

THIRD  
EDITION

# Cognitive Illusions

Intriguing Phenomena in Thinking,  
Judgment, and Memory

Edited by RÜDIGER F. POHL



# Cognitive Illusions

*Cognitive Illusions* explores a wide range of fascinating psychological effects in the way we think, judge, and remember in our everyday lives. In this volume, Rüdiger F. Pohl brings together leading international researchers to define what cognitive illusions are and discuss their theoretical status: Are such illusions proof of a faulty human information-processing system, or do they only represent by-products of otherwise adaptive cognitive mechanisms?

The book describes and discusses 26 different cognitive illusions, with each chapter giving a profound overview of the respective empirical research including potential explanations, individual differences, and relevant applied perspectives. This edition has been thoroughly updated throughout, featuring new chapters on negativity bias, metacognition, and how we respond to fake news, along with detailed descriptions of experiments that can be used as classroom demonstration in every chapter.

Demonstrating just how diverse cognitive illusions can be, it is a must read for all students and researchers of cognitive illusions, specifically, those focusing on thinking, reasoning, decision-making, and memory.

**Rüdiger F. Pohl** is retired Professor of Psychology at the University of Mannheim, Germany. His research interests include cognitive illusions, heuristics and decision-making, and autobiographical memory. Teaching psychology, he held lectures in all areas of Cognitive and Developmental Psychology as well as in History and Methods of Psychology.



Taylor & Francis  
Taylor & Francis Group  
<http://taylorandfrancis.com>

# **Cognitive Illusions**

Intriguing Phenomena in Thinking,  
Judgment, and Memory

*Third edition*

**Edited by**  
**Rüdiger F. Pohl**

Third edition published 2022  
by Routledge  
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

and by Routledge  
605 Third Avenue, New York, NY 10158

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

© 2022 selection and editorial matter, Rüdiger F. Pohl; individual chapters,  
the contributors

The right of Rüdiger F Pohl 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.

First edition published by Psychology Press 2004  
Second edition published by Routledge 2016

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

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

ISBN: 978-0-367-72425-2 (hbk)  
ISBN: 978-0-367-72424-5 (pbk)  
ISBN: 978-1-003-15473-0 (ebk)

DOI: 10.4324/9781003154730

Typeset in Bembo  
by Newgen Publishing UK

# Contents

<i>List of contributors</i>	viii
<b>Introduction</b>	1
1 What are cognitive illusions? RÜDIGER F. POHL	3
<b>PART I</b>	
<b>Thinking</b>	25
2 Conjunction fallacy JOHN E. FISK	27
3 Base-rate neglect GORDON PENNYCOOK, CHRISTIE NEWTON, AND VALERIE A. THOMPSON	44
4 Framing ANTON KÜHBERGER	61
5 Confirmation bias – myside bias HUGO MERCIER	78
6 Illusory correlation KLAUS FIEDLER, KAROLIN SALMEN, AND FLORIAN ERMARK	92
7 Causality bias HELENA MATUTE, FERNANDO BLANCO, AND MARÍA MANUELA MORENO-FERNÁNDEZ	108
8 Illusions of control SUZANNE C. THOMPSON	124

9 Wason selection task	140
JONATHAN ST. B. T. EVANS	
10 Belief bias in deductive reasoning	154
JONATHAN ST. B. T. EVANS, LINDEN J. BALL, AND VALERIE A. THOMPSON	
<b>PART II</b>	
<b>Judgment</b>	173
11 Availability	175
ANINE RIEGE AND ROLF REBER	
12 Judgments by representativeness	191
KARL H. TEIGEN	
13 Anchoring effect	209
ŠTĚPÁN BAHNÍK, THOMAS MUSSWEILER, AND FRITZ STRACK	
14 Illusory truth effect	225
LENA NADAREVIC	
15 Mere exposure effect	241
ROBERT F. BORNSTEIN AND CATHERINE CRAVER-LEMLEY	
16 Halo effects	259
SIMON M. LAHAM AND JOSEPH P. FORGAS	
17 Assumed similarity	272
ISABEL THIELMANN AND BENJAMIN E. HILBIG	
18 Overconfidence	287
ULRICH HOFFRAGE	
19 Metacognitive illusions	307
MONIKA UNDORF, SOFIA NAVARRO-BÁEZ, AND MALTE F. ZIMDAHL	
20 Fake news and participatory propaganda	324
STEPHAN LEWANDOWSKY	
21 Positivity biases	341
CARLA A. ZIMMERMAN AND W. RICHARD WALKER	

<b>PART III</b>	
<b>Memory</b>	357
22 Moses illusion	359
FELIX SPECKMANN AND CHRISTIAN UNKELBACH	
23 Survival processing effect	371
MEIKE KRONEISEN AND EDGAR ERDFELDER	
24 Labeling and overshadowing effects	386
RÜDIGER F. POHL	
25 Associative memory illusions	402
HENRY L. ROEDIGER, III, AND DAVID A. GALLO	
26 Misinformation effect	419
EMMA PECONGA, JACQUELINE E. PICKRELL, DANIEL M. BERNSTEIN, AND ELIZABETH F. LOFTUS	
27 Hindsight bias	436
RÜDIGER F. POHL AND EDGAR ERDFELDER	
<i>Author index</i>	455
<i>Subject index</i>	471

# Contributors

## **Štěpán Bahník**

The Prague College of Psychosocial Studies, and

Prague University of Economics and Business, Prague, Czech Republic  
<https://orcid.org/0000-0002-0579-6808>

## **Linden J. Ball**

University of Central Lancashire, Preston, Great Britain

<https://orcid.org/0000-0002-5099-0124>

## **Daniel M. Bernstein**

Kwantlen Polytechnic University, Surrey, British Columbia, Canada

<https://orcid.org/0000-0003-2716-2344>

## **Fernando Blanco**

University of Granada, Granada, Spain

<https://orcid.org/0000-0003-1283-8313>

## **Robert F. Bornstein**

Adelphi University, Garden City, NY, USA

<https://orcid.org/0000-0001-6203-225X>

## **Catherine Craver-Lemley**

Elizabethtown College, Elizabethtown, PA, USA

<https://orcid.org/0000-0002-8578-3539>

## **Edgar Erdfelder**

University of Mannheim, Mannheim, Germany

<https://orcid.org/0000-0003-1032-3981>

## **Florian Ermak**

University of Heidelberg, Heidelberg, Germany

<https://orcid.org/0000-0002-2788-856X>

## **Jonathan St. B. T. Evans**

University of Plymouth, Plymouth, Great Britain

<https://orcid.org/0000-0003-2679-573X>

**Klaus Fiedler**

University of Heidelberg, Heidelberg, Germany  
<https://orcid.org/0000-0002-3475-0868>

**John E. Fisk**

University of Central Lancashire, Preston, Great Britain  
<https://orcid.org/0000-0002-2981-0870>

**Joseph P. Forgas**

University of New South Wales, Sydney, Australia  
<https://orcid.org/0000-0001-8279-8367>

**David A. Gallo**

University of Chicago, Chicago, IL, USA  
<https://orcid.org/0000-0001-7669-6276>

**Benjamin E. Hilbig**

University of Koblenz-Landau, Landau, Germany  
<https://orcid.org/0000-0002-1470-1882>

**Ulrich Hoffrage**

University of Lausanne, Lausanne, Switzerland  
<https://orcid.org/0000-0002-5222-1564>

**Meike Kroneisen**

University of Koblenz-Landau, Landau, Germany  
<https://orcid.org/0000-0003-4325-5364>

**Anton Kühberger**

University of Salzburg, Salzburg, Austria  
<https://orcid.org/0000-0001-5786-5943>

**Simon M. Laham**

University of Melbourne, Melbourne, Australia  
<https://orcid.org/0000-0002-9101-9553>

**Stephan Lewandowsky**

University of Bristol, Bristol, Great Britain, and  
University of Western Australia, Perth, Australia  
<https://orcid.org/0000-0003-1655-2013>

**Elizabeth F. Loftus**

University of California, Irvine, CA, USA  
<https://orcid.org/0000-0002-2230-6110>

**Helena Matute**

University of Deusto, Bilbao, Spain  
<https://orcid.org/0000-0001-7221-1366>

**Hugo Mercier**

Institut Jean Nicod, Département d'études cognitives, ENS, EHESS,  
PSL University, CNRS, Paris, France  
<https://orcid.org/0000-0002-0575-7913>

**María Manuela Moreno-Fernández**

University of Granada, Granada, Spain  
<https://orcid.org/0000-0002-1747-4128>

**Thomas Mussweiler**

London Business School, London, UK  
<https://orcid.org/0000-0001-9373-4668>

**Lena Nadarevic**

University of Mannheim, Mannheim, Germany  
<https://orcid.org/0000-0003-1852-5019>

**Sofia Navarro-Báez**

University of Mannheim, Mannheim, Germany  
<https://orcid.org/0000-0002-6467-654X>

**Christie Newton**

University of Regina, Regina and Saskatoon, Saskatchewan, Canada  
<https://orcid.org/0000-0002-9280-2988>

**Emma PeConga**

University of Washington, Seattle, WA, USA  
<https://orcid.org/0000-0003-1831-3403>

**Gordon Pennycook**

University of Regina, Regina und Saskatoon, Saskatchewan, Canada  
<https://orcid.org/0000-0003-1344-6143>

**Jacqueline E. Pickrell**

University of Washington, Seattle, WA, USA  
<https://orcid.org/0000-0003-1250-3089>

**Rüdiger F. Pohl**

University of Mannheim, Mannheim, Germany  
<https://orcid.org/0000-0001-5354-8465>

**Rolf Reber**

University of Oslo, Oslo, Norway  
<https://orcid.org/0000-0002-6669-7109>

**Anine Riege**

University of Oslo, Oslo, Norway  
<https://orcid.org/0000-0002-2460-8665>

**Henry L. Roediger, III**

Washington University in St. Louis, St. Louis, MO, USA  
<https://orcid.org/0000-0002-3314-2895>

**Karolin Salmen**

University of Heidelberg, Heidelberg, Germany  
<https://orcid.org/0000-0001-7819-3920>

**Felix Speckmann**

University of Cologne, Cologne, Germany  
<https://orcid.org/0000-0002-6790-1693>

**Fritz Strack**

University of Würzburg, Würzburg, Germany  
<https://orcid.org/0000-0003-1892-7739>

**Karl H. Teigen**

University of Oslo, Oslo, Norway  
<https://orcid.org/0000-0001-6165-9034>

**Isabel Thielmann**

University of Koblenz-Landau, Landau, Germany  
<https://orcid.org/0000-0002-9071-5709>

**Suzanne C. Thompson**

Pomona College, Claremont, CA, USA  
<https://orcid.org/0000-0003-4105-9415>

**Valerie A. Thompson**

University of Saskatchewan, Saskatoon, SK, Canada  
<https://orcid.org/0000-0003-1676-9458>

**Monika Undorf**

University of Mannheim, Mannheim, Germany  
<https://orcid.org/0000-0002-0118-824X>

**Christian Unkelbach**

University of Cologne, Cologne, Germany  
<https://orcid.org/0000-0002-3793-6246>

**W. Richard Walker**

Colorado State University, Pueblo, CO, USA  
<https://orcid.org/0000-0003-4405-7335>

**Malte F. Zimdahl**

University of Mannheim, Mannheim, Germany  
<https://orcid.org/0000-0001-5808-2967>

**Carla A. Zimmerman**

Colorado State University-Pueblo, Pueblo, CO, USA  
<https://orcid.org/0000-0002-9585-4679>

# **Introduction**



Taylor & Francis  
Taylor & Francis Group  
<http://taylorandfrancis.com>

# 1 What are cognitive illusions?

Rüdiger F. Pohl

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. Starting in the 1970s, continuing until today, a number of researchers have published collections of such biases and fallacies. The sheer number of these phenomena led some people to endorse a rather pessimistic view of human cognitive abilities. This, in turn, led others to doubt the experimental procedures or normative models, suggesting that the observed biases may have been artificially produced by tricking participants or by applying wrong standards. In addition, it was questioned whether these “illusions” have any important real-world consequences. These arguments have triggered a lively debate that is still ongoing and reflected on by most of the authors of this book.

In this introductory chapter, I first describe what constitutes the domain of cognitive illusions. Then, I discuss how cognitive illusions can be characterized, how they can be explained, and what follows for applied and functional views. Finally I describe the intentions of this book and what has changed from the last edition.

## The domain of cognitive illusions

### Categories of cognitive illusions

Cognitive illusions can be roughly ordered 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 focused 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 to be responsible not only for biases in judgment but also for several memory illusions.

Illusions of *thinking* (Chapters 2 to 10) are 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 compute a probability from given values, to verify a logical conclusion, or to discover a hidden rule.

In illusions of *judgment* (Chapters 11 to 21), 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. Tversky and Kahneman's (1974) heuristics – availability, representativeness, and anchoring (Chapters 11 to 13, 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 22 to 27) 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, to-be-remembered material proceeds under experimental control (except for the Moses illusion, Chapter 22). In these cases, it is easier to compare the given materials to the later produced recollections.

### **Other collections**

Several other books focus on cognitive illusions or biases. One such comprehensive collection with examples from thinking, judgment, and memory was published in Germany almost 30 years ago by Hell et al. (1993). In addition to some of the illusions included here, Hell et al. also covered problems of attributing causality (see Chapter 7) and developmental misconceptions of physics. Much more recently, Benson (2016; see also Gardner, 2017) presented a collection of 188 fallacies in his "Cognitive bias cheat sheet", including phenomena from all three areas, that is, thinking, judgment, and memory. In addition he sorted these biases into four broad areas, according to their underlying problems, namely, "too much information", "not enough meaning", "need to act fast", and "what should we remember?" Other collections focused on illusions that were related either to both thinking and judgment or to memory alone and are described in the following sections.

### *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 titled "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 et al. (1982) under the same title as the 1974 paper. Later developments of this approach are documented in "Heuristics and biases: The psychology of intuitive judgment" edited by Gilovich et al. (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 titled “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 et al. (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), Baron (2007), and Evans (2007) listed about 50 different biases each.

More recently, Dobelli (2014) presented 100 ways in which humans allegedly behave irrationally. Similarly, Arp et al. (2019) edited a book on *Bad arguments*, covering 100 logical fallacies in western philosophical thinking. The fallacies are organized into two parts with further sub-categories, namely formal fallacies (propositional and categorical logic) and informal ones (fallacies of relevance, ambiguity, and presumption).

With respect to medical decision-making, Howard (2019) looked at a number of “Cognitive errors and diagnostic mistakes”, several of which are also covered in this book. He discussed them in the context of clinical case-studies. Taking a different view on cognitive biases, Korteling and Toet (2021) discussed 46 fallacies in judgment and decision-making, but less from a psychological viewpoint (in which they see some shortcomings), than from a neuro-evolutionary one. They argue that the observed fallacies “originate from the inherent design characteristics of our brain” (p. 1) that evolved to enable survival of our remote ancestors.

### *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.

More than a century later, Schacter et al. (1995) edited a book on *Memory distortions: How minds, brains, and societies reconstruct the past*, and Roediger (1996) introduced a special issue of the *Journal of Memory and Language* focusing 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 Molz & Pohl, 2017).

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).

More recently, Shaw (2017) presented a number of memory shortcomings, most of which may not only have personal significance, but may also have implications in applied contexts (e.g., in forensic investigations). As implied by the title of her book (*The memory illusion: Remembering, forgetting, and the science of false memory*) the focus of the book is more on lay conceptions of how memory functions than on all potential cases of systematic errors. In addition, she looks at the interplay between autobiographical memory and personal identity.

## **The status of cognitive illusions**

The quite diverse and comprehensive collections mentioned above already suggest that there is no agreed-upon fixed set of cognitive illusions. In other words, it is not clear which phenomena may count as an illusion (or a bias) and which not. In this section, I start by delineating what could be considered criteria for a definition of cognitive illusions in a stricter sense. However, such a definition quickly reveals that many, if not most, of the cognitive illusions do not conform. This reminds one of the massive critique most notably put forward by Gigerenzer (1991, 1996, 2004; Gigerenzer et al., 2008), especially against the so called “heuristics and biases” approach (see, e.g., Gilovich et al., 2002; Tversky & Kahneman, 1974). According to that approach, mental shortcuts were responsible for many biases. Gigerenzer, however, argued that many of the reported biases are based on faulty experimental procedures and questionable theoretical assumptions. He even suspected a “bias bias”, that is, a tendency to observe biases where actually are none (Gigerenzer, 2018). As a consequence, the domain of cognitive illusions apparently needs to be understood as a field of rather diverse phenomena. But let us take one step at a time.

### **Defining features**

The term “cognitive illusion” has evolved in analogy to the better known domain of “optical illusions” (see Hell, 1993; Roediger, 1996; see also Gigerenzer, 2008), thus suggesting that the observed phenomena are clear, robust, universal, and impossible to avoid. To define what should count as a cognitive illusion (in a stricter sense) and what not, I discuss five criteria (see Text box 1.1). Note that these criteria slightly deviate from those given in the last edition of this book (Pohl, 2017).

