369.

Maillat, D. (2013). Constraining context selection: On the pragmatic inevitability of manipulation. *Journal of Pragmatics, 59*(Part B), 190–199.

Maillat, D., & Oswald, S. (2009). Defining manipulative discourse: The pragmatics of cognitive illusions. *The International Review of Pragmatics, 1*, 348–370.

Maillat, D., & Oswald, S. (2013). Biases and constraints in communication: Argumentation, persuasion and manipulation [Special issue]. *Journal of Pragmatics, 59*(Part B).

March, J. G., & Simon, H. A. (1958). *Organizations.* New York: Wiley.

McDougall, W. (1918). Some general or non-specific innate tendencies. In W. McDougall (Ed.), *An introduction to social psychology* (pp. 93–124). Boston, MA: Luce.

McInnes, K., Howard, J., Miles, E. G., & Crowley, K. (2009). Behavioural differences exhibited by children when practising a task under formal and playful conditions. *Educational and Child Psychology, 26*, 31–39.

Michael, R. B., Garry, M., & Kirsch, I. (2012). Suggestion, cognition, and behavior. *Current Directions in Psychological Science, 21*, 151–156.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67,* 371–378.

Molz, G. (2004). Behavioural decision making and suggestional processes: Some reflections on common features. *Imagination, Cognition and Personality*, *24*, 139–149.

Molz, G., & Gielnik, M. (2006). Does the introduction of the Euro have an effect on subjective hypotheses about the price-quality-relationship? *Journal of Consumer Behaviour, 5*, 204–210.

Moscovici, S. (1976). *Social influence and social change*. London: Academic Press.

Musch, J., & Wagner, T. (2007). Did everybody know it all along? A review of individual differences in hindsight bias. *Social Cognition, 25*(1), 64–82.

Park, H., & Reder, L. M. (2004). Moses illusion. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (pp. 275–291). Hove, UK: Psychology Press.

Peter, B., & Netter, P. (2011). In memoriam: Vladimir Gheorghiu. *International Journal of Clinical and Experimental Hypnosis, 59*, 143–145.

Pohl, R. F. (2000). Suggestibility and anchoring. In V. De Pascalis, V. A. Gheorghiu, P. W. Sheehan, & I. Kirsch (Eds.), *Suggestion and suggestibility: Theory and research. Hypnosis International Monographs* (Vol. 4, pp. 137–151). München: M.E.G.-Stiftung.

Pohl, R. F., Bender, M., & Lachmann, G. (2002). Hindsight bias around the world. *Experimental Psychology*, *49*, 270–282.

Polczyk, R., Frey, O., & Szpitalak, M. (2013). The relationships between suggestibility, influenceability, and relaxability. *International Journal of Clinical and Experimental Hypnosis, 61*, 319–335.

Read, J. D., & Winograd, E. (1998). Individual differences and memory distortions [Special issue]. *Applied Cognitive Psychology, 12*(7).

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22*, 36–71.

Seashore, C. E. (1895). Measurements of illusions and hallucinations in normal life. *Studies from the Yale Psychophysiolocigal Laboratory*, *2*, 1–67.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99–119.

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition, 52*, 3–39.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. Oxford: Blackwell.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

Stanovich, K. E., Sa, W. C., & West, R. F. (2004). Individual differences in thinking, reasoning, and decision making. In J. P. Leighton & R. J. Sternberg (Eds.), *The nature of reasoning* (pp. 375–409). New York: Cambridge University Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*, 645–665.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. C*urrent Directions in Psychological Science, 22*(4), 259–264.

Taylor, W. S. (1954). Suggestion. In W. S. Taylor (Ed.), *Abnormal and social psychology* (pp. 464–497). New York: American Book Company.

Tobacyk, J. J. (1991). Superstition and beliefs about the predictability of future events. *Psychological Reports, 68*, 511–512.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Wilson, A. E., & Ross, M. (2004). Illusions of change or stability. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (pp. 379–396). Hove: Psychology Press.

Young, K. (1927). Imitation, suggestion, sympathy, and compensation. In K. Young (Ed.), *Source book for social psychology* (pp. 242–244). New York: Knopf.

AUTHOR INDEX

Aaker, J. L. 93
Abelson, R. P. 139
Abramson, L. Y. 137, 138–139, 140, 141, 145, 147
Ackerman, R. 199
Adams, J. K. 292
Adams, P. A. 292
Alba, J. W. 295, 298, 310
Alexander, D. N. 263
Allan, L. G. 129
Alloy, L. B. 129, 137, 138–139, 140, 141, 143, 145, 147
Alogna, V. K. 377
Alter, A. L. 285
Amir, Y. 258
Anderson, C. 34
Anderson, M. C. 447, 448, 452, 454, 460
Andreou, C. 145
Angle, S. T. 140
Aristotle 167, 205, 300
Arkes, H. R. 246–248, 252, 254
Armor, D. A. 327
Arndt, J. 398
Arrow, K. J. 70
Asch, S. 278, 285
Aslan, A. 453, 456, 459
Aßfalg, A. 351, 353
Åstebro, T. 309, 311
Ayeroff, F. 139
Ayers, M. S. 413, 421

Bacon, F. 99
Bacon, F. T. 246, 254
Baddeley, A. 478
Bajo, M. T. 452, 454, 457
Baker, M. 308, 311
Barber, B. C. 140
Barber, B. M. 308
Barber, S. J. 470
Barbey, A. K. 60
Barlas, S. 307
Barnes, V. E. 481
Barston, J. L. 172
Bar-Tal, D. 264
Bartels, R. D. 83
Bartlett, F. C. 7, 14, 378
Bassok, M. 102
Bauer, R. M. 268
Bäuml, K. H. 453, 455, 456, 457, 459, 460
Bayen, U. J. 431
Bayesianism 14
Baym, C. L. 419
Beach, L. R. 481
Beck, J. R. 31
Begg, I. 246, 248, 254
Bell, R. 363, 365
Belli, R. F. 380
Ben-David, I. 309
Benjamin, A. S. 199
Benoît, J. P. 305
Berlyne, D. E. 269, 270
Bernstein, D. M. 323, 342, 351, 353, 415–416, 437, 440
Berscheid, E. 279
Biais, B. 308
Billings, F. J. 414
Biner, P. M. 140
Binet, A. 469

Birmingham, L. 285
Birnbaum, M. H. 34
Bjork, R. A. 448
Blank, H. 435
Blavatskyy, P. R. 300
Bless, H. 197, 200
Boehm, L. 248
Bolger, F. 311
Boo, H. C. 237
Bornstein, M. H. 379
Bornstein, R. F. 261, 262, 263, 264, 266, 270, 273
Boswell, C. P. 419
Botvinick, M. M. 146
Boucher, J. 315, 330
Bouts, P. 141
Bower, G. H. 191
Bowers, K. S. 472
Braun, K. A. 380, 415
Brewer, N. T. 231
Brigham, J. C. 385, 388
Brody, N. 262
Brown, J. D. 144
Bruce, A. C. 300
Bruce, D. 403
Bruner, J. 402
Brunswik, E. 301
Buchner, A. 363, 365, 427
Budescu, D. V. 303
Bugg, J. M. 360
Burns, D. J. 362–363, 366
Burns, S. A. 363
Busmeyer, J. R. 35
Butler, A. C. 364

Cain, D. M. 299, 303
Camerer, C. 308
Campbell, J. D. 430
Cara, F. 478
Carlson, B. W. 29, 40
Carmichael, L. 375–378, 387
Carroll, M. 385
Carstensen, L. L. 324
Cesarini, D. 300
Chaiken, S. 83
Chan, J. C. K. 412
Chapman, J. P. 115
Chapman, L. J. 115
Chase, V. M. 30–31, 34
Chassin, L. 286
Chater, N. 154, 158
Cheng, P. W. 158–159
Christensen-Szalanski, J.J.J. 429–430, 431, 435, 440, 481
Chromiak, W. 246, 254
Cialdini, R. B. 476–477
Clements, C. M. 143
Clifasefi, S. L. 415–416
Cohen, L. J. 160
Collett, M. E. 353
Combs, B. 190
Conway, M. 470
Coolin, A. 431
Cooper, W. H. 289
Corkill, A. J. 12
Corominas, J. 73
Cosmides, L. 159
Costello, F. 37–38, 40
Cotel, S. C. 398
Cox, J. R. 155
Coyle, J. T. 7
Craik, F. I. M. 364
Craver-Lemley, C. 263
Crump, M. C. 129
Curran, T. 198, 201
Czikszentmihalyi, M. 329