#### **Text box 1.1**

Features of a potential definition of cognitive illusions (in analogy to optical illusions)

1. Deviation from reality (i.e., an accepted normative standard)
2. Systematic deviation from the standard (i.e., in a predictable direction)
3. Involuntary (unconscious) production of the illusion
4. Impossibility to avoid (or reduce) the illusion
5. Universal appearance of the illusion

The first feature of a phenomenon to count as an illusion 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. So, for example, in optical illusions, Line A is consistently perceived as longer than Line B (not shorter). Or, in illusions of thinking, a conditional probability is consistently estimated as larger than it normatively is (not smaller). However, given that cognitive processes are complex and involve a number of mental steps, not each and every answer shows the expected deviation. A systematic and robust bias may thus only become evident if the data are summed across a larger number of trials or participants. Presumably, the range of these answers is typically much smaller for optical illusions than for cognitive ones, suggesting that the latter include more moderating factors than the former.

A third aspect of cognitive illusions is that they appear *involuntarily*, that is, without specific instructions or deliberate will. People do not consciously aim to produce biased answers. These answers just happen. This is analogous to what has been found in research on suggestions (see Molz & Pohl, 2017): 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 et al., 2002, for further discussion). Another aspect is that persons who fell 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, illusional persons are still convinced they 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 appears to be much weaker for cognitive ones. For some illusions, most (if not all) attempts to overcome the effect have failed (as an example, see Pohl & Hell, 1996, and Chapter 27), thus conforming to the proposed criterion. For others, however, research has shown that certain experimental manipulations may reduce or even eliminate the illusion (as an example, see Gigerenzer et al., 2008). These results again suggest that cognitive illusions involve more malleable processes than optical illusions and that they do not reflect pure effects of our cognitive hardware.

Finally, the fifth criterion to be considered is that an illusion is *universal*. That is, it should be observed for all people independently from any personality features (e.g., intelligence). Only then could we conclude that the illusion reflects a basic feature of the human cognitive architecture and is not based on some idiosyncratic peculiarities of specific persons. Yet, research has consistently found a number of individual differences leading to varying degrees of cognitive illusions. Let us take the age effect in hindsight bias as an example: Children and older adults typically exhibit larger hindsight bias than middle-aged adults do (see Chapter 27). Interestingly, some optical illusions, too, depend on age (Leibowitz & Gwozdecki, 1967; Leibowitz & Judisch, 1967). Thus, developmental (and possibly other) individual differences in our mental architecture can have an impact on optical as well as cognitive illusions.

In sum, the attempt to define cognitive illusions in analogy to optical illusions has shown some striking differences between the two domains (see Gigerenzer, 2008). Presumably, only a few, if any, cognitive illusions fulfill the given definition in a strict sense, whereas most of them appear much more influenceable and dependent on a number of factors. In a similar vein, but for other reasons, Gigerenzer (1991, 1996, 2004, 2018; Gigerenzer et al., 2008) has repeatedly questioned the existence of several cognitive illusions, especially those that were supposed to be based on the classical heuristics of availability, representativeness, and anchoring (see Chapters 11, 12, and 13, respectively; Tversky & Kahneman, 1974). Gigerenzer's arguments can, however, be extended to the whole domain, especially when considering the surprisingly huge number of "illusions" reported in several publications. It would be quite informative to know which ones represent "hard-core" illusions and which ones simply reflect clever tricks (or wrongdoing) of the experimenter. So let us turn to Gigerenzer's points next.

### **Gigerenzer's critique**

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 approach led to an enormous number of studies investigating the intricacies of human judgment (e.g., Kahneman et al., 1982, or later, Gilovich et al., 2002), which in turn affected scholarship in psychology, 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”, leading to a one-sided view of human rationality (as being profoundly faulty) and a lack of explanatory power (by only redescribing the observed phenomena). Another major point of his criticism asserted that many of the reported demonstrations of cognitive biases are simply artificial (based on several questionable experimental features). Text box 1.2 summarizes the main critical points (see also the replies of Kahneman & Tversky, 1996, and Gilovich & Griffin, 2002). Kelman (2011) collected a number of arguments of both sides in his book on the heuristic debate (see also Polonioli, 2016). I focus here on the accusation of artificiality.

### **Text box 1.2**

Main critical points from Gigerenzer (1991, 1996, 2004; Gigerenzer et al., 2008) against the “heuristics and biases” approach that may apply to other cognitive illusions as well

1. One-sided view
  - Focusing too much on errors instead of successful solutions
  - Leading to a too pessimistic view of human thinking
2. Lack of explanatory power
  - Only redescribing the phenomena with post-hoc “explanations”
  - Not making clear predictions, thus not enabling rigorous testing
3. Artificiality of the experimental situation
  - Intentional deception
  - Inadequate material selection
  - Inadequate problem representation
  - Missing knowledge/ability
  - Questionable normative standard

The main reason leading to artificial illusions, according to Gigerenzer, was that the experimental paradigms and materials were themselves artificial. This includes (1) that participants were tricked into biased behavior (e.g., by strong suggestions or misleading information), (2) that materials were selectively, instead of representatively, sampled, (3) that problems, even if appropriate, were presented in an inadequate statistical format (e.g., with probabilities instead of frequencies), (4) that participants may simply lack the knowledge or practice to reach the correct answer (but could easily be taught to), and finally (5) that participants’ behavior was compared to wrong normative standards.

### Misleading information

Examples for strongly misleading or at least highly suggestive procedures are certain rule-discovery tasks (see Chapters 5, 9, and 10). In a typical such task, participants are asked to discover a hidden rule underneath, for example, a series of numbers, like 2–4–6 (Wason, 1960). To test the suspected rule, they are asked to produce further number series which will then be classified as correct or wrong by the experimenter. Typically, participants produce highly similar series (like 8–10–12), which are confirmed as “correct”, but then propose the apparent rule “ascending even numbers” that is, however, declined as “wrong”. The trick is that the hidden rule is much more general, namely simply “ascending numbers”. The given example “2–4–6”, however, suggests a rather specific rule and is thus highly misleading and not representative of the hidden rule. Another example of this sort is given in Text box 1.3 (adapted from a German version by Dario Bagatto; [www.hirnwindungen.de](http://www.hirnwindungen.de)). A further highly suggestive procedure is used in some labeling tasks (see Chapter 24). For example, after viewing a video depicting a car accident, participants were asked “About how fast were the cars going, when they *contacted* each other?” (Loftus & Palmer, 1974). In another condition, participants were asked “About how fast were the cars going, when they *smashed* into each other?” Not too surprisingly, speed estimates were higher in the second as compared to the first condition. The wording here was highly suggestive, thus biasing the estimates.

### Text box 1.3

A misleading example in a rule-discovery task

- Medieval age: A spy wants to enter a city, but the city gate is guarded and a password is needed. So he hides and observes the scene.
- A salesman comes by and as he approaches the gate the guard asks him: “12 – what is your answer?” The salesman answers “6” – and is allowed to enter.
- A short time later, a priest approaches and is questioned by the guard: “6 – what is your answer?” The priest replies “3” and the guard lets him pass into the city.
- The spy feels certain that he understands how the password works. So he walks over to the guard who asks him: “24 – what is your answer?” But as the spy answers “12”, he is immediately arrested and put into jail.
- What should he have answered to safely enter the city?

### Inadequate presentation format and material selection

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 & Ortman, 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”, may not be very noteworthy. The result just demonstrates that the experimenter was successful in fooling his participants, but not that the underlying cognitive processes are biased in any way. In a different approach, Pohl et al. (2017) have shown that participants based their inferences on underlying features of the material’s domain, not on any biased selection,

so that participants' behavior did not correspond to the features of the selected materials. 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 2), base-rate neglect (see Chapter 3), and overconfidence (see Chapter 18).

### *Missing knowledge*

Missing knowledge is another reason that could lead to artificial biases (Gigerenzer, 2004). Of course, in the domain of judgments under uncertainty, missing knowledge is an inherent part of the problem that can't be overcome easily. However, in other tasks, missing knowledge could easily be remedied by teaching participants the appropriate knowledge or skills, thus avoiding subsequent biases. As an example, consider the conjunction fallacy (Chapter 2). Everyone can quickly learn how two independent probabilities are combined (namely, by computing their product). In that case, any bias would not reflect basic cognitive hard-wired deficits, but simply certain states of ignorance or incapability. To study these biases could, of course, be justified in their own right, but they surely represent something different than cognitive illusions in a stricter sense.

### *Wrong normative standards*

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. Many researchers have accordingly discussed which normative models should be used as standard of comparison, eventually leading to some changes (see further below).

### *Further causes for artificial results*

In a later publication, Gigerenzer (2004) extended his arguments to even more phenomena, 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 the problematic features 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, and probability matching. Especially regression to the mean seems to be an influential factor that is often not taken into account and thus may lead to observation of "biases" where actually there are none, that is, to misinterpret random error as systematic error (Gigerenzer, 2018). Friedman (1992) had already suspected that "the regression fallacy is the most common fallacy in the statistical analysis of economic data" (p. 2131), but it certainly pertains also to other domains.

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 from the heuristics and biases domain.

### ***Normative standards***

Researchers have continuously disputed which models might be used as norms and which not (Caverni et al., 1990b, pp. 8–9):

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.

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 detailed 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 researches 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). In my view the 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.

### ***Summary of the debate***

Research has abundantly shown that many different factors influence whether a cognitive bias is observed and how large it is, whereas optical illusions can hardly be influenced

or trained to disappear. The chapters in this book present a wealth of empirical evidence showing which moderators influence the strength of any illusion and which individual differences exist. Thus it is clear that cognitive illusions span a wide range of rather diverse phenomena and that many of them are much more malleable than optical illusions are (cf. the “heterogeneity hypothesis” from Polonioli, 2016; see also Arkes, 1991). Probably only a few of them represent “hard-wired” illusions in a stricter sense (as defined above). In addition, several illusions may profit from rather artificial conditions of eliciting them or wrong normative standards (as Gigerenzer has argued). So it would probably be a good idea to strip any observed phenomenon from its artificial superstructure and reduce it to its more or less “pure” cognitive core (if any). This would also help in clarifying potential explanations (see below). However, differently from Gigerenzer who simply dismissed all “artificial” biases, I would argue that all biases, regardless of their epistemological status, should be considered, albeit under increased scrutiny concerning their origin, and that all of them can be informative about human information processing. Even magicians who play unfair tricks on us can tell us a lot about attention, memory, and thinking. So, in general, the fact that we can be influenced by artificial situations is in itself a noteworthy finding (even if it tells us little about human rationality).

## **Explanations of cognitive illusions**

In the past 30 years, many researchers endorsed a dual-process approach to explain cognitive illusions. However, the used concepts and definitions are quite diverse and are still under strong debate (e.g., Evans, 2019; DeNeys, 2021, and Dewey, 2021). A much simpler approach is to take general information-processing characteristics into account that govern our everyday thinking, judgment, and memory (Benson, 2016; see also Gardner, 2017).

### **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). Evans (1989) had originally labeled these two systems as “heuristic” and “analytic” and later changed them into “intuitive” and “reflective”. 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, pp. 16–17).

Such an approach appears quite plausible at first glance. However, many and quite diverse dual-process models have been proposed (and repeatedly changed), thus perhaps adding more confusion than clarity. 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 (potentially faulty) heuristics to single systems even more difficult (see Evans, 2019).

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 (“one-process”) 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 heuristic processing may occur in both systems and, as a consequence, cognitive biases may stem from both. In a similar vein, DeNeys (2021) argued that dual- and single-process models are indistinguishable and, even if they were, would not advance our understanding of human thinking (but see the reply by Dewey, 2021).

### ***Information-processing system***

Kruglanski and Ajzen (1983) categorized biases in human judgment as either motivational or cognitive, thus reflecting the various influences discussed above: “Motivational biases are characterized by a tendency to form and hold beliefs that serve the individual’s needs and desires” (p. 4), whereas cognitive biases “originate in the limitations of otherwise reasonable information-processing strategies” (p. 4). More specifically, they assumed that these strategies

direct people’s attention to some types of information and hypotheses and [...] lead to an underestimation or disregard of other information and hypotheses which, although relevant to the judgment in question, have no ready place in the information-processing strategy that is being employed.

(p. 4)

Stanovich (2008) suggested a taxonomy of biases and classified them according to his dual-process approach (see above) 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).

#### **Text box 1.4**

Information-processing mechanisms proposed by Benson (2016)

(1) Information overload

“There is just too much information in the world, we have no choice but to filter almost all of it out.”

(2) Lack of meaning

“The world is very confusing, and we end up only seeing a tiny sliver of it, but we need to make some sense of it in order to survive. Once the reduced stream of

information comes in, we connect the dots, fill in the gaps with stuff we already think we know, and update our mental models of the world.”

### (3) The need to act fast

“We’re constrained by time and information, and yet we can’t let that paralyze us. Without the ability to act fast in the face of uncertainty, we surely would have perished as a species long ago. With every piece of new information, we need to do our best to assess our ability to affect the situation, apply it to decisions, simulate the future to predict what might happen next, and otherwise act on our new insight.”

### (4) How to know what needs to be remembered

“We can only afford to keep around the bits that are most likely to prove useful in the future. We need to make constant bets and trade-offs around what we try to remember and what we forget.”

*Note:* Quotations retrieved from an online source without page numbers: <https://betterhumans.pub/cognitive-bias-cheat-sheet-55a472476b18#.8nptqnepc>

More recently, Benson (2016) proposed just four mechanisms (or problems) of human information processing that he considered responsible for no less than 188 cognitive biases (see also Gardner, 2017)! These mechanisms are given in Text box 1.4. Although the descriptions appear a bit sloppy they nevertheless target the central mechanisms: (1) We generally need to extract and select information for further processing. Several mechanisms influence this process, like previously encoded information, discrepant information, or goal-related importance. We also abstract and simplify information, thus losing further details. (2) We always need to look under the surface to extract meaning. For example, we use schemata, stereotypes, and prior experiences to fill in the gaps. In addition, we interpret information based on our own (possibly faulty) knowledge and beliefs and tend to project this understanding into the past and the future, as well as onto other people. (3) There is not always enough time to consider just any arguments or information. In addition, we also often prefer quick decisions or judgments. As a consequence, the results of these processes under time constraints might deviate from an optimal solution. However, we are usually quite confident in our behavior (may be more than is warranted). (4) In addition to the selection of information (the first point), another problem concerns deciding what to remember and then storing it in long-term-memory. Again, several factors influence these processes, like available knowledge, motivation, or emotion. Further mechanisms govern (and possibly distort) the recall and reconstruction processes.

In sum, all these mechanisms and processes are well known from research on information processing and they are apparently sufficient to explain a multitude of cognitive illusions, without any need to postulate additional specific processes. They also underline the conclusion already drawn, namely that the domain of cognitive illusions consists of rather heterogeneous phenomena based on different processes during information processing.

## Implications of cognitive illusions

### *Applied perspectives*

With respect to their ecological validity, cognitive biases and illusions 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; Shaw, 2017). Most of the authors in the present book take up this issue in their chapter and provide compelling evidence. And the importance in applied contexts even holds if the illusion does not conform to the strict characterization given above, but is rather based on artificial deceit.

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 et al., 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 et al. (2000) aimed to reduce false memories or misattributions (cf. Chapter 25) 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 (e.g., Gigerenzer et al., 2008), whereas others proved their robustness (e.g., Pohl & Hell, 1996). In the domain of health-related judgments and decision-making, Ludolph and Schulz (2017) presented an overview of 87 studies showing that most debiasing techniques were at least partially effective. Sellier et al. (2019) recently showed that the confirmation bias (Chapter 5) can be effectively reduced in a field setting through a one-shot intervention.

Two further examples come from clinical psychology. For some time, the obsessive-compulsive disorder (OCD) has been linked to cognitive illusions (e.g., Moritz & Pohl, 2009). Dettore 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. (2016) 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.

### *Functional views*

Closely linked to normative issues (as discussed above) 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 (or costs) of otherwise functional processes, or (c) 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 (2017, p. 128; see also Chapter 6) put it. Similarly, Tversky and Kahneman (1974) acknowledged that heuristics typically lead to useful results, but they may, of course, fail at times. Similarly, memory-based illusions (like hindsight bias; Chapter 27) can be understood as a by-product of the highly functional process of knowledge updating (see Hoffrage et al., 2000; Nestler et al., 2012). Schacter et al. (2011) considered several such illusions, namely the associative memory illusion (Chapter 25), the effect of imagination inflation (Chapter 26), and the misinformation effect (also Chapter 26), 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; Polonioli, 2016).

Some researchers, however, went one step further and asserted that the illusions themselves could possess adaptive value (cf. Chapters 18 and 21; Molz & Pohl, 2017), not just their underlying mechanisms. 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.