D'Agostino, P. R. 262, 270
Daniel, T. C. 378
Davey, S. J. 7
DeCapito, A. 379
Dechêne, A. 254
Deese, J. 397
Delton, A. W. 361
Demarest, I. H. 378
DeMartino, B. 95
Dempster, F. N. 12
Derbish, M. H. 360
Descartes, R. 242
Dickins, T. E. 361
Dion, K. K. 279, 280, 282, 285, 289
Dodds, P. S. 319
Dodson, C. S. 13
Doherty, M. E. 208
Donovan, S. 33
Donovan, W. L. 144
Dougherty, M. R. P. 304
Druckman, J. N. 85
Dubra, J. 305
Dudukovic, N. M. 453
Dunning, D. 325, 330

Eagly, A. H. 289
Edwards, K. 107
Edwards, W. 6, 63–64
Efran, M. G. 282
Egloff, B. 435
Elqayam, S. 14, 161
Elster, J. 478
Englich, B. 223

Engstler-Schooler, T. Y. 383, 385
Epley, N. 237
Epstein, S. 33
Erdfelder, E. 198, 364, 386, 427
Erev, I. 303, 304
Esfandiari, A. 72
Evans, J. St. B. T. 7, 12, 14, 30, 110, 152–153, 160, 161, 162, 165, 168, 172, 173, 178, 180

Fantino, E. 72
Fast, N. J. 144
Fazio, L. K. 470
Feist, M. I. 379
Fenton-O'Creevy, M. 144
Ferrell, W. R. 302
Fiedler, K. 5, 125, 127, 129
Fischbach, G. D. 7
Fischhoff, B. 190, 300, 436
Forgas, J. P. 277, 289
Franco, R. 35
Frederick, S. 218–219, 220, 231, 237
Frenda, S. J. 421
Freudian theory 150
Freund, T. 264
Frey, O. 479
Friedland, N. 141
Friedmeyer, M. T. 286
Frigo, L. C. 349
Funder, D. C. 7
Furnham, A. 237

Gäbel, A. 453
Gaeth, G. J. 80, 83
Gaissmaier, W. 73
Gallagher, K. M. 87
Gallo, D. A. 395, 397, 398, 403
Gambara, H. 86
Garb, H. N. 213
Garry, M. 414, 415, 473
Gavanski, I. 214
Gentner, D. 379
Getz, S. J. 146
Gheorghiu, V. 467–468, 474, 476, 481
Gifford, R. K. 116–117
Gigerenzer, G. 6, 8–10, 12, 15, 216, 296, 302, 303
Gilovich, T. 6, 7, 8, 220, 237
Girotto, V. 478
Glaser, M. 308
Godwin, P. 311
Goel, A. M. 309
Goldsmith, R. W. 34
Gómez-Ariza, C. J. 452, 454, 457
Gonsalves, B. D. 419
Gonzalez, R. 220
Gonzalez-Vallejo, C. 307
Goodnow, J. J. 72
Graham, J. R. 309
Green, J. D. 321
Greene, R. L. 342, 348–349, 352, 353
Gregory, R. L. 264
Greifeneder, R. 197
Greve, K. W. 268
Griffin, D. 6, 220
Griggs, R. A. 155
Grosjean, S. 113
Groß, J. 431
Grundgeiger, T. 455
Gude, C. 245, 254
Gudjonsson, G. H. 469, 476
Guilbault, R. L. 429, 440

Ha, Y. 113
Haddock, G. 195
Hahn, U. 14
Hall, L. 113
Hamilton, D. L. 116–117
Hanawalt, N. G. 378
Handal, A. 448
Handley, S. J. 58, 165, 168
Hannah, S. D. 129
Hanslmayr, S. 456
Harari, H. 285
Harley, E. M. 437
Harper, C. 168
Harris, A. J. L. 14
Harris, G. 246
Harrison, A. A. 252
Hart, J. 362–363
Hartnett, J. 322
Harvey, C. R. 309
Hasher, L. 246, 254
Hastie, R. 440
Hawkins, S. A. 440
Healy, P. J. 291
Heaps, C. M. 417–418, 420
Hell, W. 5
Hertwig, R. 30–31, 34, 198, 201, 249
Herz, H. 309
Herz, R. S. 382, 387
Herzog, S. M. 198
Hicks, J. L. 350
Higgins, E. T. 83, 92, 186
Hilbert, M. 9
Hilbig, B. E. 198
Hilton, D. 308
Hirshman, E. 398
Hoch, S. J. 437
Hockley, W. E. 349, 350

Hodgson, R. 7
Hoffrage, U. 296
Holyoak, K. J. 158–159
Holzworth, R. J. 208
Howe, M. L. 360, 368
Hübner, M. 476
Huff, M. J. 403
Hume, D. 205
Husband, T. H. 414
Hutchinson, J. W. 295, 298, 310
Hyman, I. E., Jr. 413, 414
Hynan, L. G. 34

Jacoby, L. L. 413
Jakab, E. 452
James, G. 71, 75
James, W. 406
Jamis, M. 412
Jensen, J. D. 87
Johannesson, M. 300
Johansson, M. 453
John, T. 459
Johnson, J. E. V. 300
Johnson, M. K. 403
Johnson-Laird, P. N. 151, 152, 155, 168
Jungermann, H. 478, 481
Juslin, P. 296, 299, 302, 304, 311

Kahn, A. S. 324
Kahn, I. 453
Kahneman, D. 5–6, 8, 10, 15, 26, 28, 29, 32, 35, 41, 46–48, 52, 58, 84, 162, 186–188, 193, 198, 199, 200, 204, 206, 208, 209, 212, 215, 218–219, 220, 227–229, 301, 436, 469, 481
Kang, S. H. 364
Kao, S.-F. 116
Kauff, D. M. 262
Keillor, G. 316
Keinan, G. 141
Kelley, C. M. 413
Kennison, R. E. 353
Keren, G. 12, 95, 219, 310
Ketcham, K. 421
Kirkpatrick, E. A. 391
Kirsch, I. 416, 473
Klauer, K. C. 180
Klayman, J. 113, 307
Klein, S. B. 361, 363
Kleinbölting, H. 296
Koehler, D. J. 71, 75, 220
Kongthong, N. 268, 273
Kool, W. 146
Kort, J. 288
Koutstaal, W. 13, 403
Koval, P. 285
Kramer, M. E. 363
Kroneisen, M. 364–365, 368
Kronlund, A. 342, 353
Krueger, J. I. 7
Kruger, J. 316, 325, 330
Kruglanski, A. W. 12, 264
Krugman, H. E. 250
Kühberger, A. 86, 95
Kuhl, B. A. 453
Kuhn, D. 113
Kühnen, U. 197
Kunda, Z. 113
Kunst-Wilson, W. R. 273

Laham, S. M. 285
Landy, D. 280, 284
Langens, T. A. 140, 143
Langer, E. J. 134, 136–137, 138, 139, 140, 142, 147
Langlois, J. H. 289
LaPaglia, J. A. 412
Laplace, P. S. 205
Larimer, M. E. 145
Latimer, A. 83
Lawrence, M. 311
Layman, E. 214
Layman, M. 190
Leavitt, L. A. 144
LeBoeuf, R. A. 231
Lee, A. Y. 93
Lehman, D. R. 417
Levin, I. P. 80, 83, 86, 95
Lichtenstein, S. 190, 200, 300, 310
Light, L. L. 353
Lilienfeld, S. O. 113
Lindsay, D. S. 415
Lloyd-Jones, T. J. 386
Löckenhoff, C. E. 324
Loftus, E. F. 380, 381, 408, 409–410, 414, 415–416, 417, 418, 421, 470
Louie, K. 95
Louie, T. A. 439
Lovallo, D. 308
Lowenstein, G. F. 437
Luo, C. R. 348
Lupfer, M. B. 214
Lupyan, G. 379, 470
Lynch, J. St. 152–153

Maillat, D. 477–478
Malmendier, U. 309, 311
Mandel, D. R. 91–92
Mandler, G. 269, 270
Manktelow, K. I. 156, 158, 159, 162

Manning, C. G. 414
Mantonakis, A. 415–416
Marsh, E. J. 470
Marsh, R. L. 350
Martin, P. J. 286
Matlin, M. W. 330
Mazurier, K. 308
Mazzoni, G. A. L. 416
McCabe, D. P. 362
McCarthy, J. D. 64
McClelland, A. G. R. 311
McCloskey, M. 411–412
McDavid, J. W. 285
McDermott, K. B. 5, 6, 392, 393, 397, 398, 400, 401, 417
McGahan, J. R. 119
McGoey, P. J. 302
McKenna, F. P. 135–136, 138–139, 143
McKenzie, C. R. M. 93, 128
Mecklinger, A. 453, 456
Meek, S. W. 419
Meissner, C. A. 385, 386, 388
Melcher, J. M. 384
Mellinger, A. E. 140
Mercier, H. 105, 113
Mesulam, M. M. 7
Metcalfe, J. 412
Meyerowitz, B. E. 83
Meyers, J. 250
Michael, R. B. 473
Mill, J. S. 214
Miller, M. B. 395
Minami, T. 268
Mitchell, K. J. 403
Mitterer, J. 246
Mochon, D. 231, 237
Moore, D. A. 291, 299, 303
Moore, J. E. 286
Moreland, R. L. 273
Mortiz, S. 145
Moxey, L. M. 93
Müller-Lyer 479
Murayama, K. 460
Murphy, A. H. 293, 300
Musch, J. 12, 430
Mussweiler, T. 223, 237

Nagae, S. 378–379
Nairne, J. S. 357–364, 366, 368
Nakauchi, S. 268
Nanda, R. 309
Napoleon, T. 286
Nash, M. 417–418, 420
Neill, C. M. 324
Neimark, E. D. 63
Nestler, S. 435
Newell, B. R. 65, 73, 75
Nicholson, N. 144
Nickerson, R. S. 113
Niewiadomski, M. W. 349, 350
Nilsson, H. 34–35, 41
Nisbett, R. E. 201, 283, 289
Norman, K. A. 403

Oaksford, M. 154, 158
O'Connor, M. 310, 311
Odean, T. 308
Oeusoonthornwattana, O. 272
Okado, Y. 418, 419
O'Keefe, D. J. 87
Olsson, H. 304, 311
Önkal, D. 311
Opitz, B. 456
Oppenheimer, D. M. 231
Ornstein, P. A. 470
Ortega, A. 457
Osgood, C. E. 315, 330
Ostrove, N. 283, 285–286
Oswald, M. E. 113
Oswald, S. 477–478
Otgaar, H. 368
Over, D. E. 156, 158, 159, 160, 162

Paivio, A. 374
Pandeirada, J. N. S. 357–364, 366
Park, J. H. 140
Parkes, C. M. 254
Penfield, W. 408
Pennycook, G. 53–54, 58, 59, 60
Pentland, J. 413
Perfect, T. J. 385
Peynircioglu, Z. F. 339, 340
Pezzo, M. V. 438, 439, 440
Pezzo, S. P. 439
Pfeiffer, P. E. 303
Phan, L. B. 327
Phillips, M. C. 419
Piaget, J. 167
Piatelli-Palmarini, M. 7
Pickrell, J. E. 417
Pinon, A. 86
Poggendorf 479
Pohl, R. F. 198, 382, 437
Polczyk, R. 479
Poletiek, F. H. 113
Pollard, P. 172
Porter, E. 315
Porter, S. 417
Pothos, E. M. 35
Pouget, S. 308

Prentice, W.C.H. 378
Price, J. R. 273
Prull, M. W. 353

Raaijmakers, J. G. W. 452
Rajaram, S. 470
Rakow, T. 65
Read, J. D. 415
Reber, R. 195
Reder, L. M. 413, 421
Reed, A. E. 323
Regev, Y. 141
Reimer, T. 198
Renner, C. H. 253
Reyes, R. M. 191
Ritchie, T. 322, 330
Rivkin, I. D. 327
Robertson, T. E. 361
Roediger, H. L. III 5, 7, 8, 360, 364, 392, 393, 397, 400, 417
Röer, J. P. 363, 365
Roese, N. J. 435, 439, 440
Róman, P. 452, 457
Ronson, J. 99
Rose, R. L. 116
Ross, L. 201
Ross, M. 201, 470
Rotello, C. M. 349
Roth, J. 140
Rothman, A. J. 83, 195, 196
Rubin, D. C. 323
Rupprecht, J. 455
Russo, J. E. 297

Salovey, P. 83
Samenieh, A. 459
Sander, E. 105
Sandewall, Ö. 300
Sanford, A. J. 93
Sanna, L. J. 437
Santos-Pinto, L. 294, 306
Savary, J. 231
Schacter, D. L. 7, 13, 401, 403
Schlehofer, M. M. 144
Schlichting, A. 457, 459
Schneider, S. L. 80, 92
Schooler, J. W. 383, 384, 385–386, 387, 388
Schooler, L. J. 73, 198
Schul, Y. 12, 219
Schulze, C. 73, 74
Schwartz, M. 254
Schwarz, N. 193–195, 196, 197, 198, 199, 200, 201, 437
Schwikert, S. R. 198, 201
Scott-Phillips, T. C. 361
Seamon, J. G. 262, 270, 398
Sedikides, C. 321
Seligman, M.E.P. 329
Shackle, G.L.S. 32
Shanks, D. R. 64, 272
Shaw, J. S. 448
Shepperd, J. A. 327, 330
Sher, S. 93
Sherman, S. J. 414
Shoemaker, P.J.H. 297
Shuford, E. H. 63
Sicoly, F. 201
Siegel, S. 129
Siegler, I. C. 323
Sigall, H. 280, 283, 284, 285–286
Simon, H. A. 473
Skowronski, J. J. 322
Sloman, S. A. 60
Slovic, P. 6, 190, 300
Smith, E. E. 107
Smith, H. D. 206–207
Sniezek, S. M. 353
Snyder, M. 101–102
Soane, E. 144
Sobel, J. 294, 306
Soderstrom, N. C. 362
Soll, J. B. 307
Soriano, M. F. 452
Sperber, D. 105, 113, 478
Spinoza, B. 242
Spitzer, B. 456
Sporer, S. L. 386
Stadler, M. A. 392
Stahlberg, D. 437
Stang, D. J. 262, 270, 315, 323, 330
Stanovich, K. E. 12, 104, 161, 178, 472, 473
Stark, C.E.L. 418, 419
Storm, B. C. 460
Strack, F. 223, 237
Sullivan, L. E. 7
Sully, J. 7
Susa, K. J. 386
Svenson, O. 294
Swann, W. B. 101–102
Szpitalak, M. 479

Tabachnik, N. 129
Tate, G. A. 309, 311
Taylor, H. A. 31
Taylor, S. E. 144, 327
Tentori, K. 31, 36–37, 41
Tesser, A. 252, 430
Thakor, A. V. 309
Thapar, A. 353
Thomas, D. R. 379