Another example is the confirmation bias (Chapter 5) that many researchers consider to possess adaptive value (e.g., Peters, 2020).

With respect to memory, 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 focused on “positive illusions” (see Chapter 21) which in their view are highly adaptive. Sutton (2009) further noted that (1) the positivity bias in autobiographical memory is related to enhanced emotion regulation, that (2) forgetting is also an adaptive response, and that (3) “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 with 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

The selection of cognitive illusions in this book is rather subjective. I included some that are more well known and covered in almost any cognitive textbook, but I also included some that are less familiar but may hold valuable insights. Of course, I also included those that I had focused on during my own academic career (Chapters 13, 26, and 27).

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 (e.g., Pohl et al., 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 experimental 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 (see also Molz & Pohl, 2017).

Besides its hoped-for 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 (like 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 is to use it as a *textbook* for classes of cognitive psychology. To this end, each chapter follows the same general outline and contains a survey of empirical findings (including individual differences), a discussion of theoretical explanations, an overview of applied perspectives, and – last not least – a ready-to-run 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 details that are necessary to conduct and analyze the described experiment. This concerns materials, instructions, procedures, and statistical tests. Thus it should be fairly easy to incorporate these experiments into a class on “cognitive illusions”.

### ***Changes to the second edition***

I first checked the literature on which of the phenomena described in the second edition still yielded currently active research, and which new fields had emerged in the past years. As a consequence, I decided to drop a few of the old chapters, namely those on probability matching (Newell & Schulze, 2017), the revelation effect (Aßfalg, 2017), retrieval-induced forgetting (Kliegl & Bäuml, 2017), and suggestibility (Molz & Pohl, 2017). Of course, this selection is highly subjective and says little to nothing about the importance of these phenomena. Two other topics were continued, but by new authors, namely the illusory truth effect in Chapter 14 (formerly the validity effect; Renner, 2017) and positivity biases in Chapter 21 (formerly the Pollyanna principle; Matlin, 2017). And finally, I included some new topics that I considered highly interesting, namely the causality bias (Chapter 7), the assumed similarity (Chapter 17), metacognitive illusions (Chapter 19), fake news and participatory propaganda (Chapter 20), and the Moses illusion (Chapter 22). The last topic had already been part of the first edition (Park & Reder, 2004). All other chapters (including this introduction) were thoroughly updated by their respective authors, reflecting on new empirical data, theoretical approaches, and applied perspectives.

### **Summary**

- There are many collections of cognitive illusions and biases in the domains of thinking, judgment, and memory.
- Cognitive illusions can be characterized with five features: (1) Behavior deviates from some standard, (2) it deviates in a systematic fashion, (3) cognitive illusions appear involuntarily, (4) they are hard if not impossible to avoid, and (5) they are universal.
- In a first approach, many biases were explained with a few judgmental heuristics, namely availability, representativeness, and anchoring.
- Gigerenzer criticized this approach as being too pessimistic about human capabilities, of little explanatory value, and often based on artificial experimental procedures.
- Cognitive illusions should be understood as spanning a wide range of phenomena, from artificial deceit up to “hard-wired” and unavoidable biases.
- Accordingly, some cognitive illusions can be reduced or even eliminated, whereas others can hardly be influenced.
- Cognitive illusions can be explained with simple information-processing strategies.
- Many cognitive illusions do have real-world consequences.
- Whereas most cognitive illusions can be seen as the (bearable) costs of an otherwise highly functional system, some illusions might even possess adaptive value.

### **Further reading**

Still a good start into the field is the classical paper by Tversky and Kahneman (1974) on “heuristics and biases” that led to an uncountable number of studies and also to substantial debate about experimental procedures and theoretical underpinnings, mainly voiced by Gigerenzer et al. (2008).

Polonioli (2016) provided a good discussion and overview concerning the explanation and diversity of biases. The dual-process approach is thoroughly treated in Kahneman's (2012) book on *Thinking: Fast and slow*, but still elicits critical discussions (e.g., DeNeys, 2021; Evans, 2019). A neuro-evolutionary framework of understanding cognitive biases, not discussed here, can be found in Korteling and Toet (2021).

## References

- Aßfalg, A. (2017). Revelation effect. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (2nd ed., pp. 339–356). London: Routledge.
- Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, 110(3), 486–498.
- Arp, R., Barbone, S., & Bruce, M. (Eds.). (2019). *Bad arguments: 100 of the most important fallacies in Western philosophy*. Oxford: Wiley Blackwell.
- Baron, J. (2007). *Thinking and deciding*. New York: Cambridge University Press.
- 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.
- Benson, B. (2016). Cognitive bias cheat sheet. *Better Humans*. Retrieved from <https://betterhumans.pub/cognitive-bias-cheat-sheet-55a472476b18#.8nptqnepc> on Aug. 11, 2021.
- Boyer, P. (2009). Extending the range of adaptive misbelief: Memory “distortions” as functional features. *Behavioral and Brain Sciences*, 32, 513–514.
- 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.
- De Neys, W. (2021). On dual- and single-process models of thinking. *Perspectives on Psychological Science*, 16(6), 1412–1427.
- Dettorre, D., & O'Connor, K. (2013). OCD and cognitive illusions. *Cognitive Therapy and Research*, 37(1), 109–121.
- Dewey, C. (2021). Reframing single- and dual-process theories as cognitive models: Commentary on De Neys (2021). *Perspectives on Psychological Science*, 16(6), 1428–1431.
- Dobelli, R. (2014). *The art of thinking clearly*. London: Harper.
- 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.
- Evans, J. St. B. T. (2019). Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, 25(4), 383–415.

- Fiedler, K. (2017). Illusory correlation. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (2nd ed., pp. 115–133). London: Routledge.
- 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.
- Friedman, M. (1992). Do old fallacies ever die? *Journal of Economic Literature*, 30, 2129–2132.
- Gardner, J. (2017). *Cognitive biases directory*. Melbourne: Sage Decisions.
- 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, UK: 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. (2008). Why heuristics work. *Perspectives on Psychological Science*, 3(1), 20–29.
- Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics*, 5(3–4), 303–336.
- 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., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4), 684–704.
- Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: 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.
- 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: Elsevier.
- Hell, W. (1993). Kognitive und optische Täuschungen [Cognitive and optical illusions]. In W. Hell, K. Fiedler, & G. Gigerenzer (Eds.), *Kognitive Täuschungen* [Cognitive illusions] (pp. 317–324). Heidelberg: Spektrum der Wissenschaften.
- Hell, W., Fiedler, K., & Gigerenzer, G. (Eds.). (1993). *Kognitive Täuschungen* [Cognitive illusions]. Heidelberg: Spektrum der Wissenschaft.
- 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: Springer.
- 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.
- Howard, J. (2019). *Cognitive errors and diagnostic mistakes: A case-based guide to critical thinking in medicine*. Cham: Springer International Publishing.
- Kahneman, D. (2012). *Thinking, fast and slow*. London: Penguin Books.
- 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.
- Kelman, M. (2011). *The heuristics debate*. New York: Oxford University Press.
- 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.

- Kliegl, O., & Bäuml, K.-H. T. (2017). Retrieval-induced forgetting. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (2nd ed., pp. 446–464). London: Routledge.
- Korteling, J. E., & Toet, A. (2021). Cognitive biases. In S. Della Sala (Ed.), *Encyclopedia of behavioral neuroscience* (2nd ed., pp. 610–619). Amsterdam: Elsevier.
- 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., & Ajzen, I. (1983). Bias and error in human judgment. *European Journal of Social Psychology*, 13(1), 1–44.
- Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, 118(1), 97–109.
- Leibowitz, H. W., & Gwozdecki, J. (1967). The magnitude of the Poggendorf illusion as a function of age. *Child Development*, 38, 573–580.
- Leibowitz, H. W., & Judisch, J. M. (1967). The relation between age and the magnitude of the Ponzo illusion. *American Journal of Psychology*, 80, 105–109.
- 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(5), 585–589.
- Ludolph, R., & Schulz, P. J. (2017). Debiasing health-related judgments and decision making: A systematic review. *Medical Decision Making*, 38(1), 3–13.
- Matlin, M. W. (2017). Pollyanna principle. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (2nd ed., pp. 315–335). London: Routledge.
- McKay, R. T., & Dennett, D. C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, 32, 493–561.
- Molz, G., & Pohl, R. F. (2017). Suggestion and cognitive illusions. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (2nd ed., pp. 467–484). London: Routledge.
- 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.
- 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.
- Newell, B. R., & Schulze, C. (2017). Probability matching. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (2nd ed., pp. 62–78). London: Routledge.
- 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* (1st ed., pp. 275–291). Hove, UK: Psychology Press.
- Peters, U. (2020). What is the function of confirmation bias? *Erkenntnis*. <https://doi.org/10.1007/s10670-020-00252-1>
- Piatelli-Palmarini, M. (1994). *Inevitable illusions: How mistakes of reason rule our minds*. New York: Wiley.
- Pohl, R. F. (2017). Labelling and overshadowing effects. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (2nd ed., pp. 373–389). London: Routledge.
- 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.
- Pohl, R. F., Michalkiewicz, M., Erdfelder, E., & Hilbig, B. E. (2017). Use of the recognition heuristic depends on the domain's recognition validity, not on the recognition validity of selected sets of objects. *Memory & Cognition*, 45(5), 776–791.

- Polonioli, A. (2016). Adaptive rationality, biases, and the heterogeneity hypothesis. *Review of Philosophy and Psychology*, 7(4), 787–803.
- Renner, C. H. (2017). Validity effect. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (2nd ed., pp. 242–255). London: Routledge.
- 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.
- Sellier, A.-L., Scopelliti, I., & Morewedge, C. K. (2019). Debiasing training improves decision making in the field. *Psychological Science*, 30(9), 1371–1379.
- Shaw, J. (2017). *The memory illusion: Remembering, forgetting, and the science of false memory*. New York: Random House Books.
- 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.
- 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.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
- Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3), 129–140.
- 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* (1st ed., pp. 379–396). Hove, UK: Psychology Press.
- Withköft, M., Kerstner, T., Ofer, J., Mier, D., Rist, F., Diener, C., & Bailer, J. (2016). Cognitive biases in pathological health anxiety: The contribution of attention, memory, and evaluation processes. *Clinical Psychological Science*, 4(3), 464–479.



Taylor & Francis  
Taylor & Francis Group  
<http://taylorandfrancis.com>

**Part I**

# **Thinking**



Taylor & Francis  
Taylor & Francis Group  
<http://taylorandfrancis.com>

## 2 Conjunction fallacy

*John E. Fisk*

Violations of the rules of probability theory and associated systematic reasoning biases have been widely demonstrated. One area that has received much attention in the research literature concerns conjunction-rule violations. Formally, the conjunction rule may be expressed as follows:

$$P(A \& B) = P(A) \times P(B | A) \quad (2.1)$$

Thus, 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:

$$P(\text{Study and Pass}) = P(\text{Study}) \times P(\text{Pass} | \text{Study}) \quad (2.2)$$

When the two events A and B are independent then Equation 2.1 simplifies to

$$P(A \& B) = P(A) \times P(B) \quad (2.3)$$

since for independent events:

$$P(B) = P(B | A) = P(B | \text{not } A) \quad (2.4)$$

The extent to which individuals make judgments 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 et al., 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; Wells, 1985; Yates & Carlson, 1986). Looking at Equation 2.1,  $P(A \& B) \leq P(A)$  since, by definition,  $P(B | A) \leq 1$  and by extension, it follows that  $P(A \& B) \leq P(B)$ . Thus the fallacy occurs when the conjunctive probability is assigned a value exceeding that assigned to one or both of the component events, that is,

$$P(A \& B) > P(A) \text{ and/or} \quad (2.5)$$

$$P(A \& B) > P(B), \quad (2.6)$$

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:

- Linda is active in the feminist movement.
- Linda is a bank teller.
- 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; Hertwig & Chase, 1998; Nilsson et al., 2009; Rogers et al., 2011; Tentori et al., 2013; Yates & Carlson, 1986).

It is important to note that the fallacy is most common in cases where the conjunction contains a likely event (e.g., feminist) paired with an unlikely event (e.g., bank teller). In other contexts, the fallacy is much reduced, for example, where the conjunction contains two unlikely events (Fisk & Pidgeon, 1996; Wells, 1985). Details for a classroom demonstration are given in Text box 2.1.

### **Text box 2.1 Classroom demonstration**

#### **Background**

Aczel et al. (2016) presented participants with a conjunction paired with either the more likely or the less likely component. They found that the value assigned to the conjunctive probability was substantially larger in the former case compared with the latter. Thus, it appeared that the probability assigned to the conjunctive event depended on which single event it was paired with, an outcome which, to the best of my knowledge, has not been previously observed. The following classroom demonstration seeks to replicate and extend this finding.

#### **Method**

The context within which the conjunctive probability is presented constitutes the independent variable with three levels: with the less likely component, with the more likely component, and with an unrelated statement. The dependent variable

is the estimated value of the conjunction. Participants will be randomly assigned to three groups with each group receiving one of the following three versions of the Linda problem. All three versions start with Linda's description (see above).

Participants are then asked: How likely are each of the following statements (for each, enter a percentage probability between 0 and 100):

Version 1 continues with the two following statements:

Linda is a feminist.

Linda is a bank teller and a feminist.

Version 2 continues with the two following statements:

Linda is a bank teller.

Linda is a bank teller and a feminist.

Version 3 continues with the two following statements:

Linda is a socialist.

Linda is a bank teller and a feminist.

The resulting three sets of conjunctive estimates can be analysed using one-way between participant ANOVA. The prediction is that the conjunctive estimate paired with the more likely component (Version 1) will be significantly larger than that paired with the less likely component (Version 2). While no prediction is made in relation to the conjunctive estimate from the third version, as will be clear having read the remainder of this chapter, I expect that it will be close in magnitude to the outcome for Version 2. Paired comparisons can be made using Tukey's test. The incidence of the conjunction fallacy can be compared between Versions 1 and 2 using chi squared analysis. Fallacies are predicted to be rare in Version 1 and commonplace in Version 2.

Over the years, there has been doubt expressed as to whether the fallacy is a real phenomenon or the product of invalid experimental approaches. By way of contrast, the dominant view is that it represents a significant violation of normative reasoning. More recently, while accepting that the fallacy is a genuine phenomenon, its real-world significance has been questioned. I shall now examine each of these perspectives in turn.

### **Is the conjunction fallacy real?**

The first question that must be addressed is whether the conjunction fallacy is a genuine phenomenon at all. Several researchers have suggested that what appears to be a reasoning error can be viewed as perfectly rational when viewed from an alternative perspective. We shall examine these differing explanations in the remainder of this section.

***The laws of probability do not apply where there is subjective uncertainty***

Maguire et al. (2018) have drawn a distinction between subjective and objective uncertainty. Objective uncertainty, they argued, is characterized by pure randomness where each event is independent and unconnected to any subsequent event (e.g., where the outcome probabilities are known *a priori* such as with the toss of a coin or the throw of a die). With subjective uncertainty, Maguire et al. maintained that events are not usually independent and may convey additional information as to the nature of the generative mechanism. In their view, potential events confronting the individual as part of their everyday lives are characterized by subjective uncertainty and rather than being viewed as straightforward possibilities they may be perceived as subjectively informative conveying some deeper meaning. As such, in considering the broader context suggested by the event in question, the individual may amend their understanding of the underlying situation going beyond the basic information provided. In the context of the conjunction fallacy, given the existence of subjective uncertainty, since Linda's description (see above) is based on those attributes chosen by the experimenter, according to the Maguire et al. argument, the participant may view it as incomplete and partial. Equally, the statements to be considered may be viewed as potentially informative. In this context, given the individual's evolving state of knowledge, the resulting conjunction may seem less surprising and subjectively more likely than the bank teller component statement on its own. Crucially, Maguire et al. maintained that this was not a fallacy since objective probabilities cannot be defined for the events in question which reflect the *evolving state of knowledge* concerning Linda.

According to Maguire et al., it is possible to change a problem from one involving subjective uncertainty to one involving objective uncertainty in which the rules of probability would apply. They produced a modified version of the Linda problem in which the participant was told that a social media database was searched by inputting the following *randomly selected parameters*:

university degree = philosophy  
 marital status = single  
 $IQ > 130$   
 name = Linda  
 age = 31

The participant was informed that this returned a single database record and was asked which was more probable:

(1) That the record states:

occupation = bank teller and political outlook = feminist

Or (2) that the record states:

occupation = bank teller

Maguire et al. found that the incidence of the conjunction fallacy was significantly reduced from 64% in the original version of the problem to 39% in the social media

version. Nonetheless the fact that the incidence of the fallacy remained close to 40% is noteworthy.

### ***Conversational implicature***

Research questioning the veracity of the conjunction fallacy emerged shortly after the phenomenon was popularized by Tversky and Kahneman (1983). One strand of the research literature focused on the notion of conversational implicature – that is to say, where a communication has a meaning going beyond its literal interpretation. For example, when examining a painting we might say that we like the frame, with the implicit implication that we did not like the painting itself. Or that the child is full of energy, when we really mean that she will not sit still.