Thompson, C. P. 321, 322
Thompson, S. C. 137, 140, 142–143, 147
Thompson, S. R. 357–364
Thompson, V. A. 53–54, 58, 178
Thompson, W. C. 191
Thorndike, E. L. 278
Titchener, E. B. 256
Todd, P. M. 10
Toplak, M. E. 12, 104
Topolinski, S. 273
Toth, J. P. 254
Trippas, D. 58, 60
Trope, Y. 102
Trouche, E. 105
Trueblood, J. S. 35
Tulving, E. 364
Tunney, R. J. 64
Tversky, A. 5–6, 8, 10, 15, 26, 28, 29, 32, 35, 41, 46–48, 52, 84, 162, 186–188, 193, 198, 199, 200, 204, 206, 208, 209, 212, 215, 220, 227–229, 301, 436, 469, 481

Underwood, B. J. 391, 397
Unkelbach, C. 252, 254
Unturbe, J. 73
Updegraff, J. A. 87
Upfold, D. 246

Vanags, T. 385
Van Avermaet, E. 141
Vendemia, J.M.C. 419
Verde, M. F. 349
Viscusi, D. 141
Vogl, R. J. 321
Vohs, K. D. 435, 439, 440
von Clef, J. 382, 387
von Collani, G. 435
von Restorff, H. 120
von Winterfeld, D. 6
Vulkan, N. 64, 75

Wade, K. A. 415
Wagner, A. D. 453
Wagner, T. 12, 430
Walker, W. R. 321, 322
Wallsten, T. S. 303
Walsh, R. O. 144
Walster, E. 279
Wänke, M. 200, 201
Wason, P. 100, 150, 155, 162
Wasserman, E. A. 116
Watkins, M. J. 339, 340
Watson, J. M. 397, 401
Watts, P. 37–38
Weber, M. 308
Weber, R. A. 309
Weinstein, N. D. 192
Weinstein, Y. 360
Weller, P. 454
Wells, B. 322
Wells, G. L. 214
West, R. F. 12, 104, 178, 472, 473
West, S. A. 361
Westerman, D. L. 342, 348–349, 352, 353
Wetzel, C. G. 288
Whittlesea, B. W. A. 273
Willham, C. F. 429–430, 431, 435, 440
Williamson, J. D. 119
Willman, P. 144
Wilson, P. R. 284, 289
Wilson, T. D. 288, 289
Wimber, M. 456
Winkielman, P. 197
Winkler, R. L. 300
Winograd, E. 403
Wlaschin, J. 83
Wolford, G. L. 31, 395
Wurgler, J. 308, 311
Wyer, R. S. 34

Xu, G. 248

Yates, J. F. 29, 30, 40, 310
Yates, M. C. 119
Young, D. 286
Yuille, J. C. 417

Zajonc, R. B. 252, 256, 269, 271, 273
Zalmanov, H. 199
Zaragoza, M. S. 411–412
Zechmeister, E. B. 245, 254

This page intentionally left blank

SUBJECT INDEX

ABC Research Group 10
accessibility 186, 478; selective 232–237
acquiescence 476
adaptive intelligence 127–129
adaptive learning 440
advertisement research *see* consumer research
affect heuristic 11
affective primacy hypothesis 271
affirmation theory 348–349
age: and decision making 324; and hindsight bias 431, 440; and the misinformation effect 414; and retrieval-induced forgetting 457, 459; and the revelation effect 353
aggression, overt vs. covert 122–127, 132
aleatoric reaction strategies 472
Alzheimer's patients 263, 268
anchoring effect 223–224; conclusions 236–237; conversational influences 230–231, 236, 237; insufficient adjustment 229–230, 236, 237; numeric priming 231–232, 236, 237; paradigms 227–229; as pervasive and robust 225–226, 232, 237; relevance of 226–227; scale distortion 235–236, 237; selective accessibility 232–235, 236, 237; theoretical accounts 229–236
anchoring heuristic 5, 6, 237; *see also* anchoring effect
anchoring paradigm 470
antisaccade task 456
antisocial personality disorder 213–214
Argument Evaluation Task (AET) 178
argumentative theory of reasoning 111–112
Asian disease problem 80, 82, 84, 85, 87–90
associative false recognition effect 396
associative learning 129, 289
associative memory illusions 15, 390, 471; activation processes 403; classroom demonstration 391–392; conclusion 402; discussion 395–396; monitoring processes 403; sample experiment (DRM paradigm) 393–396; theories and data 396–397; typical results 394–395; *see also* DRM (Deese–Roediger–McDermott) effect
associative network theories 287–288
associative tradition 390–391
attitude polarization effect 234, 247, 252
attractiveness halo 289; *see also* physical attractiveness
attribute framing 80, 81, 83, 94
attribute substitution 71, 218
autobiographical memory 322, 478
availability: vs. accessibility 186; classroom demonstrations 187–190; definition of 186; famous-name experiment 187–188; letter-frequency experiment 188–190; *see also* availability heuristic
availability heuristic 5, 6, 11, 185–186; adopting the perspective of others 192–193; applications of 190–192; biased encoding and retrieval of information 190–191; conclusions 199–200; ease or amount of recall 193–194; and the

effects of processing intensity 196–197; malleability of the judgment 195–196; naïve theories 197–198; recall as diagnostic experience 194–195; relevance of the feeling for the judgment at hand 195; and retrieval fluency 198–199; salience of the feeling 197; summary 193; vividness of information 191; *see also* availability
averaging 29, 33–34
averaging model of conjunctive probability judgment 34
averaging models 34, 39

Backward Associative Strength (BAS) 397
Barron-Welsh Art Scale 266, 268
base-rate effect 299
base-rate fallacy 45
base-rate neglect 44–45, 205, 206, 211, 220; causality in 51; conclusions 59–60; conditions from Pennycook and Thompson 53–54, 58; congruent problems 54–55; and dual process theory 57–59; frequently investigated varieties 45; implicit base-rate problems 45–46; incongruent problems 56–57; lawyer-engineer problem 45–46, 52–53, 58; mammography problem 45, 46–47, 52; results and discussion 47; theoretical accounts 48, 51–53
Bayesian principles 14, 73, 473
Bayes' theorem 5, 31–32, 47, 52, 54, 209, 303, 305, 473
behavioral finance 311
belief bias 165–167; belief by logic interaction 173; classroom demonstration 170–172; conclusions 179; in conditional inference 176–177; in informal reasoning 178; old and new paradigms 175–176; syllogistic 167–170, 172–174; *see also* beliefs
Belief Index 172
beliefs: balancing 288; based on imperfect measurement 305; causal 215; false 415, 469–470; information based on 58; materialist 59; and misbeliefs 15; mistaken 109–111, 211; naïve 197–198; overoptimistic 295, 310; polarization of 295; prior 85, 111, 116, 119, 155, 162, 218; rational 305, 307; revision of 14; statistical 215; subjective 9, 14, 291, 310, 471; supernatural 159, 480; unrealistic 307; *see also* belief bias
better-than-average effect 294–295, 303, 306, 325, 328
betting *see* gambling
bias and biases 5; artificial or inflated 9; cognitive 6–7, 256; debate on 8; ecological validity of 12–13; and evolutionary selection 15; fading affect 321–322; hindsight 15, 226, 249; in human reasoning 159–162; lack of explanatory power 10; matching 152; one-sided view of 8–9; optimistic 325; reduction of 216–217, 288–289; self-serving 306–307; systemic reasoning 25; taxonomy of 7; truth 252–253; and Type 1 and Type 2 processes 155; Wason's view of 150; *see also* belief bias; confirmation (myside) bias; heuristics; heuristics and biases; hindsight bias
bidirectional framing effect 86, 88; *see also* framing effect
blocking account 450–454, 459–460
boredom, proneness to 270, 273
Brier score 292–293
Brunswikian error 304
bubbles, in finance markets 308

calibration 6, 292–293, 296–297
calibration paradigm 295, 305
calibration-in-the-large 293, 310
calibration-in-the-small 293
causality heuristic 11
causality judgments 214–215
ceiling rule 30–31, 34
CEST (cognitive experiential self-theory) 29, 33, 39
CHARM theory 413
Chief Executive Officers (CEOs), overconfidence in 309
Chief Financial Executives (CFOs), overconfidence in 309
childhood sexual abuse, memories of 12
chi-squared test 28
choice blindness 108, 113
chronic pain 439
classroom demonstrations: anchoring effect 224–225; associative memory illusions 391–392; availability experiment (1) 187; availability experiment (2) 188–190; belief bias in syllogistic reasoning 170–172; base-rate neglect 48–50; cognitive illusions 17; confirmation (myside) bias 103–105; conjunction fallacy 26–28; framing 87–90; halo effect 281–282, 284–285; hindsight bias 431–433; illusions of control (computer) 137–138; illusions of control (driving) 135–136, 144; illusions of control