Given the laws of probability, when ranking the component and conjunctive statements, Linda's description is irrelevant since the conjunction cannot be more likely than either of its components. The individual might then wonder why the description was included. In including the description, did the experimenter intend to convey information relevant to the “true” meaning of the bank teller component statement? Given this possibility and when set alongside the conjunctive event, is the “Linda is a bank teller” statement interpreted as “Linda is a bank teller and not a feminist”? If this were to be the case then ranking this statement as less likely than the feminist–bank teller conjunction would not be a fallacy. However, while conversational implicature might account for a small proportion of conjunction rule violations, the consensus is that the concept does not offer a comprehensive explanation of the phenomenon (Donovan & Epstein, 1997; Morrier & Borgida, 1984; Pagin, 2019; Yates & Carlson, 1986).

### ***Frequentist interpretations***

Some, including many from the frequentist tradition in probabilistic inference, have argued that probabilities cannot be defined for single unique events such as whether, or not, Linda is likely to be a feminist. From this perspective probabilities are only defined for repeated events where a reference class can be defined, for example, how likely is it on a randomly selected day of the year that the temperature will drop below zero. Accordingly, it has been proposed that the fallacy will be greatly reduced when responses are elicited in terms of frequencies (Fiedler, 1988). For example, of a group of 100 people all resembling Linda's description: “How many out of 100 would be bank tellers?” versus “How many out of 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 sometimes greatly reduced in that most people indicate that there are a greater number of bank tellers than there are feminist bank tellers. However, presenting problems in terms of frequencies does not always reduce the incidence of the fallacy (e.g., Costello, & Watts, 2014; Evans et al., 2000; Fisk et al., 2019).

### ***Applying a different probabilistic rule***

Some researchers have claimed that the way in which many 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 or more 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  $P(\text{feminist and bank teller} \mid \text{Linda})$  and  $P(\text{bank teller} \mid \text{Linda})$  participants are actually evaluating  $P(\text{Linda} \mid \text{feminist and bank teller})$  and  $P(\text{Linda} \mid \text{bank teller})$ . Crucially, making judgments such that

$$P(\text{Linda} \mid \text{feminist and bank teller}) > P(\text{Linda} \mid \text{bank teller}) \quad (2.7)$$

does not necessarily violate any normative rule. While this suggestion is an appealing one, the results of one of my own studies (Fisk, 1996) cast doubt on this interpretation of the fallacy.

### ***Quantum probability***

The conjunction fallacy is only a genuine phenomenon if it is assumed that the laws of classical probability theory apply. Among the other more recent theories of probability judgment, one of the most complex and conceptually demanding is in terms of quantum probability, according to which the fallacy may not be a fallacy at all (e.g., Busemeyer et al., 2011). To illustrate the quantum framework, I 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.

In quantum probability, an event is represented by a pair of orthogonal base vectors of unit length, one representing the likelihood that the event will happen, the other that it will not. In the example set out in Figure 2.1, the likelihood of a yes or a no response to Bill playing jazz for a hobby is measured by the orthogonal base vectors  $J_y$  and  $J_n$  respectively. Similarly, the likelihood of a yes or a no response to Bill being an accountant is measured by the orthogonal base vectors  $A_y$  and  $A_n$ . In the context of the accountant and jazz questions, our knowledge concerning Bill (i.e., his description) is represented by a state vector labelled  $\Psi$ . This state vector is projected onto the two sets of axes. The squared length of these projections,  $\{J_y \mid \Psi\}$  and  $\{A_y \mid \Psi\}$ , determine the probability that Bill plays jazz for a hobby,  $P(J)$ , on the one hand and the probability that he is an accountant,  $P(A)$ , on the other hand, that is,  $P(J) = \{J_y \mid \Psi\}^2$  and  $P(A) = \{A_y \mid \Psi\}^2$ . However this is only true when each event is considered in isolation.

The quantum situation is rendered more complex in that the process of considering different possibilities, that is, that Bill is an accountant, changes the individual's original knowledge state. Should they accept this possibility as true then the state vector rotates from its original position  $\mid \Psi \rangle$  so as to coincide with the base vector  $A_y$  and the new state vector is now  $\mid \Psi_A \rangle$ . These order effects have been cited as a possible explanation for the

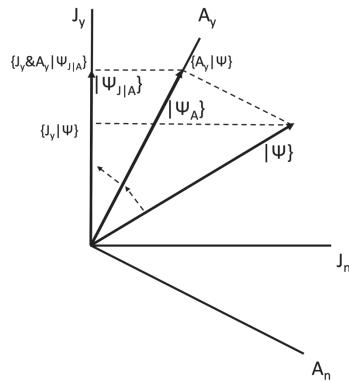


Figure 2.1 The conjunction fallacy in terms of quantum probability.

conjunction fallacy. When considering the jazz possibility the state vector projects onto  $J_y$  with a length of  $\{J_y|\Psi\}$  yielding a probability  $P(J) = \{J_y|\Psi\}^2$ . When considering the conjunction that Bill is an accountant and plays jazz, the state vector, having first rotated to the  $A_y$  base vector, then rotates to the  $J_y$  state vector producing a projection of length  $\{J_y \& A_y | \Psi_{J|A}\}$  and a conjunctive probability  $P(A \& J) = \{J_y \& A_y | \Psi_{J|A}\}^2$ . Since length  $\{J_y \& A_y | \Psi_{J|A}\} > \{J_y | \Psi\}$ , this implies  $P(A \& J) > P(J)$ , that is, a conjunction fallacy.

It is noteworthy that the base vectors representing the jazz question and the accountancy question are at oblique angles to each other. In the quantum sense this renders the two questions incompatible so that an answer to the jazz question cannot be determined until the accountancy question has been answered and vice versa. Compatibility and incompatibility are properties of vector spaces in quantum probability and the distinction has not really been defined in psychological terms.

While quantum probability has risen into prominence as an account of the conjunction fallacy and lay probability judgment in general, the axioms underpinning quantum probability generate several corollaries which, if it is to provide an account of the fallacy, should be confirmed in relation to the type of judgments which have characterized the conjunction fallacy literature. To test this proposition, Boyer-Kassem et al. (2016) administered versions of four tasks all of which have featured in the conjunction fallacy literature including the Linda and Bill vignettes.

Discussion of these axioms is beyond the scope of this chapter, but the results obtained by Boyer-Kassem et al. suggest that their participants' responses violated three of these key axioms: the grand reciprocity law, the QQ equality, and the presence of order effects in incompatible vector spaces. Thus, despite its prominence there are questions as to whether quantum probability adequately describes the kind of judgments that characterize the conjunction fallacy.

### **The conjunction fallacy is real and constitutes a potentially serious judgmental bias**

While many researchers acknowledge that the fallacy exists, there is a debate in the literature as to its significance. The dominant view is that the fallacy demonstrates that

individuals often produce erroneous judgments with potentially serious consequences for themselves. I shall now examine the alternative perspectives characterizing this viewpoint.

### ***Representativeness***

In their pioneering study, Tversky and Kahneman (1983) attempted to explain the fallacy in terms of the representativeness heuristic (see Chapter 12). Considering the Linda vignette above, her description is *representative* of the feminist concept but is unrepresentative of the characteristics of being a bank teller. However, the conjunctive concept also contains the attributes characteristic of feminism and so, over all, Linda's description 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 judgments are based on differences in representativeness, the conjunctive event is therefore judged to be more likely than the bank teller event. This will potentially result in the conjunctive probability being substantially overestimated.

### ***Signed summation***

“Signed summation” was a heuristic procedure proposed by Yates and Carlson (1986) whereby individuals represent degrees of likelihood in terms of a “qualitative likelihood index” (QLI). This yielded positive values for likely events and negative for unlikely. For example, subjectively, a likely event might be assigned a value of +27, an unlikely event - 12. The conjunctive QLI value is then the signed sum of these two values:

$$\text{Conjunctive QLI} = +27 - 12 = +15 \quad (2.8)$$

As is often the case with representativeness, signed summation can potentially result in greatly overestimated conjunctive probabilities. Where the unlikely event is only marginally unlikely, the negative QLI will be small in absolute terms and the conjunctive QLI will be similar in magnitude to that of the likely event.

### ***Inductive confirmation***

Tentori et al. (2013) introduced the concept of inductive confirmation to explain the fallacy. They noted that a hypothesis can either be strengthened or weakened by some additional evidence. In the former case, it is said to be inductively confirmed. It has already been noted that the fallacy is most commonly found where the conjunction pairs a likely with an unlikely event. Many accounts of the fallacy propose that the probability assigned to 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. Thus, an event will more often give rise to the fallacy when it is judged likely through being inductively confirmed by the evidence relative to when it is judged likely due to it having a high prior probability. By way of an example, one of their scenarios involved an individual, O. The conjunctions used contained the proposition that “O is an expert mountaineer” ( $h_1$ ) together with one of two component events, either that “O gives music lessons” ( $h_2$ ) or that “O owns an umbrella” ( $h_3$ ). The evidence, “e”, provided in the vignette, that

“O has a degree in violin performance” clearly provides inductive confirmation for  $h_2$ . In general terms, using their notation, given that  $h_1$  is true, the degree of inductive confirmation conferred by evidence “e” on hypothesis  $h_2$  is given by the expression  $\text{Conf}(h_2 | h_1)$ .

Using the “O gives music lessons” and other scenarios, Tentori et al. measured the degree to which the evidence confirmed the rival hypotheses. Crucially they found that  $\text{Conf}(h_2 | h_1) > 0$  and that  $\text{Conf}(h_3 | h_1) \leq 0$ . Since the degree of confirmation was positive in the former case and negative or zero in the latter, the implication of this was that the fallacy would be more prevalent with the  $h_1 \& h_2$  conjunction. Consistent with this, Tentori and co-workers’ results showed that, even though participants’ responses were such that  $P(h_2) < P(h_3)$ , nonetheless the vast majority of cases were such that  $P(h_1 \& h_2)$  was ranked higher than both  $P(h_1 \& h_3)$  and  $P(h_1)$ . Thus, most instances of the conjunction fallacy occurred with the conjunction containing the inductively confirmed event. There is, however, a flaw in the way in which Tentori et al. measured the relative incidence of the fallacy. In all their experiments, participants were asked to select which of  $(h_1)$ ,  $(h_1 \& h_2)$ , and  $(h_1 \& h_3)$  was most probable. The majority of participants selected  $(h_1 \& h_2)$  as most probable, thereby clearly committing the fallacy. However, in these cases we do not know how the participants in question viewed the relative likelihoods of  $(h_1)$  and  $(h_1 \& h_3)$  since they were not asked. It is possible that they might have viewed  $P(h_1 \& h_3) > P(h_1)$  thereby also committing the fallacy in this case. Thus the incidence of the fallacy in relation to  $P(h_1 \& h_3)$  is likely to have been underestimated by Tentori et al., perhaps to a substantial degree.

More recently Crupi et al. (2018) argued that the concept of confirmation could account for the presence of double fallacies where the conjunction is assigned a probability greater than the probabilities assigned to both component events. They demonstrated that in a scenario where the evidence supported both  $h_1$  and  $h_2$ , and where  $h_1$  and  $h_2$  provided support for each other, 62% of participants committed double fallacies. Rather differently, in a second scenario where both  $h_1$  and  $h_2$  were inconsistent with the evidence, nonetheless 49% of participants committed double fallacies. This occurred because the disparity between  $h_1$  and the evidence was reconciled when  $h_1$  was in conjunction with  $h_2$  and vice versa. Participants in both scenarios were physicians and the scenarios involved medical diagnosis, in one case the likelihood of symptoms given a diagnosis was assessed and in the other case the likelihood of different diagnoses given the symptoms.

### **Dual-process theories**

At least some individuals do appear to avoid the conjunction fallacy, some demonstrating a basic grasp of extensional reasoning, others an actual understanding of the conjunction rule (Yates & Carlson, 1986). Thus, a capacity for rational thought may sit alongside a reliance on more heuristic based processing. This possibility has parallels with the notion of dual-process theory. The theory argues for two types of reasoning and two forms of rationality (Evans & Over, 1996; see also Chapter 9 of this volume). The first of these, System 1, is based on a preconscious automatic, intuitive, rapid heuristic-based system, which allows decisions to be made quickly without any great cognitive effort utilizing criteria such as similarity or representativeness. The second, System 2, is a more analytical, effortful, rational, conscious system which, under appropriate conditions, is capable of generating normative solutions consistent with formal logic or probability theory. The

default is System 1 but where conflicts arise, System 2 processes may override the default position.

In the context of the conjunction fallacy, although the Linda problem is concerned with probabilities, Kahneman and Frederick (2002) argued that the problem elicits heuristic processing and that the judgment is made by substituting the representativeness attribute for the more challenging probability concept. In these cases, System 2 processes are ineffective in monitoring the output of System 1, resulting in fallacious judgments. Nonetheless, there is evidence that individuals may be aware of the conflict between the outputs of the two systems. In the context of the conjunction fallacy, a basic grasp of extensional reasoning would allow the individual to avoid the fallacy but the fact that the evidence is more representative of the conjunction than the less likely event creates conflict. This conflict has been said to decrease the individual's confidence in the judgment (De Neys et al., 2011). However, it has been argued that having to distinguish between probabilities that are in close proximity to each other (i.e., the less likely event and the conjunctive probabilities) is what potentially undermines confidence (Aczel et al., 2016).

Early research suggests that it is possible to prime and facilitate System 2 processes (e.g., Agnoli & Krantz, 1989; Fisk & Pidgeon, 1997). More recently, Bakhti (2018) utilized different priming interventions designed to elicit reflective thinking (System 2) on the one hand and more intuitive (System 1) based thinking on the other. The priming manipulation involved invoking a mind-set focused either on religious concepts or reflective analytical thinking. Previous research has revealed that religious beliefs are negatively associated with an analytical thinking style (Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014) and that belief in paranormal phenomena is associated with a greater propensity to commit the conjunction fallacy (Rogers et al., 2011). Bakhti (2018) found that religious priming was associated with a higher incidence of the fallacy relative to reflective priming.

Also utilizing the dual-process concept, Epstein's cognitive-experiential self-theory (CEST) has been used to investigate the fallacy. Specifically, Donovan and Epstein (1997) argued that conjunction problems may be classified along two distinct dimensions (natural-unnatural and concrete-abstract). They found that the fallacy was more common with concrete unnatural problems, such as the Linda vignette, which tend to elicit inappropriate (System 1) reasoning strategies.

Having features in common with the distinction between System 1 and System 2 processes, Donati et al. (2019) examined the conjunction fallacy phenomenon through the perspective of Bruner's notion of two modes of thought: narrative and paradigmatic. As narratives are readily shared with others and may evolve as they are shared, Donati et al. (2019) investigated whether the incidence of the fallacy might be affected when vignettes are discussed within small groups prior to an individual decision being made. The results revealed that the group condition was associated with significantly fewer fallacies, namely 53% relative to 88%. Some participants provided open-ended written explanations for their judgments and these were classified according to whether they exhibited narrative or paradigmatic thought processes. The results revealed that those demonstrating paradigmatic thought produced only 25% fallacious responses while those using the narrative style generated 87% fallacies.

Hertwig and Chase (1998) found that response format can elicit different reasoning strategies. When participants were asked to rank statements in order of probability, they appeared to utilize a cue-based strategy with rankings determined by the degree

of evidential support for each particular statement, perhaps consistent with System 1 processes. In relation to System 2 processes, when participants were asked to provide numerical estimates, they adopted rule-based strategies. There appeared to be at least some utilization of the conjunction rule among those with statistical training while statistically naïve individuals adopted a “ceiling rule” setting the conjunction exactly equal to the less likely probability.

Age-related changes in the prevalence of the conjunction fallacy may also reflect changes in the salience of System 1 and System 2 processes. Morsanyi et al. (2017) found that children exhibited similar fallacy rates to adults in a child-friendly version of the Linda scenario. However, when the vignette contained base rate information for the component and conjunctive events, children were more likely to avoid the fallacy relative to adults. Morsanyi et al. argued that adults are influenced by context, background information, and beliefs while children were more influenced by the explicitly presented information. Interestingly, adults who received a maths training intervention were also more likely to rely on the base rate information.

### ***Source memory***

Most investigations of the fallacy have been in the domain of reasoning performance. However, conjunction errors have been identified in the context of source memory judgments (Nakamura & Brainerd, 2017). Specifically, when choosing in which of two lists a target word had been presented, individuals sometimes indicated that the proportion of words that appeared both in List 1 *and* in List 2 was greater than the proportion of words believed to have appeared in one or both of the individual lists. In a separate experiment, individuals committed disjunction fallacies. In this case individuals were asked whether the words appeared either in List 1 *or* List 2 as well as being asked about their occurrence in each list individually. Here the proportion of words identified as having been in List 1, plus the proportion identified as having been in List 2, exceeded the proportion identified when individuals were asked whether a word had been in either List 1 or List 2. In fact, I have some doubt as to whether the probabilities were calculated correctly, but this is beyond the scope of the present chapter.

### ***The fallacy may have serious consequences for human decision-making***

The conjunction fallacy has been identified in the context of medical diagnosis where conjunctions of symptoms were judged more likely than individual symptoms (e.g., Tversky & Kahneman, 1983). As noted above, analogous results were observed by Crupi et al. (2018) who observed double fallacies when physicians made diagnostic judgments. In the area of social cognition, it is often the case that conjunctions of personal attributes are judged to be more likely than the individual attributes themselves. In a study examining the profile of homicide victims, the likelihood of combinations of victim attributes (e.g., gender, age, and race) were consistently overestimated relative to the actual joint probabilities. In comparing the likelihood of pairs of victim attributes occurring together (e.g., young and male) and three attributes occurring together (e.g., young, male, and black) the latter was rated more likely by 27% of the participants (Dearden, 2018).

Knowledge concerning the characteristic movement of everyday objects, for example, the factors governing changes in direction and momentum, are acquired through the process of interacting with the world around us utilizing those learned sensory and

perceptual processes which allow us to predict everyday outcomes. This intuitive grasp of physics might allow us to evaluate conjunctions of physical events without succumbing to fallacious conclusions. However, even here conjunctions involving the trajectory and resting place of objects, their density, and structural stability were judged more likely than their constituent event (Ludwin-Peery et al., 2020).

## **The conjunction fallacy is real but its real-world consequences may have been previously overstated**

### ***Configural weighted averaging***

Nilsson et al. (2009) have proposed a weighted averaging model in order to explain both conjunctive and disjunctive reasoning errors. They 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 judgment is conjunctive or disjunctive. Nilsson et al. maintained that the less likely event has the larger weight in determining the probability assigned to the conjunctive event while the more likely component has the larger weight in determining the probability assigned to the disjunctive event. Given that the weights for each component, although different, sum to one, in its simplest form, Nilsson and co-workers' model will always result in fallacies being committed.

To account for the fact that some conjunctive and disjunctive judgments were not fallacious, Nilsson et al. proposed that individuals' base estimates 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 several studies, Nilsson et al. obtained results consistent with the CWA model. I include the Nilsson et al. account in this section because the less likely event assumes more importance in determining the conjunctive probability. The reduced role for the more likely event (compared to accounts such as signed summation and representativeness) is more in line with probability theory in that changes in the less likely event have a greater impact on the conjunctive probability relative to equivalent changes in the more likely event.

### ***Random noise or sampling errors***

Costello and Watts (2014, 2017) adopted a radically different perspective in which the fallacy is viewed as a by-product of an error-prone sampling process in an otherwise generally normative and rational process. In their view, when an event happens, its occurrence is stored in episodic memory and, when considering the likelihood of that particular event, these memories are searched and the number of times the event occurred is retrieved. However, this process is subject to error in that on a small number of occasions an event is thought to have occurred when it did not and similarly, on occasions, the occurrence of an event is missed. A similar situation arises when the individual attempts to retrieve the frequency of joint events, conjunctive and disjunctive, however, in this case the search through memory is subject to a greater degree of error. Given that the search through memory is error prone, there will be outcomes where the probability of the less likely

component event is underestimated, while that of the conjunctive event is overestimated. Since the error rate for joint events is greater than that for component events, this tendency will be further accentuated. In these cases, where the errors are sufficiently large, the conjunction fallacy will occur.

On the positive side, Costello and Watts (2014) maintained that these errors appear to cancel each other out. For example, the addition rule in probability defines a fundamentally important equality:

$$P(A) + P(B) - P(A \& B) - P(A \text{ or } B) = 0 \quad (2.9)$$

The rule stipulates that for any two events, A and B, the sum of the component probabilities minus the two corresponding joint probabilities (conjunctive and disjunctive) is equal to zero. Costello and Watts (2014) demonstrated that although the individual's estimates of the single event and the joint probabilities were prone to error, these errors cancelled each other out and the addition rule was satisfied. In fact, Costello and Watts (2017) acknowledged that the greater propensity for error when retrieving joint probabilities sometimes resulted in the addition rule being violated. Given the assumptions underpinning their model, they demonstrated that when the component probabilities summed to more (less) than one, the product of the above equality (Equation 2.9) was positive (negative). However, they claimed that the margin of error was constrained and small in magnitude being no larger than  $\pm 2\Delta d$ , where  $\Delta d$  was a measure of the increase in the error rate for sampling joint events relative to the error rate for single events. Costello and Watts (2017) provided a reanalysis of their (2014) data demonstrating that the product of the above equality was broadly consistent with this expectation, with the outcome being positive for  $P(A) + P(B) - 1 > 0$ , and negative for  $P(A) + P(B) - 1 < 0$ . By way of contrast, in Study 2 of my paper (Fisk et al., 2019), I collected estimates of the probabilities of the same weather events as those investigated by Costello and Watts. However, a reanalysis of my data yielded results that contradicted those obtained by Costello and Watts in that while the addition rule produced an overall average significantly greater than zero, the values of  $P(A) + P(B) - 1$  were significantly less than zero. Crupi and Tentori (2016) also questioned the adequacy of the Costello and Watts (2014) model pointing out that the results of the Tentori et al. (2013) study were inconsistent with Costello and Watts's account.

The reason that Costello and Watts's work is included in this section is that, while the fallacy may sometimes occur, the implication is that since the addition rule is satisfied individual judgments are essentially rational. That said, it can be demonstrated that the model proposed by Nilsson et al. also generates outcomes that are consistent with the addition rule and that my own model (Fisk et al., 2019, and see below) can sometimes produce outcomes broadly consistent with the rule.

### ***Joint probabilities are quasi-random adjustments from fixed reference points***

During the 1990s and early 2000s, I 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 degree of surprise experienced when contemplating prospective events. 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 the other of the components, the other one having no obvious influence. For conjunctive events, 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).

My thinking has now coalesced into a model of joint probability judgment, conjunctive and disjunctive (see Fisk et al., 2019). The assumptions underpinning my model involve a two-stage process. First, based on the principles of potential surprise, the individual selects a reference point for determining the joint probability. Usually, but not invariably, the reference point is the less (more) likely component in the case of the conjunctive (disjunctive) probability although in rare cases the other component may act as the reference point.

In the second stage of the judgment the individual makes a quasi-random adjustment from the reference point, either positive or negative. The magnitude of the adjustment is apparently random in so far as it appears to be unrelated to any of the other variables of interest that we have tested. Thus, in the conjunctive case where the reference point is the less likely component probability, the magnitude of the adjustment appears to be unrelated to the magnitude of the more likely event probability, unrelated to the strength of the conditional probability  $P(\text{more likely}|\text{less likely})$ , and unrelated to the degree of inductive confirmation. The adjustment is quasi-random in the sense that it is constrained to reside within the interval between the reference point and the ends of the probability continuum (0, 1).

The direction of the adjustment does not appear to be random in that those who make a negative (positive) adjustment in the conjunctive (disjunctive) case, thereby avoiding the fallacy, tend to do so with a degree of consistency across different judgment problems. We have argued that the second stage of the judgment process may be influenced in an all or nothing manner by the application of some System 1 or System 2 process. For example, those avoiding the fallacy may have some basic grasp of extensional reasoning without having knowledge of, or being able to apply, the multiplicative conjunction rule. Thus, they may know that the conjunction must be less probable than the less likely component without being able to determine the actual magnitude of the conjunctive probability itself. Hence the negative adjustment is essentially all or nothing and unrelated to the magnitude of the more likely event.

In the model proposed by my co-workers and myself, the addition rule (Equation 2.9) is generally violated and our results across two experiments were broadly consistent with this. While the addition rule can also be violated in the model proposed by Costello and Watts (2017), as we pointed out above, in our case the inequality was in the opposite direction to that predicted by them. There are circumstances where our model does predict outcomes broadly consistent with the addition rule. This will occur where the fallacy rate and the absolute magnitude of the quasi-random adjustments are similar in the conjunctive and disjunctive cases.

As intimated above, in the conjunctive case, given the level of uncertainty that characterizes most judgments and the ever evolving state of knowledge, basing the conjunctive probability on the less likely component may constitute a reasonable approximation and may be adequate for most everyday decisions.

## Conclusion

The fallacy continues to generate research interest. There remains no real agreement as to whether it is a genuine phenomenon and alternative explanations appear to come into fashion and then retreat into the background as do theoretical accounts of the fallacy itself. It is to be hoped that new theoretical accounts will be better grounded in empirical data and generate continuing research interest and critical appraisal. Perhaps it would be of value to focus on what the data have revealed about the fallacy over the last 40 years and attempt to integrate the empirical results into some sort of unified theory.

## Summary

- Since the fallacy was first popularized, some have questioned whether, or not, it constitutes a genuine reasoning fallacy.
- Initially the notion of conversational implicature was invoked to account for what appeared to be, but it was argued was not, non-normative reasoning.
- More recently, it has been questioned whether subjective probability judgment should follow the norms of classical probability.
- Many continue to maintain that the fallacy *does* constitute a potentially serious reasoning bias.
- It has been argued that processes such as representativeness, signed summation, and inductive confirmation can serve to substantially distort conjunctive probability judgments.
- This is often seen through the lens of dual-process theory where errors may be attributed to the effects of System 1 processes.
- More recently, while acknowledging the existence of the conjunction fallacy, some have questioned whether it has important consequences in the course of everyday judgment.

## Further reading

Of the three broad perspectives that I have outlined, Maguire et al. (2018) addresses the issue of whether the fallacy is a meaningful construct. Tentori et al. (2013) reveal that the manner in which the likely event derives its probability (conditioned on evidence or due to a high prior probability) is a determinant of the fallacy. My own research (Fisk et al., 2019) attempts to provide a critical analysis of much of the existing literature as well as provide new insights.

## References

- Aczel, B., Szollosi, A., & Bago, B. (2016). Lax monitoring versus logical intuition: The determinants of confidence in conjunction fallacy. *Thinking & Reasoning*, 22, 99–117.
- Agnoli, F., & Krantz, D. H. (1989). Suppressing natural heuristics by formal instruction: The case of the conjunction fallacy. *Cognitive Psychology*, 21, 515–550.

- Bakhti, R. (2018). Religious versus reflective priming and susceptibility to the conjunction fallacy. *Applied Cognitive Psychology*, 32, 186–191.
- Boyer-Kassem, T., Duchene, S., & Guerci, E. (2016). Quantum-like models cannot account for the conjunction fallacy. *Theory and Decision*, 81, 479–510.
- 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.
- 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.
- Costello, F., & Watts, P. (2017). Explaining high conjunction fallacy rates: The probability theory plus noise account. *Journal of Behavioral Decision Making*, 30, 304–321.
- Crupi, V., Elia, F., Apra, F., & Tentori, K. (2018). Double conjunction fallacies in physicians' probability judgement. *Medical Decision Making*, 38, 756–760.
- Crupi, V., & Tentori, K. (2016). Noisy probability judgement, the conjunction fallacy, and rationality: Comment on Costello & Watts (2014). *Psychological Review*, 123, 97–102.
- Dearden, T. E. (2018). The conjunction fallacy in profiles of victims of homicide. *Journal of Investigative Psychology and Offender Profiling*, 15, 187–199.
- De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE*, e15954.
- Donati, C., Guazzini, A., Gronchi, G., & Smorti, A. (2019). About Linda again: How narratives and group reasoning can influence conjunction fallacy. *Future Internet*, 11, 210.
- 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.
- Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, UK: Psychology Press.
- 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., Marshall, D. A., Rogers, P., & Stock, R. (2019). An account of subjective probability judgement for joint events: Conjunctive and disjunctive. *Scandinavian Journal of Psychology*, 60, 405–420.
- 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.
- 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.
- 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.
- Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2020). Broken physics: A conjunction fallacy effect in intuitive physical reasoning. *Psychological Science*, 31, 1602–1611.
- Maguire, P., Moser, P., Maguire, R., & Keane, M. T. (2018). Why the conjunction effect is rarely a fallacy: How learning influences uncertainty and the conjunction rule. *Frontiers in Psychology*, 9, 1011.

- Morrier, D. M., & Borgida, E. (1984). The conjunction fallacy: A task specific phenomenon? *Personality and Social Psychology Bulletin, 10*, 243–252.
- Morsanyi, K., Chiesi, F., Primi, C., & Szűcs, D. (2017). The illusion of replacement in research into the development of thinking biases: The case of the conjunction fallacy. *Journal of Cognitive Psychology, 29*, 240–257.
- Nakamura, K., & Brainerd, C. J. (2017). Disjunction and conjunction fallacies in episodic memory. *Memory, 25*, 1009–1025.
- 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.
- Pagin, A. (2019). Exploring the conjunction fallacy in probability judgment: Conversational implicature or nested sets? *Journal of European Psychology Students, 10*, 12–25.
- 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.
- 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.
- 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.
- 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.
- 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.

### 3 Base-rate neglect

Gordon Pennycook, Christie Newton, 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 an 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 et al., 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 3.1. The purpose of the first 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 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 et al., 2012). Thus, at least about 80% of the students in the study completely neglected the base-rates, but probably more did as well.

#### Text box 3.1

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 1,000 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 40 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)

Consider now the second example, a base-rate problem modeled on Kahneman and Tversky's (1973) "lawyer-engineer problem" (described subsequently). This problem is very similar to the first example, 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. Similarly, Bago and De Neys (2017) found that across four experiments that used a modified version (IT engineers and professional boxers; Pennycook et al., 2014b) with extreme base-rate ratios (997/3, 996/4, 995/5), average accuracy was 40%. 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 (Bago et al., 2018; De Neys & Glumicic, 2008; Pennycook et al., 2012).

The third example in Text box 3.1 – 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.

### Text box 3.2

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 ten 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 3.1). 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.

### **Results and discussion**

The experiment described in Text box 3.2 was the first in Kahneman and Tversky's (1973) seminal work on base-rate neglect. Their results can be found in Table 3.1. 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 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.

*Table 3.1* Estimated base-rates of the nine areas of graduate specialization and summary of similarity and likelihood ratings for Tom W.

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

Source: Kahneman and Tversky (1973).

## 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 3.1, 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 centered 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 3.2) from Kahneman and Tversky (1973) is a particularly striking example of this. In the case of the more complex mammography problem (Text box 3.1) (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), when given problem with extreme versus moderate base-rates (Newman et al., 2017), when base-rates are made central (Wu & Emery, 2021), or if given problems where the base-rates come *after* individuating information (e.g., stereotypes; Krosnick et al., 1990) or if no individuating information is present (Gualtieri & Denison, 2018). 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 3.3. 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% 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 needs to be integrated  $[(0.8/0.2) \star (0.15/0.85)]$ . 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 3.3, 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).

**Text box 3.3**

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% 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 four cases out of five, a faulty motor creates a weaker field, but in one case out of five, 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?

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. Imagine that you were presented a new representative sample of women at age 40 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  $\sim 3$  (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 Tverksy'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). A modified version of this problem was given to children as young as 3 years old which revealed that base-rate weighting occurs early in development (Gualtieri & Denison, 2018). At 3 years old, children use base-rate information more than individuating information to make decisions. Children aged 4–5 years old begin incorporating individuating information and by 6 years old (when they start learning stereotypes), their susceptibility to base-rate neglect matches adults'. 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.

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 3.4): 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 3.4

Conditions from Pennycook and Thompson (2012)

- (1) In a study 1,000 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 1,000 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 1,000 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 3.4 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 3.1 and 3.2., 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 3.1, 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 50–90% 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.