(gambling) 136–137, 142, 145; labelling effect 376–377; misinformation effect 409–411; overconfidence (calibration) 296–297; overconfidence (precision of numerical estimates) 297–298; Pollyanna Principle 316–319; probability matching (1) 66–68; probability matching (2) 68; retrieval-induced forgetting 448–450; revelation effect 344–348; selection task as decision making 157–158; survival processing effect 358–359; validity effect 243–245
clinical depression 328, 439
clinical psychology 13, 213–214, 374
cognition: creative 460; social 288
cognitive biases 12, 256; *see also* biases; heuristics
cognitive experiential self-theory (CEST) 29, 33, 39
cognitive functions, classification of 473
cognitive illusion(s): categories of 5–6; cross-referencing of 16; communicative and learning functions 479–480; defining features 3–5; earlier collections of 6–8; functional views 14–15; new research on 16; normative issues 13–14; status of 13–15; *see also* cognitive illusions debate; cognitive illusions research
cognitive illusions debate: dual-process models 11–12; fast and frugal heuristics 10–11; heuristics and biases 8–9
cognitive illusions research 468–469, 481; and the domain of suggestionality 474–477; examples 469–471; on the function of cognitive illusions 478–479; and the history of research on suggestions 468–469; and linguistic pragmatics 477–478; linking suggestions and illusions 471–474; points of contact 469
cognitive miserliness 7
cognitive processing 217
cognitive psychology 17, 154, 413
Cognitive Reflection Test 104
colour, memory for 379–380
communication, and cognitive illusions 479–480
compatibility 6
compensational strategies 480
competitive framing 85; *see also* framing; framing effects
computers, illusions of control on 137–138
conditional inference 174–175
confidence judgments 6, 310
confidence-frequency effect 299–300
confidence-frequency theory 302
configural weighted averaging (CWA) method 34–35
confirmation (myside) bias 99–100, 107, 113, 152, 276, 301; debiasing or making the best of 112–113; demonstrations of 100–108; explaining 110–112; hypothesis testing 100–103; limits of 108–110; vs. myside bias 107; potential usefulness of 111–112; thought listing 106–108; Wason selection task 103, 105–106, 112; *see also* halo effects
congruity effect 363–364
conjunction fallacy 25–26, 206, 211, 218, 220, 475; applying a different probabilistic rule 31–32; averaging 33–35; classroom demonstration 26–28; cognitive experiential self-theory 33; conclusions 38–41; frequentist interpretations and response format 30–31; inductive confirmation 36–37; linguistic interpretation 29; quantum probability 35–36; random noise or sampling errors 37–38; signed summation 29–30; surprise theory 32–33
conjunction rule 25
consumer research 128, 227, 374
consumption, and overconfidence 306–307
contaminated mindware 7
contamination 6
control, perceived *see* illusions of control
control heuristic 142
coping mechanisms 479
counterfactuals 6, 214
creative cognition 460
criterion-flux hypothesis 350–351
cue independence 455
cue validities 302–303
cumulative prospect theory 85, 91

death-scenario conditions 362–363
debiasing 112, 252
debiasing effect 169
decision-making 6; and age 324; from descriptions 9; dichotomous 260; from experience 9, 74; medical 12; psychology of 227, 236
decrement-to-familiarity hypothesis 349–350
deduction paradigm 175
deductive reasoning: belief bias in 165–167; and conditional inference 174–175; *see also* belief bias

Deese–Roediger–McDermott effect *see* DRM (Deese–Roediger–McDermott) effect
defensive strategies 480
depression 328, 439
depressive realism 147
developmental psychology 271
devil effect 276; *see also* halo effects
die problem 64–70
discrepancy-attribution hypothesis 351–352
distinctiveness account 120
distinctiveness heuristic 401
distortions 5, 128, 226–227; judgment 8–9, 382; memory 7–8, 12–13, 382, 390, 406–407, 414, 416, 419, 420–421, 435; scale 229, 235–237
domain expertise 213–214
dream interpretations 414
drinking age problem 155–156
driving, illusions of control in 135–136, 144
DRM (Deese–Roediger–McDermott) effect: association-based theories 397–398; fluency-based attributions 400; neural mechanisms of 402; processes that reduce 401–402; sample experiment 393–397; similarity-based theories 398–400
dual-coding theory 374
dual-process theory 154–155, 473; and base-rate neglect 57–59
dying-to-remember (DTR) effect 363

ease of recall 193–194, 199
Eastern cultures 328
ecological models 302–303, 310
ecological validity 249–253
economic forecasting 439
economics 80, 128
EEG (electroencephalography) tests 402, 419, 453
effort minimization 478
egocentrism 7, 307
electrodermal responses 259
electroencephalography (EEG) 402, 419, 453
emotional arousal 362
emotion regulation 15
emphasis framing 80, 81, 94, 95
encoding 5, 7, 13, 83, 119, 120–121, 127, 200, 263, 271, 359, 361–363, 378, 385–387, 390, 406, 407–408, 419; biased 190–191, 193; of correct judgment vs. original judgment 430, 435–437; richness of 364–367
engineer-and-lawyer problem 209
engram 408
entrepreneurship 308–309, 311
Equation 176
equivalency framing 80
error independence 301
error models 303–304, 310
errors: of commission 5, 8; fundamental attributive 192; halo 277, 278, 289; of omission 4–5, 7, 8; sampling 12, 29, 37–38
event encoding *see* encoding
event-related potentials (ERPs) 198, 418
evolution, and the survival processing effect 357–360
experiments *see* classroom demonstrations; research
expertise effect 300
exposure effect *see* mere exposure effect
external validity 139, 161
extrasensory perceptions tasks 139
eye-movement tracking 152
eyewitness testimony 12, 295, 374, 418, 460

faces, memory for 360
facial expression 284
fading affect bias 321
fallacies 168–169, 173; base-rate 45; ecological validity of 12; gambler's 207, 211, 218, 220; inverse 209, 222; planning 291, 307, 327, 439; *see also* biases; conjunction fallacies
false beliefs 415, 469–470
falsification principle 5, 100, 107, 111
false memories 8, 13, 391, 398, 402; compared to true 420; false recall 396, 397, 400; neurophysiology of 418–419; *see also* associative memory illusions; false beliefs; false recognition
false recognition 391–393, 395–398, 400; *see also* associative memory illusions; false memories
familiarity 134, 139, 142, 145, 198, 248–254, 256, 259, 269–272, 341–343, 348–350, 353, 398, 416
familiarity attribution 416
familiarity-decrement hypothesis 349–350
familiarity hypothesis 329
famous-name experiment 187–188
feature-positive effect 120
feeling-of-knowing (FOK) 341–342
figure-ground drawing, reversible 264
filial regression 212
finance, overconfidence in 308
fixation 39–41, 155, 431
flashbulb memories 414
fluency 351–352; retrieval 198–199

fluency heuristic 11, 198–199
fluency illusions 7
fMRI (functional magnetic resonance imaging) techniques 268, 402, 418, 419, 420, 453, 456
food preferences 382
forecasting 6, 16; judgmental 311; weather 300
forensic psychology 374
four cards selection task *see* Wason selection tasks
framing 79–80; attribute 80, 81, 83, 94; broader construal of 94–95; competitive 85; emphasis 80, 81, 94, 95; equivalency 80; examples of 81–82; of goals 80, 82, 83–84; issue 80; and the pragmatics of language 93–94; risky-choice 80, 82, 84–86, 92, 93; varieties of 80–85; *see also* framing effect
framing effect 17, 79–80, 95; bidirectional 86, 88; cognitive accounts 91–92, 113; explaining 90–91; size of 85–87; formal characteristics 91; goal 80, 82, 83–84; government expenditures 81; hierarchy of 94; humanitarian 81; lean vs. fat 81–82; moderators and mediators 90; motivational accounts 92–93, 113; unidirectional 85–86, 88; *see also* framing
free recall 359
frequentism 29, 30–31, 217
functional Magnetic Resonance Imaging (fMRI) 268, 402, 418, 419, 420, 453, 456
fundamental attribution error 192
future events: assessment of 146; control of 141; planning for 362; simulation of 15
fuzzy trace theory 91–92, 398, 399, 401

gambler's fallacy 207, 211, 218, 220
gambling 300; illusions of control in 136–137, 142, 145
Gestalt psychology 279, 285, 287, 376
gist representation and traces 14–15, 91, 93, 246, 253, 398–399, 402, 407
global-matching hypothesis 349, 350
goal framing 80, 82, 83–84
government expenditures frame 81
Gudjonsson Suggestibility Scale (GSS) 476
guessing 7, 119, 392, 396
gut feelings 161, 217–218

halo biases *see* halo effects
halo effects 17, 276–277; boundary conditions 285–286; and celebrity opinions 286–287; classic experiments 279–280; consequences and practical importance 286–287; history 278–279; judicial 282–283, 286; limits to 283; moderators and bias reduction 288–289; negative 276; other examples 284–285; physical attractiveness 279–284; theoretical explanations and psychological mechanisms 287–288; typical experiment and class illustration 277–278; in the workplace 286; and writer competence 277–278
halo error 277, 278, 289; *see also* halo effects
happiness: and age 324; and childhood 198; and country of residence 325; and physical attractiveness 279–280; rating 323–325; and wealth 324, 325
hard-easy effect 298–299
HB13 model 427–428, 436
Herring illusion 264
heuristics 5–6; affect 11; anchoring 5, 6, 237; applied perspectives 12–13; artificiality of 9–10; availability 5, 6, 11; causality 11; control 142; criticism of 216; distinctiveness 401; dual-process models 11–12; fast and frugal 10–11, 12, 15; fluency 11, 198–199; general-purpose 11; individual differences 12; lack of explanatory power 10; one-sided view of 8–9; recognition 11, 272; representativeness 5, 6, 38, 57, 186; similarity 11; surprise 11; Take-The-Best 11; unrealistic optimism 192; *see also* availability heuristic; bias and biases; heuristics and biases; representativeness heuristic
heuristics and biases program (Tversky and Kahneman) 301
hindsight 307
hindsight bias 15, 226, 249, 424, 470; applied perspectives 439; classroom demonstration 431–433; cognitive processes 435–436; components view 435; computational models 437; conclusions 439–440; correct judgment (CJ) 425–430, 433–436; definition 428–429; discussion 434–435; empirical evidence 429–430; examples 425; experimental designs 425; extensions 430; individual differences 430–431; measures of 425–427; and meta-cognition 437–438; model-based measures 427–428, 434; motivational processes 438–439; original judgments (OJ) vs. recall of original judgments (ROJ) 425–430, 433–436; recollection-reconstruction theory of 436–437;

research results 433–434; theoretical accounts 435–439
Homer's *Odyssey* 212
horns effect 276; *see also* halo effects
hubris 295; *see also* overconfidence
humanitarian frame 81
hypnosis 7
hypochondriasis 13
hypothesis testing 5, 232; for confirmation (myside) bias 100–103

illusions: of change or stability 17; memory-based 15; of perception 7; positive 15; of reality 7; of source monitoring 7; of transparency 227; of validity 213; visual 479
illusions of control 134, 295; classroom demonstrations 135–138; in financial investment 144–145; implications of 143; and the intrusion of reality 141; methods of researching 134, 138–139; and mood 141; need or desire for the outcome 140–141; negative consequences 144–145; neural research 146–147; in parenting 144; positive consequences of 143–144; and power 141; reconciling positive and negative effects 146; and regulatory focus 142; skill-related factors 139; success or failure emphasis 139–140; theories of 142–143
illusory correlation 471; conclusions 127–128; definitions 117–118; discussion 127; expectancy-based 116, 117, 119–120, 124–125; experiment by Hamilton and Gifford 117; experiment by Kao and Wasserman 116; experimental results 125–127; experimental task and dependent measures 118–119; phenomenon of 115–116; sample experiment 122–125; similarity-based 120, 128; theoretical accounts of 119–122; unequal opportunity to learn 117, 120; unequal weighting of information 116, 117, 120
imagination inflation 15, 414
implicit associative response (IAR) 397
impression formation 277, 278, 287, 374
inductive confirmation 29, 36–37
inference(s): conditional 174–175; conversational 230–231, 236, 237; informal 174; pragmatic 106; probabilistic 227; schema-based 5; transitive 374; unwarranted 288
influenceability 479
information: belief-based 58; misleading 7; tagging of 242
information-leakage account 93
information processing: social 287; variables 288
inhibition account 450–457, 459–460
insufficient adjustment 229–230, 236, 237
Intensity Principle 320–321
Interaction Index 172
interference 7–8, 385–386, 390, 451–452, 455–457, 459–460
interference theory 406
intergroup behaviors 258, 271
interquartile index 294
interrogations 113, 469
intuition 155; about random sequences 206–208; about sample sizes 208–209, 211
intuitive physics 6
intuitive reaction strategies 472
inverse fallacy 209, 222
irrationality: in human beings 128, 161; and overconfidence 301; Wason's view of 150
issue frames 80; *see also* emphasis framing

Japanese culture 328
judgment(s): absolute 224, 228, 231, 234–235; comparative 224, 233–234, 236; confidence 6, 310; and the conjunction rule 25–26; distortion in 8–9, 382; everyday 6; expert 6; forced-choice preference 259; heuristic 219; illusions in 5–7; influences on 224; of learning 359; psychology of 227; remember 400; by representativeness 204–205 (*see also* representativeness heuristic); and the revelation effect 339–342, 352–353; sequential 228, 236; social 227, 277; of taste and odour 381–383; of vulnerability 196–197
Judgment and Decision Making (journal) 11
judgmental anchoring 227, 237

knew-it-along effect *see* hindsight bias
knowledge: judgment-relevant 234; missing 7; previous 470; prior 119, 128, 439; schematic 408; stereotypical 116; updating 15
Korsakoff's syndrome 263, 268

labelling effect 373–374, 387; applied domains where it might occur 374; classroom demonstration 376–377; examples 373–374; judgments of taste

and odour 381–383; Law of Prägnanz 376; memory for colour 379–380; memory for speed 381; memory for visual objects 375–376, 378–379; theoretical accounts 375–376; *see also* verbal overshadowing
Lake Wobegon effect 316–318, 320, 325–328, 330
language pragmatics 93–94, 477–478
law of large numbers 213
Law of Prägnanz 376
law of small numbers 213, 220
lean frame 81
learning: associative 129, 289; incidental 295
legal testimony *see* eyewitness testimony
letter-frequency experiment 188–189
likely and unlikely (LU) events 37
Likert ratings 243, 259–260
linguistic interpretation 29
linguistic pragmatics 93–94, 477–478
local mental models 302
logic 5–6, 93–94, 100, 104, 110, 327, 481
Logic Index 172
ludic strategies 480