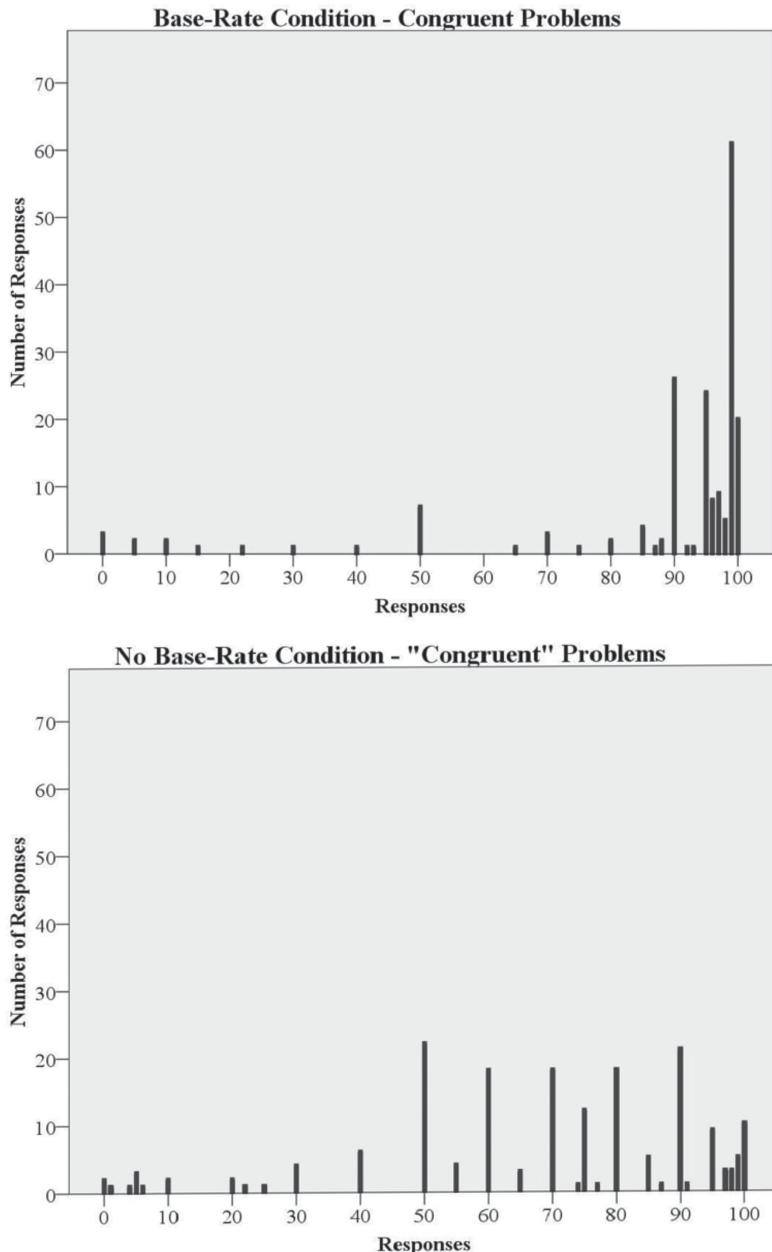


Figure 3.1 Distribution of probability estimates for congruent problems (from “Reasoning with base rates is routine, relatively effortless, and context dependent” by Gordon Pennycook and Valerie A. Thompson, 2012, *Psychonomic Bulletin & Review*, 19, 531, © Psychonomic Society, Inc. 2012. Adapted with permission of the publisher). For BR condition, high responses are consistent with both stereotypes and base-rates.

Note: For NoBR condition, high responses are consistent with stereotypes. Note that problems were not “congruent” in the NoBR 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 counter-balancing was such that, in the no base-rate condition, the problems would have been “congruent” or “incongruent” if base-rates had been included.

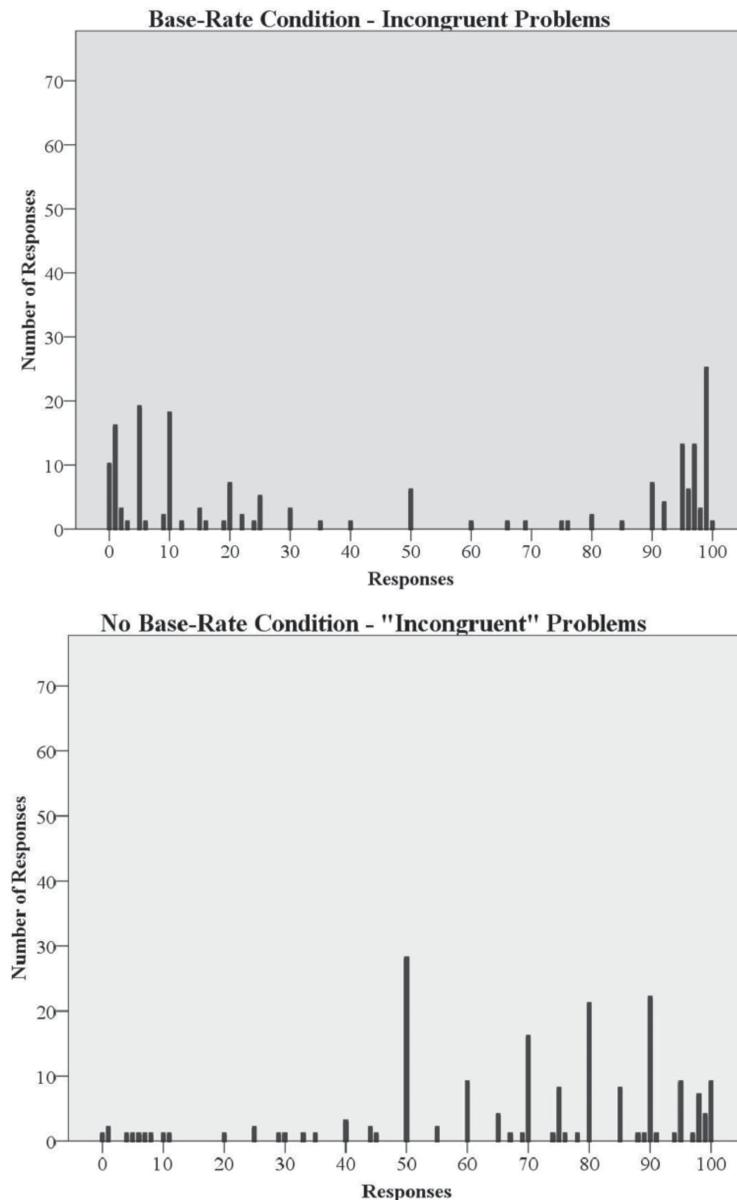


Figure 3.2 Distribution of probability estimates for incongruent problems (from “Reasoning with base rates is routine, relatively effortless, and context dependent” by Gordon Pennycook and Valerie A. Thompson, 2012, *Psychonomic Bulletin & Review*, 19, 532, © Psychonomic Society, Inc. 2012. Adapted with permission of the publisher).

*Note:* 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.

An entirely different pattern was evident when base-rates and stereotypes were inconsistent (incongruent). In the BR condition (Figure 3.2), 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). This contrasts starkly with the responses for the NoBR 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 moderated their judgments by using the base-rates, most of the probability judgments would be clustered somewhere in the middle of the distribution. As is evident from Figure 3.2, 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 12 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. This explanation is consistent 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 et al., 2014a).

Although most researchers apparently agree that individuating information like stereotypes are very intuitive sources 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 3.4). 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 et al. (2014b). 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 participants' ability to respond based on their 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. Similarly, Newman et al. (2017) demonstrated that belief-based reasoning can be slow and rule-based (Type 2) reasoning can be fast. This finding challenges the common speed-asymmetry explanation that fast responses are belief-based and precede slow rule-based responses (Evans, 2009; Newman & Thompson, 2017; Thompson & Johnson, 2014). A more recent study also suggests that fast base-rate responses can be given with high confidence, despite conflicting heuristic responses (Bago & De Neys, 2017). There is neuropsychological evidence to support the claim that base-rates can be processed fast circumventing slow, reflective Type 2 processing (Bago et al., 2018; Banks & Hope, 2014). 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 et al., 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, 2018; 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 et al., 2014a) 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 et al., 2014a). 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 et al., 2014a). 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.

## Conclusion

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 et al., 2014b). 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 et al., 2014a). 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 et al. (2014b) 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 and the Social Sciences and Humanities Research Council (SSHRC) of Canada.

## References

- Bago, B., & De Neys, W. (2017). Fast logic? Examining the time course assumption of dual process theory. *Cognition*, 158, 90–109.
- Bago, B., Frey, D., Vidal, J., Houdé, O., Borst, G., & De Neys, W. (2018). Fast and slow thinking: Electrophysiological evidence for early conflict sensitivity. *Neuropsychologia*, 117, 483–490.
- Banks, A. P., & Hope, C. (2014). Heuristic and analytic processes in reasoning: An event-related potential study of belief bias. *Psychophysiology*, 51, 290–297.
- 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: Psychology Press.
- De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. *Cognition*, 113, 45–61.
- De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, 106, 1248–1299.
- 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. 249–267). New York: Cambridge University Press.
- Evans, J. St. B. T. (2009). How many dual-process theories do we need? One, two, or many? In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 33–54). Oxford: Oxford 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., & Nisbett, 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.
- Gualtieri, S., & Denison, S. (2018). The development of the representativeness heuristic in young children. *Journal of Experimental Child Psychology*, 174, 60–76.
- 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: Psychology Press.
- Newman, I. R., Gibb, M., & Thompson, V. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43, 1154–1170.
- Newman, I. R., & Thompson, V. A. (2017). Logical intuitions and other conundra for dual process theories. In W. De Neys (Ed.), *Dual process theory 2.0* (pp. 121–136). Abingdon, UK: Routledge.
- Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014a). 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. (2014b). Base-rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40, 544–554.
- Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking and Reasoning*, 24, 423–444.
- 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.), *Psychology of learning and motivation* (Vol. 60, pp. 35–75). Burlington, VT: 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.
- Wu, S., & Emery, C. (2021). American base-rate neglect: It is not the math, but the context. *Journal of Behavioral Decision Making*, 34, 116–130.

## 4 Framing

*Anton Kühberger*

Is telling half a lie morally worse than telling half the truth? Does a glass described as half full contain the same amount of liquid than 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 communicating similar things differently. The difference, however, lies in the words used rather than in the underlying reality. 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”. Framing research shows that seemingly innocent differences in wording can have important consequences.

That some state of affairs can be described in 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 peoples’ responses even in equivalent situations. The framing effect is a reliable empirical finding, reported in hundreds of papers. Framing research gained its momentum from the interpretation that it violates a basic normative principle, frequently 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 consequentialism: Variations of form should not affect judgment and choice as long as they do not change actual outcomes. Obvious as this principle may be at first sight, it may be violated in many ways, all instances of framing. Here are some examples of framing that may or may not influence judgment and choice. If product A is more expensive than product B, then B is cheaper than A: Does the difference in frame influence your preference? If one succeeds in solving half of the items of a test or fails to solve half of the items, is this person evaluated equally? If your diary is half full, is it half empty at the same time? Does it matter whether a sample is described as consisting of a minority of females or of a majority of males? Would you rather buy a product priced €1.99 or ¢199? Is an incidence rate of 0.1 in 100 worse than 1 in 1,000? And why is a p-value of 0.055 described as “close to significance” rather than as “marginally insignificant”?

### Varieties of framing effects

Framing effects come in many varieties:

- 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 emails can be emphasized as a matter of public safety, or as a matter of individual privacy. For another example see Text box 4.1. In contrast to equivalency framing, emphasis framing does not involve logically equivalent ways of making the same statement. Equivalency framing actually 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 from around 380 papers published on framing between 1997 and 2007 about 60% were on emphasis framing, while only about 20% were on equivalency framing using the experimental method (Borah, 2011).

### ***Attribute framing, goal framing, and risky-choice framing***

Levin et al. (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.

#### **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. 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; cf. Chapter 24 on labeling). This leads to a valence-consistent shift in the evaluation of the entire issue.

#### **Goal framing**

In goal framing, messages are framed to 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 introduced by Meyerowitz and Chaiken (1987; see Text box 4.1). Notice the difference to 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. Indeed, a further distinction here is between portraying the action (e.g., doing X v. avoiding X) and the kernel state itself (e.g., desirable v. undesirable). This amounts to four possible frames, namely (i) doing something desirable, (ii) avoiding something undesirable, (iii) doing something undesirable, (iv) avoiding something desirable. Research on the effects of these combinations is lacking.

Explanations of goal-framing effects resolve around two issues: emphasizing the construal of a behavior as risky or uncertain, as opposed to safe and certain (Rothman & Salovey, 1997); and emphasizing people's dispositional sensitivity to favorable or unfavorable outcomes (Mann et al., 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 et al. (2008) have proposed a somewhat different account for goal-framing effects: They proposed that the self-regulatory orientations introduced by Higgins (1997) – prevention focus (being concerned with safety, security, the fulfilment of obligations, and the absence of unfavorable outcomes), and promotion focus (being concerned with accomplishments and ideals, and the presence of favorable 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 favorable outcomes especially salient, and persuasive. In contrast, illness-detection behaviors (e.g., mammography) induce a prevention-focused mindset, rendering the presence or absence of unfavorable outcomes especially salient, and persuasive.

### Risky-choice framing

The prototypical example of risky-choice framing is 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  $v(+200) > \frac{1}{3}v(+600)$ , and people prefer the sure gain. In the negative frame, the (often assumed rather than measured) 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 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 cognitive explanation of risky-choice framing effects. However, there is a variety of related explanations that take the weighting of probabilities also 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.

### 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?*

## 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; Steiger & Kühberger, 2018), and meta-analyses specifically on message-framing effects (Gallagher & Updegraff, 2012; Nabi et al., 2020; O’Keefe & Jensen, 2008, 2009), and on risky-choice framing effects (Gong et al., 2013; Kühberger et al., 1999).

Before discussing the framing effect size (see also Text box 4.2) 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 one option over another, in addition to expectation. An “unframed” condition, or a condition framed in both ways (termed *competitive* in research by Chong & Druckman, 2007b) might provide a plausible benchmark for comparison, however. For example, Druckman (2011), 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 showed that exposure to competitive frames makes framing effects disappear. However, research on competitive framing is rare, and may not be feasible in many cases.

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. 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, defined relative to two different framing conditions, showing increased risk seeking in the negative framing condition in relation 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 gain), as well as risk seeking for losses (i.e., more than 50% choices of the risky loss). Specifically, in our example the bidirectional test (discrepancy from the 50:50 expectation in either condition) results in no framing effect for gains (10 out 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 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). The literature usually reports the unidirectional test.

Some meta-analyses report effect sizes for framing effects. The first comprehensive meta-analysis (Kühberger, 1998) summarized 136 studies with 230 single effect 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  $d = .31$ . This analysis was repeated and extended by Steiger and Kühberger (2018), taking possible publication bias into account. They found a corrected overall effect size of  $d = 0.52$ , considerably higher than the size reported originally. No evidence was found of intense p-hacking. Reassuringly, this corrected effect size estimate is very similar to the effect size reported in the Many Labs Replication Project (Klein et al., 2014) on gain-loss framing ( $d = 0.60$ ). 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. Finally, Kühberger and Steiger (2018) conducted a meta-analysis of studies published in 2016 and again found a similar effect size ( $d = 0.56$ ).

For the subgroup of risky-choice framing tasks, a meta-analysis on more than 40 such experiments reported the bidirectional framing effect (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 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 – indeed non-significant – effects of framing on the persuasiveness of health messages when the persuasiveness was assessed as either attitudes, or 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. A recent meta-analysis (Nabi et al., 2020) similarly found weak effects of message framing on attitudes ( $d = 0.02$ , across  $k = 13$  studies), behavioral intent ( $d = 0.08, k = 20$ ), and behavior ( $d = 0.18, k = 2$ ).

Taken together, there is good evidence for a framing effect of medium size in attribute and risky-choice framing tasks of about  $d = .50$  ( $r = .24$ ). To illustrate its practical importance, a binomial effect-size display (Rosenthal & Rubin, 1982) may be helpful. By

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 62% ( $.50 + r/2$ ), and negative framing would decrease the percentage of participants choosing risk-aversive to 38% ( $.50 - r/2$ ). In contrast to attribute framing and risky-choice framing, the evidence for goal-framing effects is still 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 the EU is preparing for the outbreak of a new variant of COVID19, which is expected to kill 6,000 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, 2,000 people will be saved.*

*If Program B is adopted, there is 1/3 probability that 6,000 people will be saved, and 2/3 probability that no people will be saved.*

*Which program would you favor?*

- Program A     Program B

Here is the task framed negatively:

*Imagine that the EU is preparing for the outbreak of a new variant of COVID19, which is expected to kill 6,000 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, 4,000 people will die.*

*If Program D is adopted, there is 1/3 probability that nobody will die, and 2/3 probability that 6,000 people will die.*

*Which program would you favor?*

- Program A     Program B

#### **Design**

Experiments vary widely in sample size, and therefore in power. Assuming an effect size of  $d = .50$ , a significance level of  $\alpha = .05$  in a one-sided test, and a power of .80, a power analysis yields a required sample size of 51 participants per group. That is, about 100 participants should be recruited.

### ***Analysis***

Participants are required to choose one of the two programs. The results can be reported in a 2 (framing: positive v. negative)  $\times$  2 (choice: sure v. risky option) frequency table. Alternatively, they can be depicted by a diagram showing the percentage of risk-seeking choices.

Two basic analyses are possible, unidirectional and bidirectional. Our power calculation pertains to the unidirectional test, comparing the two framing conditions directly in a  $2 \times 2$  table. Do a  $\chi^2$  ( $df = 1$ ) -test. The bidirectional test is whether the choice proportions differ from 50% in either condition. Again, run  $\chi^2$  ( $df = 1$ ) -tests.

## **Results**

Chances are good to find a unidirectional effect, showing a difference between framing conditions. Chances to find both bidirectional effects are less good.

### **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. 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 small (recall that expected values are identical). Empirically, allowing for indifference reduces the effect.