magnetoencephalography 145
management, overconfidence in 309, 311
market entry 308–309, 311
matching bias 152
maximax strategy 92
medical diagnosis 439
memory and memories: age-related positivity effect 322–323; autobiographical 322, 478; of childhood sexual abuse 12; for colour 379–380; distinguishing true from false 416–418; distortions in 7–8, 12–13, 382, 390, 406–407, 414, 416, 419, 420–421, 435; effects of verbalization on 383–386; episodic 361, 366–367; of faces 360; flashbulb 414; impairment of 412, 440; implicit 263, 271, 455; increasingly positive 321–322; indirect 360; long-term 323; neural mechanisms of 402; normative issues 14; positivity effect 322–323; recognition 359; and retrieval expectations 401; schematic 407; source 360; source of activation (SAC) 413; for speed 381; and survival 357; traumatic 418; for visual objects 375–376, 378–379; *see also* associative memory illusions; encoding; false memories; free recall; working-memory capacity (WMC)
Memory Assessment Procedure (MAP) 417
Memory Characteristics Questionnaire (MCQ) 400
memory illusions 5–6, 7–8; susceptibility to 8, 12, 412; *see also* associative memory illusions
memory traces 398
mental arithmetic 6
mental disorders 13
mental models 165
mental models theory 173
mere exposure effect (MEE) 256, 315–316; affective primacy model 271; arousal model 269; behavioral measures 259–260; classroom demonstration 264–266; conclusions 271–272; examples 256–257; exposure variables 261–262; imagery effects 263; individual differences 263–264; in intergroup behavior 258; measures of 259–260; moderating variables 260–261; neurological correlates 268–269; nonspecific activation model 269; overexposure effect 268; participant variables 262–263; perceptual fluency/attributional model (PF/A) 270, 272; in political elections 257–258, 259; in preference for novel types of art 257; real-world implications 272; relevance 257–258; in repetition and liking for music 257; research methods 258–259; results of study 267–268; reverse 271; in sales and marketing research 257; stimuli 266–267; stimulus complexity 261; stimulus variables 260; theoretical accounts 269–271, 273; two-factor model 270; and unfamiliar people 257
message-framing effects 85, 86–87, 93; *see also* framing; framing effects
meta-cognitions 437–438, 440
meta-memory 295, 341
mindware gaps (missing knowledge) 7
MINERVA–DM model 304
misattributions 13, 437
misbeliefs 15; *see also* beliefs
miscalibration 310
miserly processing account 59
misinformation effect 15, 406–409, 418, 469–470; behavioural evidence 416–418; conclusion 419–420; consequences 415–416; manipulations of 413–415; neuroscientific evidence 418–419; possible mechanisms 416; theoretical accounts 411–413
misinformation study 409–411
Monte Carlo simulations 128

Moses illusion 17
motivation, and overconfidence 307
multiTree software 428, 433
myside bias 107, 113; *see also* confirmation (myside) bias

naïve theories 197–198
negative affective states 13
negative thoughts, intrusive 13
negotiations 226–227, 259
Nekker illusion 264
neuroimaging techniques 268, 402, 418, 419, 420, 453, 456
new paradigm 14, 175, 180
non-regressive predictions 211–212, 220
normative system problems 160
norms 6
numeric priming 231–232, 236, 237
numerical estimates, precision of 293–294, 297–298

obsessive-compulsive disorder (OCD) 13
Odyssey (Homer) 212
optimism 6, 295; *see also* overconfidence
optimistic bias 325
orientation 479–480
orientation illusion 17
overconfidence 291, 439; and the base-rate effect 299; and calibration 292–293, 310; classroom demonstrations 295–298; combined error models 304; and the confidence-frequency effect 299–300; consumption value of 306–307; ecological models 302–303, 310; in entrepreneurship and market entry 308–309; and error independence 301; error models 303, 310; and the expertise effect 300; in finance and trading 308; and format dependence 300–301; functions of 306–308; and the hard-easy effect 298–299; heuristics and biases 301; and the importance of sampling 299; in management and CEOs 309; motivation value of 307; and the overconfidence effect 298, 310; placements in rank orderings 294–295, 310; and the precision of numerical estimates 293–294, 310; rational beliefs based on imperfect measurement 305; related methods and measures 295; relevance of 308–309; sensory sampling model 304–305; signal-detection approach 302; signaling value of 307–308; and subjective multidimensional assessment 305–306; tasks and measures 291–292; theoretical accounts 301–306; and underconfidence 297, 301, 303, 305
overconfidence effect 298, 310
over-discounting hypothesis 352
overestimation 147, 186, 291–292; *see also* overconfidence
overexposure effect 268
overoptimistic beliefs 295; *see also* overconfidence
overplacement 292; *see also* overconfidence
overprecision 292; *see also* overconfidence
override failure 7

paradigms, old and new 175–176; *see also* new paradigm
pathological health anxiety 13
perceptual fluency model (mere exposure effect) 270
permanence theory 408
personality, implicit theories of 279, 287
personality disorders 213–214
personality profiles 414
pessimism, retroactive 438–439
physical attractiveness 279–284, 286
physics, intuitive 6
placebo effect 308, 476, 480
planning fallacy 291, 307, 327; *see also* overconfidence
playful strategies 480
Poggendorf illusion 264
point of subjective equality (PSE) 379–380
Pollyanna Hypothesis 315
Pollyanna Principle 315–316; accuracy of recall for daily events 320–321; accuracy of recall for personal information 321–322; age-related positivity in memory 322–323; examples in language 319–320; examples in perception 319; exceptions for 328; explanations for 328; judgment and age 324, 329; judgment and country of residence 325, 329; judgment and wealth 324, 329; memories for events 321–322; overview 319; and psychology's negative emphasis 328–329; rating of personal abilities and characteristics 325–326; research study 316–319; unrealistic optimism effect 327–328
The Pollyanna Principle: Selectivity in Language, Memory, and Thought (Matlin and Stang) 315, 323, 325, 330
positive illusions 15
positive psychology 329, 330
positive testing 121
positivity effect 322–323

pragmatic reasoning schemas 158, 162
preciseness paradox 214
preference-reversal effects 227
prejudice 374, 472; *see also* stereotyping
probabilistic mental models (PMMs) 302, 304
probabilistic rules 29, 31–32
probability and probabilities: assessment of 5–6; and base-rate neglect 44–45; vs. frequencies 217; frequentistic 217; judgments of 6; maximizing 66
probability matching 17; as an adaptive response 73–74; conclusions 74–75; definition 62–63; and the die problem 64–70; as an error 70–73; when probabilities are unknown 63–64; when probabilities are known 64–66
probability theory 5, 25
probability weighting 85
problem solving 6, 115, 177, 180, 383, 480
processes: activation vs. monitoring 403; automatic vs. controlled 386; intuitive 60; reflective 60
processing intensity 196–197
prospect theory 84–85, 91
protection 144, 479, 480
prototype effects 276; *see also* halo effects
pseudocontingencies 129
psychography 7
psychology and psychologists 80, 95, 155, 161, 162, 258, 272; clinical 13, 213–214, 374; cognitive 17, 154, 413; developmental 271; emphasis on the negative 328–329; forensic 374; Gestalt 279, 285, 287, 376; of judgment 227; mathematical 213; positive 329, 330; of reasoning 150, 155, 174–175; social 128, 129, 277, 374, 472
Psychology Information Test (PIT) 250–251
psychopathology 13