#### ***Design: between versus 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 counter-intuitive 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 that is stronger than the desire to appear consistent in one's own choices.

#### ***Description of options***

The sure option only states that 2,000 people will be saved; it says nothing about the remaining 4,000 people. Adding the implicit complement (i.e., that 4,000 people will not be saved) spoils the effect. The same applies to the negative framing condition: Adding "*2,000 people will not die*" to "*4,000 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 number 600 in conjunction with the probability levels of 1/3 and of 2/3 is used frequently in many risky-choice framing experiments. Here we recommended using 6,000 in order to increase plausibility of the scenario. There is no deeper theoretical reason for using any number. Be careful: The use of small numbers (e.g., six people in danger, two 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 et al., 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. Rather, in the gambling domain, numbers are treated as exact. Nevertheless, framing effects with monetary outcomes are robust and similar in size to hypothetical 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 picture 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 cognitive characteristics, others in affective and motivational processes, still others in the pragmatics of language and communication.

### ***Cognitive accounts***

(Cumulative) prospect theory is the most influential cognitive model of framing. Although it succeeds in explaining a variety of framing results, there is increasing doubt about 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 & Grädl, 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., withholding the info that *with 2/3 probability nobody will be saved*) make the framing effect disappear. Taken together, the (cumulative) prospect theory idea that framing effects follow from the specifics of outcome valuation (and probability weighting) can account for the basic finding in the classic formulation of the task, but fails to account for variability in the findings due to small changes in task and procedure. Framing effects have more than just this single source.

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 gist of the risky option being *some people will be saved or no one will be saved*. The choice thus boils down on the simple contrast between *some people will be saved* (sure gain) and *some people will be saved or no one will be saved* (risky gain). Because saving some people is clearly better than saving some or 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 to risky-choice framing tasks, it is however not clear how it could be applied to attribute or to goal framing.

### **Motivational and emotional 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, the 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 et al., 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. Since, 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).

Recent accounts focus on the emotional effects of framing. Recall that there are two major sources how emotions can influence decisions (Loewenstein et al., 2001): by *expected* emotions (expected on realizing the outcome of the decision), and by *immediate* emotions (felt while contemplating the decision – often unrelated to the decision itself, i.e., incidental emotions). The role of incidental emotions on framing effects is still unresolved, because of equivocal findings. However, with respect to immediate emotions there is a stunningly simple – yet plausible – assumption: Compared with an equivalent risky option, sure gains are emotionally attractive, while sure losses are emotionally aversive. There is empirical evidence for this (e.g., Gosling & Moutier, 2017, 2019; Nabi et al., 2020). In addition, neurophysiological evidence (DeMartino et al., 2006) indicates that framing susceptibility is related to amygdala activation. Much contemporary research follows this line of inquiry.

### **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 others 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 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, 2015; Teigen & Nikolaisen, 2009; van Buiten & Keren, 2009).

In essence, the story about the source of framing effects revolves around the notions of selection and salience of information: (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 psychophysical processes transforming the input; and (v) framing can lead to different motivational and emotional processes (see also 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**

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 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. If communicators select descriptions on reason, Level 1 information equivalence does not apply. Equivalent descriptions of proportions like in our meat example are Level 2 information equivalent,

however, as they convey logically equivalent attribute descriptions. These descriptions are logically equivalent, but they can leak information about what a communicator expects, or prefers. At Level 3 lies the risky-choice framing task: The options 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 as relative (e.g., by assuming reference points) suffices to introducing inequality. At Level 4, we are leaving the traditional domain of framing and are turning 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 diverse 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 et al., 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, which highlight different aspects of topics. 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.
- 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 weak effects at best.
- Framing effects are explained by a variety of models: cognitive, motivational, emotional, 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 communication. 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. Teigen (2015) thoroughly discusses the framing of numerical qualities.

## 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.
- DeMartino, B., Kumaran, D., Seymour, B., & Dolan, R. (2006). Frames, biases, and rational decision making in the human brain. *Science*, 313, 684–687.
- Druckman, D. (2011). What's it all about? Framing in political science. In G. Keren (Ed.). *Perspectives on Framing* (pp. 279–301). New York: Taylor & Francis.
- 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.
- Gong, J., Zhang, Y., Yang, Z., Huang, Y., Feng, J., & Zhang, W. (2013). The framing effect in medical decision-making: A review of the literature. *Psychology, Health and Medicine*, 18(6), 645–653.
- Gosling, C. J., & Moutier, S. (2017). High but not low probability of gain elicits a positive feeling leading to the framing effect. *Frontiers in Psychology*, 8, 81.
- Gosling, C. J., & Moutier, S. (2019). Is the framing effect a framing effect? *Quarterly Journal of Experimental Psychology*, 72(6), 1412–1421.
- 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. (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.
- 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.
- Klein, R. A., Rathliff, K. A., Vianello, M., Adams, R. B. Jr., Bahník, Š., Bernstein, M. J. . . . Cemalcilar, Z. (2014). Investigating variation in replicability. *Social Psychology*, 45, 142–152.
- 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.
- Kuhberger, 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.
- Kuhberger, 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.
- Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, 127, 267–286.
- 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.
- Nabi, R. L., Walter, N., Oshidary, N., Endacott, C. G., Love-Nichols, J., Lew, Z. J., & Aune, A. (2020). Can emotions capture the elusive gain-loss framing effect? A meta-analysis. *Communication Research*, 47(8), 1107–1130.
- 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. *Psychothema*, 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). Mahwah, 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., & Kühlberger, 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.
- Steiger, A., & Kühlberger, A. (2018). A meta-analytic re-appraisal of the framing effect. *Zeitschrift für Psychologie*, 226(1), 45–55.
- Teigen, K. (2015). Framing of numerical quantities. In G. Keren & G. Wu (Eds.), *The Wiley handbook of judgment and decision making* (pp. 568–589). Hoboken, NJ: Wiley-Blackwell.
- 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(5), 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 in 1620:

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)

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.”<sup>1</sup>

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 off, 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 in some details over 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 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 number 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 got 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 number 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 participant’s strategy is called, there seems to be a problem with it, since 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 rule-discovery 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 extroverted, 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.

### **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 in Figure 5.1: “If there is a vowel on one side, then there is an even number on the other side.”

Give them five 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 five minutes).



*Figure 5.1* The four cards.

### **Group phase**

Form groups of students, ideally at random and of four or five 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 et al. (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 tend 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 et al. (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).

### ***Wason selection task***

The Wason selection task – or four-card 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 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 confirmation 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 et al., 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 et al., 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, that 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 has 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 rather than 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 had been 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 *mymode 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 et al., 2012). After they handed back the questionnaire, the experimenter performed a sleight of hand which 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 than 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, less 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 explanations suggest 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 been 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 et al., 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 et al., 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-card 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.

## Note

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

## Further reading

Nickerson (1998) provides a good overview with many intriguing examples from outside the lab. Oswald and Grosjean (2004) provide an earlier review which 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 Lilienfield et al. (2009).

## 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? *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). Harmondsworth: Penguin.

# **6 Illusory correlation**

*Klaus Fiedler, Karolin Salmen, and Florian Ermak*

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 diagnoses. Yet, 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 equivocal. Although many findings testify to humans' and animals' high sensitivity to differential event frequencies (Alloy & Abramson, 1979; Malmi, 1986; Zacks et al., 1982), another sizeable 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. Let this term not only apply to overestimations of zero correlations but to all kinds of systematic biases in subjective correlation assessment.

Some "classical" examples help to illustrate the phenomenon. In a seminal 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 broad shoulders, whereas patients characterized as suspicious would often highlight the eyes in the drawings, although in fact, all combinations occurred 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 trait types, participants believed to have seen mostly expected pairings, such as accountant/perfectionist or doctor/helpful.

While the illusions obtained in the two aforementioned 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. In a typical task set-up of this kind, illusory correlations arise because present causes or effects are given more weight than absent causes and effects. Text box 6.1 presents one typical, often-cited example of such a study 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 v. 7, respectively, both with and without the fertilizer). In contrast, when the absolute blooming frequency was low (e.g., 7 v. 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. Their thought-provoking finding that the same high rate of positive behavior is worth more in a large group (majority) than in a small group (minority), which was replicated in countless other experiments even with children (Primi & Agnoli, 2002), has obvious implications for the creation of minority devaluation and discrimination.

### **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.

### Definitions

From these various examples, it is evident that the term “illusory correlation” does not always refer to the same phenomenon. It is useful to distinguish four different variants that may produce illusory correlations independently or in conjunction: *expectancy-based* illusory correlations, illusions arising from *unequal weighting of present versus absent information*, illusory correlations due to *unequal sample size* (e.g., more opportunity to learn about a majority than about a minority), and the *alignment of skewed base-rate distributions* (pseudocontingencies).

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, let 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 a  $2 \times 2$  table as in Table 6.1. Cell entries  $a, b, c, d$  indicate the joint frequencies with which good and bad weather co-occur 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

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

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

	<i>Attribute <math>y^+</math> Effect present</i>	<i>Attribute <math>y^-</math> Effect absent</i>
Attribute $x^+$	Cell A	Cell B
Cause present	Frequency $a$	Frequency $b$
Attribute $x^-$	Cell C	Cell D
Cause absent	Frequency $c$	Frequency $d$

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 phi-coefficient

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

Although the choice of an appropriate 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. However, 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 information load 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 consists of 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* into 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,  $\Delta, \phi$ , etc.). Yates et al. (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 are sensitive to 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

Let us now consider the different theoretical explanations that have been advanced to account for illusory correlations: *expectancies*, *unequal weighting of present versus absent information*, *unequal sample size* (e.g., more opportunity to learn about a majority than about a minority), and the *alignment of skewed base-rate distributions* (pseudocontingencies).

### *Expectancy-based illusory correlations*

A major domain of *expectancy*-driven illusory correlations is the study of 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 good weather generally improves mood. When participants are uncertain of the number of smiling and frowning faces associated with good versus bad weather, they may follow their prior expectancies to guess the number and then make a judgment. Expectancies can particularly influence the initial perception and encoding, especially when stimulus observations are ambiguous so that prior knowledge is required to classify observations according to  $x$  and  $y$ . One may also assume that expectancies facilitate the learning of 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, Greve et al., 2019).

*Expectancy*-based illusory correlations are often confused 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 a common perceptual feature (e.g., same color, common symbols, smile in both the face and the sun) while holding expectancies constant. Such overlap in common features may enhance the experienced correlation (Fiedler et al., 2008; Plessner et al., 2000), but unlike expectancy effects, this reflects a stimulus-driven encoding influence.

### *Unequal weighting of present versus absent information*

Alternatively, biased correlation assessments may reflect the *unequal weighting* of the four cells in Table 6.1. More weight is typically 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 et al., 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 et al., 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). Consequently, two formally equivalent 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). Researchers often attribute such unequal weighting to enhanced salience of present features in 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, enhanced salience and a memory advantage have also been attributed to rare or distinctive attribute levels, according to the famous von-Restorff (1933) effect. As explained earlier in Text box 6.2, the same high proportion of 75% desirable behaviors in a minority leads to less positive impressions than the same proportion observed in a majority, even though the constant proportion implies a zero correlation. The *distinctiveness* account (Hamilton & Sherman, 1989) 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.

### ***Unequal sample size***

However, the available evidence does not support an encoding and memory advantage of the most infrequent cell or stimulus combination (Fiedler et al., 1993; Klauer & Meiser, 2000; Ernst et al., 2019). The phenomenon can be explained more simply as a *sample-size effect*. Usually, we encounter the majority more often than the minority and desirable behaviors more often than undesirable ones (Fiedler, 1996). Learning theories predict that learning increases with the number of trials, and it is even logically justified to draw stronger inferences from larger samples following the Bayesian rule of succession (Costello & Watts, 2019).

The sample-size account can be set apart from the distinctiveness account when illusory correlations are studied in a hypothesis-testing paradigm (see Chapter 4 in this volume). Translating pertinent findings (cf. Fiedler et al., 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. For instance, participants can search 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 the rate of good mood 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 or strategy of information search, rather than an encoding focus on the rarest events, which are by definition the least attended stimuli. Therefore,

an illusory inference arises that good mood was more likely on sunny than on rainy days could only be due to the enhanced opportunity to learn from a larger sample. It could not be attributed to the enhanced salience or attention given to the rarest events.

### ***Alignment of skewed base-rate distributions***

For the accounts discussed so far, it is necessary to assume that reasoners represent joint frequencies ( $a, b, c, d$ ) in memory. This is easy to imagine in our simplified example of weather and mood but becomes increasingly difficult when correlations between many different attributes are assessed simultaneously or when attributes can take on more than two levels. With three levels of weather and mood (e.g., adding a neutral category), we must keep track of nine joint frequencies. In an easily available, more intricate example of multivariate correlation assessment, when mood, speed, and accuracy covary with weather, fatigue, day time, and social settings, the required number of joint frequencies may be no longer manageable. In these cases, the *pseudocontingency heuristic* (e.g., Fiedler et al., 2009, 2013) offers a fast and frugal algorithm to infer contingencies from the alignment of frequency distributions, independently of joint frequencies. In our example, when individuals only track the marginal frequencies (e.g., how frequent is good and bad weather, how often does good and bad mood occur), they infer a positive contingency when frequent good and infrequent bad mood is aligned with frequent good and infrequent bad weather. In contrast, they infer a negative contingency when mostly good (bad) mood is misaligned with mostly bad (good) weather.

Thus, when individuals in the prior example learn that in 30 diary entries, 21 reports show good mood and only nine bad mood, and 20 days have good weather, they assume that mood and weather are positively related, even when no joint observations are presented (e.g., Fiedler & Freytag, 2004; Meiser, 2006; Meiser & Hewstone, 2004; Meiser et al., 2018; Vogel et al., 2013). It would be possible for the same marginal frequencies that there was good mood on all bad-weather days, and that bad mood occurred only on good-weather days, indicating a negative relationship. Nevertheless, the alignment of mostly good mood with mostly good weather would induce a positive illusory correlation.

### ***Illusory-correlation classroom demonstration***

The following demo experiment is intended to elucidate the interplay of top-down expectancies and bottom-up stimulus effects in correlation assessment. It resembles a simplified combination of the two studies conducted by Fiedler et al. (1999).

Correlation assessment is embedded in a hypothesis-testing task. Participants are asked to test the hypothesis that “partnership problems arise when male aggression is overt and female aggression is covert”. Accordingly, they are presented with a series of images, 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  $2 \times 2$  distribution (male v. female  $\times$  present v. 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 (75%), so that aggression is more often present than absent both in males and in 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. When larger samples provide more opportunities to learn the 75% confirmation rate for male overt aggression and female covert aggression than for the contrary, the stimulus-driven learning effect should reinforce the expectancy-driven illusion.

In contrast, when absolute sample size is larger for male covert and female overt aggression than for male overt and female covert aggression, the impact of opportunity to learn should 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.

### **Text box 6.3**

#### **Procedural details for the classroom demonstration**

##### ***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.

##### ***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 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 the male and female target.

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 assessed by asking participants to rate the degree to which they believe that males and females tend to exhibit overt and covert aggression, and the participants' own gender should be assessed.

The cover story embeds the hypothesis in a diagnostic task setting: The male and female target are 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.

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 the male than the female partner, whereas covert aggression is more likely to be paired with the female than the male partner."

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 eight or ten 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, for example, 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.

Two 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; and (b) impression ratings of

the male and female target on adjective scales related to overt (violent, aggressive, hot-headed, irritable) and covert (insidious, sanctimonious, scheming, dishonest) aggression. The order of the two dependent measures may vary.

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

## **Results**

Let us first summarize the theoretical predictions before we turn to the analysis of the experimental results. Two possible findings can be expected. On the one hand, expectancy-based illusory correlations could be manifested in a general tendency to confirm the gender stereotype: Responses to all dependent measures could, 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 could produce a different kind of illusory-correlation effect: 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 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).

The results reported below are borrowed from the Fiedler et al. (1999) experiments, which slightly differed from the present experiment in terms of the specific sample size. 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,  $t(73) = 4.18, p < .001$ .

Turning to the dependent measures proper, let us consider the results 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,  $F(1, 31) = 16.98, p < .001$ .