qualitative likelihood index (QLI) 29–30, 38–39
quantum probability 29, 35–36

RAFT (Reconstruction After Feedback with Take-the-best) 437
randomness: diagnosing 210–211; intuitions about 206–207
random noise 29, 37–38
ranking paradigm 303
rationalisation 155
rationality, in human reasoning 159–162
rational reaction strategies 472
reaction patterns, immediate vs. mediate 471–472
reasoning 6, 108–109; argumentative theory of 111–113; indirect 166; judgments 340–342, 344; pragmatic 158, 162; probabilistic 25–26; *see also* deductive reasoning; syllogistic reasoning
recognition 250, 262, 340, 359, 367–370, 383–386, 394–396, 471; false 391–393, 395–398, 400; item 452, 454–456; latency 198; memory 246, 253; test of 6, 339, 341, 345, 347–348, 400–402, 410–411, 418, 419
recognition heuristic 11, 272
recognition memory 359
recollection bias 440
recollection-reconstruction theory of hindsight bias 436–437
Reconstruction After Feedback with Take-the-best (RAFT) 437
regression effects 4, 303, 438–439
regulatory-focus theory 93
reiteration effect 249, 296
relatedness effects 7
relevance theory 478
representativeness, defined 215
representativeness heuristic 5, 6, 38, 57, 186, 204–205, 218; alternative explanations 218; biases are not errors 217; biases can be reduced 216–217; and causality judgments 214–215; conceptual vagueness 216; criticisms of 216–218; demonstration 1 (intuition about random sequences) 206–208; demonstration 2 (intuition about sample sizes) 208–209; diagnoses based on sample size 211; diagnosing randomness 210; and expert judgments 213–214; prediction vs. diagnosis 209–210; probabilities vs. frequencies 217; revisiting 218–219; wider frameworks for 218–219
research: using electrodermal responses 259; using eye-movement tracking 152; using electroencephalography (EEG) 402, 419, 453; experimental 258–259; using fMRI 268, 402, 418, 419, 420, 453, 456; on illusory control 146–147; using magnetoencephalography 145; naturalistic 258; on suggestions 468–469; *see also* cognitive illusions research
retrieval fluency 198–199
retrieval-induced forgetting (RIF) 18, 446; and age 457, 459; blocking accounts of 460, 451, 453–455; classroom demonstration 448–450; cue independence 454; divided attention 452–453; experimental tasks 446–448;

final test phase 454–456; individual differences 456–457; inhibition accounts of 450–453, 457, 460; item recognition and implicit tests 454–455; mechanisms of 450–451; neural correlates 453, 455–456; retrieval specificity 451–452; retrieval-practice phase 451–453; and selective memory retrieval 457–459; working memory capacity 456–457
revelation effect 17; and affirmation theory 348–349; classroom demonstration 344–348; conclusions 352–353; and criterion-flux hypothesis 350–351; and decrement-to-familiarity hypothesis 349–350; and discrepancy-attribution hypothesis 351–352; experiment results 347–348; and global-matching hypothesis 349, 350; measures of 343–344; and no-task conditions 320, 340, 342–351; preceding tasks that cause 342–343; and recognition judgments 339–342; and task conditions 340, 342–350, 352–353; test-item presentation time 348; theoretical accounts 348–352
revision 6; of beliefs 14
Rhetoric (Aristotle) 205
richness-of-encoding (RE) hypothesis 364–365
risk aversion 84–86, 94
risk manipulation 86
risky-choice framing 80, 82, 84–86, 92, 93
rTMS (repetitive transcranial magnetic stimulation) 418
rule discovery task 102, 113

sample-size effect 121, 128
sampling: biased 436–437; importance of 297, 299, 303, 304, 310; natural 52; representative 9; selective 9, 213; sensory 304–305
sampling errors 12, 29, 37–38
sampling procedure 299
SARA (Selective Activation and Reconstructive Anchoring) 437
scale distortion 229, 235–237
schema and schemata 5, 7, 158–159, 219, 359; in memory 407–408; pragmatic reasoning 158, 162
schema priming effects 271
security motive 83, 92
selection tasks 157–158; *see also* Wason selection tasks
Selective Activation and Reconstructive Anchoring (SARA) 437
selective attention 91
Selective Processing Model 173
Selective Scrutiny Model 173
self-esteem 438–439
self-fulfilling prophecy 286, 307
self-illusion 480
self-presentation 430, 438
self-regulation 92
semantic priming 232
semantic priming models 398
sensory sampling model 304–305
Sensory Suggestibility Scale (SSK) 476
Sensory Suggestibility Test (SST) 476
sensory-discrimination tasks 304
sequential judgment paradigm 228, 236
shallow processing 478
signal detection 129
Signal Detection Theory (SDT) 173, 302, 343–344, 348
signaling, and overconfidence 307–308
signed summation 29–30, 38–39
similarity heuristic 11
size constancy tasks 301
social cognition 288
social judgment 227, 277
social psychology 128, 129, 277, 374, 472
sociology 80, 95
software: experiment-presentation 137, 449; multiTree 428, 433; Super Lab 137
source of activation confusion (SAC) 413
speed, memory for 381
stereotype effects 276; *see also* halo effects
stereotypes 57–59; gender 122–127; and halo effects 299; social and diagnostic 119
stereotypical knowledge 116
stereotyping 374; theories of 128, 129
stigmatization 374
stimulation models 16
stimuli: subliminal 262, 270, 271, 272; simple vs. complex 266; unfamiliar 269
stochastic fluctuations 304
stress 362, 480
Stroop task 456
subliminal stimuli 262, 270, 271, 272
substitution 7; attribute 71, 218
suggestibility 7, 475–476; and linguistic pragmatics 477–478
suggestibility scale 476
suggestion 7; research on 481
suggestionability 481
suggestionality 471, 474–475; suggestive situations 474–475; suggestive techniques 476–477; theory of 477; *see also* suggestibility

suggestive influences, taxonomy of 476–477
suggestive reaction strategies 472
suggestive situations 474–475
SuperLab 137
support 479–480; social 324
support theory 6, 26–28
surprise 438; conditional values 32
surprise heuristic 11
surprise index 294
surprise theory 29, 32–33
survival processing effect 17, 357–358; affective explanation 362–363, 367; classroom demonstration 358; evolutionary nature of 357–360; item congruity 363–364, 367; and planning 361–362, 367; possible underlying mechanisms 361–366; proximate explanations 361–362; and the richness of encoding 364–365, 367; single-item and relational processing 365–366, 368
syllogistic reasoning 5, 167–169; belief bias in 167–174; four kinds of 172
systematic framing effect 79

2–4–6 task 150–151
Take-The-Best heuristic 11
tasks: cognitive 456; experimental 446–448; extra-sensory perception 139; intellectual 6; output-interference 446, 458; perception 301; sensory-discrimination 304; size constancy 301; working-memory 351; *see also* Wason selection tasks
theories of personality, implicit theories of 279, 287
thinking illusions *see* cognitive illusions
thought listing 106–108
Thurstonian error and uncertainty 304
trading: and investment 311; overconfidence in 308
trait attributes 128
trait specificity 279
transient global amnesia 268
truth, perception of 247
truth bias 242, 252–253
truth effect 252–253
uncertainty, sources of 304
unconscious thinking 155
underconfidence 297, 301, 303, 305
unidirectional framing effect 85–86, 88; *see also* framing effect
unrealistic optimism effect 327–328
unrealistic optimism heuristic 192

validity: ecological 249–253; illusion of 213
validity effect 242, 272; as automatic process 245–246; classroom demonstration 243–245; in the college classroom 250–252; conclusion 253; and ecological validity 249–253; and hindsight bias 249; overview 243, 245; role of expertise 248–249; and the role of repetition 246, 248, 250, 253; role of source of information 247–248; types of information producing 246–247
verbal overshadowing 7, 383, 387, 470–471; conclusions 387; moderators of 386; replication project 386; on other senses 384–385; theoretical accounts 385–386; on visual objects 383–384; *see also* labelling effect
verbalization, effects of on memories 383–386
visio-spatial attention 253
visual illusions 479
vividness of information 191
voice recognition 384–385
von Restorff effect 120
vulnerability, judgments of 196–197

Wason selection tasks 103, 105–106, 112, 113, 150, 478; abstract 151–152; bias and rationality in human reasoning 159–162; as decision making 157–158; and dual-process theory 154–155; matching bias 152–154; realistic 155–156; theories developed with 158–159
'what is beautiful is good' effect 280, 283
working-memory capacity (WMC) 456–457, 459
writer competence, and the halo effect 277–278
Wundt's converse of the Herring illusion 264

Zollner illusion 264