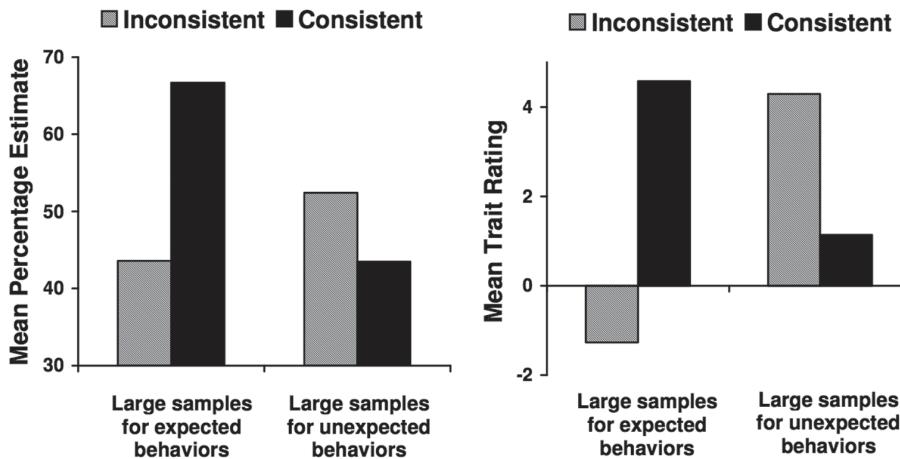


Figure 6.1 Mean percentage estimates (left chart) and mean impression ratings (right chart) for expectancy-consistent (black bars) and expectancy-inconsistent (shaded bars) attributes as a function of sample-size conditions.

Likewise, on the impression-rating measure, the mean rating for targets on expectancy-consistent traits (e.g., male-violent, female-dishonest) was higher (+4.58) than for inconsistent traits (-1.27),  $F(1, 31) = 8.13, p = 0.008$ . 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).

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 (52.3%) 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 for 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 every other task setting. 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 payoff 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 objects in the environment is an important module of adaptive intelligence and behavior. Although humans and animals seem to have the competence to assess environmental correlations quite accurately, their performance is often impaired under suboptimal 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 back-side 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 has to be accurate but also quick. It 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  $\phi$  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 marginal frequencies (using the pseudocontingency heuristic explained above) provides a perfect estimate of the correlation. Aligning skewed base-rates will only be misleading when marginal frequencies are unequal. 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 irrationality.

## Summary

- The detection and assessment of environmental correlations is an important module of adaptive intelligence.
- The phenomenon of illusory correlations refers to failure and inaccuracy in correlation assessment.
- Different types of illusory correlations reflect different underlying cognitive processes: Illusions can be based on *expectancies, unequal weighting of present versus absent information, unequal sample size* (e.g., more opportunity to learn about a majority than about a minority), and the *alignment of skewed base-rate distributions* (pseudocontingencies).
- 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, economic 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. Another approach by Allan et al. (2008) complements associative learning with a psychophysical signal-detection framework that takes payoff structures and response strategies into account. More recently research in clinical psychology has connected increased susceptibility to illusory correlations with schizophrenia patients (Balzan et al., 2013; Beer et al., 2012). Alloy and Tabachnik (1984) offer an intriguing view on human and animal performance on correlation-assessment tasks. The notion of pseudocontingencies is explained in Fiedler, Kutzner, and Vogel (2013). Fiedler (2000) provides a review of different variants of illusory correlations within the same connectionist framework.

## Acknowledgment

The authors’ own research on illusory correlations was supported by several DFG grants awarded to the first author.

## 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.
- Balzan, R. P., Delfabbro, P. H., Galletly, C. A., & Woodward, T. S. (2013). Illusory correlations and control across the psychosis continuum: The contribution of hypersalient evidence-hypothesis matches. *Journal of Nervous and Mental Disease*, 201(4), 319–327.
- Beer, K., Moritz, S., & Lincoln, T. M. (2012). Illusory correlations in paranoid schizophrenia: Another cognitive bias relevant to delusions? *Journal of Experimental Psychopathology*, 3(4), 661–672.
- Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, 72, 193–204.
- Costello, F., & Watts, P. (2019). The rationality of illusory correlation. *Psychological Review*, 126(3), 437–450.
- Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin*, 90(2), 272–292. <https://doi.org/10.1037/0033-2909.90.2.272>
- Ernst, H. M., Kuhlmann, B. G., & Vogel, T. (2019). The origin of illusory correlations. *Experimental Psychology*, 66(3), 195–206.
- 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., Bluemke, M., Freytag, P., Unkelbach, C., & Koch, S. (2008). A semiotic approach to understanding the role of communication in stereotyping. In Y. Kashima, K. Fiedler, & P. Freytag (Eds.), *Stereotype dynamics: Language-based approaches to the formation, maintenance, and transformation of stereotypes* (pp. 95–116). Mahwah, NJ: Lawrence Erlbaum.
- Fiedler, K., & Freytag, P. (2004). Pseudocontingencies. *Journal of Personality and Social Psychology*, 87(4), 453–467. <https://doi.org/10.1037/0022-3514.87.4.453>
- Fiedler, K., Freytag, P., & Meiser, T. (2009). Pseudocontingencies: An integrative account of an intriguing cognitive illusion. *Psychological Review*, 116(1), 187–206.
- 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.
- Greve, A., Cooper, E., Tibon, R., & Henson, R. N. (2019). Knowledge is power: Prior knowledge aids memory for both congruent and incongruent events, but in different ways. *Journal of Experimental Psychology: General*, 148(2), 325–341.
- 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 analysis 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.
- 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.
- Meiser, T. (2006). Contingency learning and biased group impressions. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 183–209). New York: Cambridge University Press.
- Meiser, T., & Hewstone, M. (2004). Cognitive processes in stereotype formation: The role of correct contingency learning for biased group judgments. *Journal of Personality and Social Psychology*, 87(5), 599–614.
- Meiser, T., Rummel, J., & Fleig, H. (2018). Pseudocontingencies and choice behavior in probabilistic environments with context-dependent outcomes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(1), 50–67.
- 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.
- Vogel, T., Kutzner, F., Fiedler, K., & Freytag, P. (2013). How majority members become associated with rare attributes: Ecological correlations in stereotype formation. *Social Cognition*, 31(4), 427–442.
- 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

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

<i>Overt aggression</i>	<i>Covert aggression</i>
4. Threatens with violence	4. Lies to get an advantage
5. Screams when s/he doesn't like something	5. Makes a pretense on 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 go 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 other 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).

## 7 Causality bias

*Helena Matute, Fernando Blanco, and  
María Manuela Moreno-Fernández*

The causality bias, also known as the illusion of causality, occurs when people believe that two events that occur together by mere chance are causally related. This happens in many situations in our daily life. Just think, for instance, of your friend in school wearing a “lucky wristband” and obtaining a good grade. You may also think of days when you had a headache, you took a remedy that your neighbor recommended, and you felt better the next day. You probably assumed that that remedy was actually the cause of your recovery. This, however, could be as illusory as the magic wristband of your friend, that is, it could be an illusion of causality. The causality bias may sometimes be innocuous, but it can also have serious consequences. For instance, people follow bogus treatments instead of evidence-based therapies, sometimes when suffering from severe health problems, which can lead to fatal consequences (Freckleton, 2012; Lim et al., 2010; US Food and Drug Administration [FDA], 2017).

So, how do psychologists know that your strong intuition of causality about taking that remedy and feeling better could actually be an illusion? In this chapter, we will show that intuitions of causality, as strong as they may seem, can often be biased. Many experiments have shown that people have the potential to detect causal relationships accurately under certain circumstances (Shanks & Dickinson, 1987; Wasserman, 1990), but experiments have also shown that the causality bias is a robust and very common phenomenon that can affect all of us (Alloy & Abramson, 1979; Chow et al., 2019; MacFarlane et al., 2018; Matute, 1996; Matute et al., 2019; Willett, & Rottman, 2021).

In real life, it is often difficult to identify when a causal relationship is illusory. In laboratory experiments, by contrast, researchers can set up the situation to make sure that there is no objective causal relationship between two events. To this end, they may program the two events to occur independently of each other. In this way, researchers can then test in which cases people develop the bias of causality and which conditions can be used to reduce this bias. Thus, the question today is not whether this bias occurs, but which conditions increase or reduce its strength.

In a standard causality bias experiment, participants are exposed to a series of trials in which a potential cause, C, and a potential outcome, O, may or may not occur in each trial, and the task of the participants is to judge the degree to which both events are causally related (see Matute et al., 2015, 2019, for reviews). For instance, participants may play the role of a physician who is observing the records of a series of fictitious patients. Each patient’s record is presented separately, one per trial. Each record shows whether that patient took a given drug (or not) and whether the patient reported feeling better (or not). After observing a number of records of fictitious patients (typically between 30

and 100, depending on the experiment), participants are asked to give their judgment of causality, that is, their subjective estimation as to whether the drug is the cause of the recovery, by using a numerical scale (usually from 0 to 100). Of course, the experiment does not always involve a drug as the potential cause and a recovery as the outcome. These experiments may involve any two events that could, in principle, be causally related. For instance, experiments have also been run using fertilizers as the potential cause and flowers blooming as the potential effect, meals as the potential cause and allergic reactions as the outcome, and many other stimuli. In any case, it is important to recall that the two target events in these experiments are always programmed to occur independently of each other, that is, the outcome occurs with the same probability regardless of whether or not the potential cause is present. This should be a clear indication that the two events are uncorrelated. However, when certain conditions are met, participants usually tend to develop the illusion that the potential cause is actually causing the outcome. Current research shows which factors can either increase or reduce causal illusions. We will address some of these factors in this chapter.

Table 7.1 shows the four possible combinations (a, b, c, and d) that can take place when a potential cause and a potential effect are sometimes present and sometimes absent (see also Chapter 6 for a related analysis). For instance, you may think of the relationship between taking a given pill and feeling better the next day. If someone told you that, out of 100 patients who took the pill, 75 felt better the next day, you may conclude that the pill is highly effective. However, this conclusion could be strongly biased. Indeed, you also would need to know what happened to another group of 100 patients who did not take the pill. This should be, at minimum, your control condition. That is, you really need to know the two probabilities before you reach any conclusion. In other words, you need to know the probability that patients feel better after taking the pill,  $P(O|C)$ , and compare this number against the probability of patients feeling better without taking the pill,  $P(O|\text{no}C)$ . Given that in our example the two probabilities are identical, we should conclude that the causal relationship is zero.

The most common index used to compute the statistical contingency (i.e., covariation) between the two events is called the  $\Delta P$  contingency index (Allan, 1980). It simply subtracts the probability of the outcome in the absence of the cause from the probability of the outcome in the presence of the cause. In our example this difference is zero, therefore the contingency between the two events is null. That is, this value suggests that there is no causal

Table 7.1 Contingency table

	<i>Outcome present</i>	<i>Outcome absent</i>	
Cause present	a	b	$P(O C) = .75$
	75	25	
Cause absent	c	d	$P(O \text{no}C) = .75$
	75	25	

*Note:* The table shows the four possible combinations that can occur (a, b, c, and d), with the potential cause and the potential outcome being present or absent. The numbers in each cell represent the frequencies of each trial type in a hypothetical situation in which there are 75 a, 25 b, 75 c, and 25 d trials. In this particular example, the probability of the outcome is the same when the potential cause is present (top row) as when it is absent (bottom row), that is, .75. Therefore, there is no contingency between the cause and the outcome in this example.

relationship between the two events. By contrast, when  $P(O|C)$  is higher than  $P(O|\text{no}C)$ , the contingency is positive. Only in these cases does the contingency value suggest that the presence of the potential cause has a positive (or generative) effect on the outcome. For instance, a pill causes recovery, because recovery is more probable when we take the pill than when we do not. On the other hand, if  $P(O|\text{no}C)$  were higher than  $P(O|C)$ , then this would mean that the contingency is negative, thus suggesting a preventive causal relationship between the two events, or, in other words, that the cause prevents the occurrence of the outcome. In any case, experiments studying the causality bias, or illusion of causality, typically use a null contingency similar to the example shown in Table 7.1. That is, in these experiments the outcome occurs with the same probability in the presence and in the absence of the potential cause. If the participants estimate that the strength of the causal relationship is higher than zero, then we say there is a causality bias or an illusion of causality.

There are two basic scenarios where the bias of causality is typically observed in the psychology laboratory (and in real life as well): passive and active procedures. In a standard passive procedure (Blanco & Matute, 2019; Matute et al., 2011; Yarritu et al., 2014), the experimental participant observes a series of trials in which two external events, C and O, can either co-occur or not, thus leading to a sequence of events of type a, b, c, and d, as in Table 7.1. At the end of these trials, the participants give their judgment on whether there is a causal relationship between the two external events. For instance, observing some patients taking a remedy and reporting feeling better the next day would be an example of the passive procedure. In the active procedure (Blanco et al., 2011; Matute, 1996; Moreno-Fernández & Matute, 2020; Yarritu et al., 2014), the outcome is also an external event, but the cause depends on what you do. For instance, rather than simply observing the patients taking a remedy and feeling better or not, you may be the one administering the remedy to some (or all) of the patients. Thus, in this case, the potential cause could be your own behavior rather than an external event.

The causality bias occurs both in the active and in the passive procedures. However, in the active procedure in which the potential cause depends on the participants' behavior, this is often known as an illusion of control (Langer, 1975). Indeed, the illusion of control can be regarded as a special type of the illusion of causality in which the participants believe not only that there is a causal relationship between the two events, but also that their own behavior is what is causing the outcome. From our perspective this is the main difference between the illusion of control and the illusion, or bias, of causality. However, and because of this additional, self-related, component, the illusion of control is an instance of the illusion of causality whose unique features others have explored in detail. Given that there is another chapter in this book devoted to the particular topic of the illusion of control (see Chapter 8), we will not cover these self-related aspects in this chapter. Instead, we will focus on the more general aspects of the bias of causality which are common to both phenomena, that is, the illusory perception of a causal relationship, regardless of whether the potential cause is the participant's behavior or an external event. Thus, we will refer to the two target events as the cause and the effect, assuming that the cause can either be the participants' behavior or an external event, and will not address those self-referential aspects that might be exclusive of the illusory control.

## **Empirical evidence**

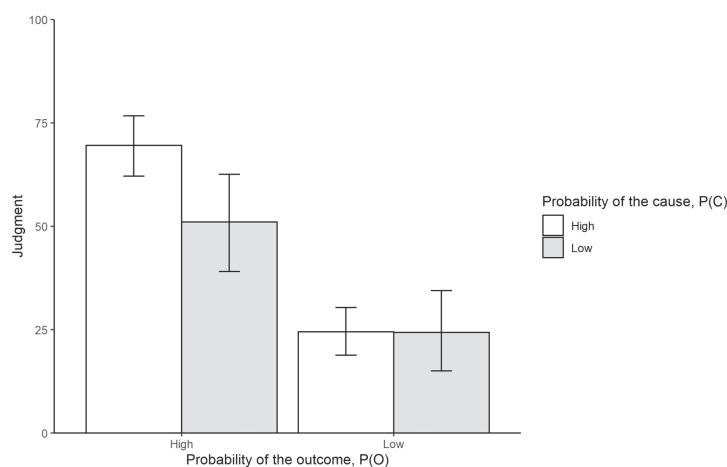
According to multiple studies, the causality biases that were described above are driven by two different effects: cause-density bias and outcome-density bias. That is, judgments of

causality tend to increase when either the target cause or the outcome appear with high probability, even if the actual contingency is null.

The evidence for the outcome-density bias is abundant. For instance, Alloy and Abramson (1979) asked participants to judge the control they exerted over the onset of a green light by pressing a button (i.e., the experiment used the “active procedure” to produce an illusion of control). Although the contingency between button and light was null, participants overestimated their control more when the light came on frequently (high outcome density) than when it came on rarely (low outcome density). A similar result has been reported repeatedly by using a variety of experimental settings, and using both the active and the passive procedures (see a review in Matute et al., 2019; see also Chapter 8 for additional examples using the active procedure).

Likewise, the cause-density bias consists of the overestimation of causality that appears when the cause is presented often, compared to when it is presented rarely. The cause-density bias has also been replicated in many different studies (Hannah & Beneteau, 2009; Matute et al., 2011; Vadillo et al., 2011), using either the passive or the active procedures (Blanco et al., 2011; Matute, 1996).

Furthermore, there is experimental evidence suggesting that the strongest of biases can be observed when the two conditions, high outcome density and high cause density, are met simultaneously (Blanco et al., 2013; see Figure 7.1). Interestingly, it is in these latter situations where it is possible to observe a large amount of type “a” events, or coincidences between the target cause and the outcome (see the corresponding cell in Table 7.1), even if the contingency is null. Hence, although other factors may also be important, the causal illusion (be it observed in either active or passive procedures) seems to rest primarily on this high number of cause-outcome coincidences.



*Figure 7.1* The combined effect of outcome- and cause-density bias in a condition of null contingency.

*Note:* Error bars depict 95% confidence intervals for the mean. When both the target cause and the outcome are presented with high probability, judgments of causality tend to show a strong bias, as shown in several experiments. This figure has been created from the data reported in Blanco et al. (2013, Experiment 1). Although both the probability of the cause and the probability of the outcome contribute to illusions, these are noticeably stronger when the two probabilities are high.

## Theoretical background

Several models that were initially proposed to account for associative and contingency learning have been applied to the above-described causal biases. These theoretical models differ in their approach and degree of formalization (see a review in Perales & Shanks, 2007).

First, normative models describe the computations that would be needed to accurately estimate causal relations. Thus, the contingency index  $\Delta P$  (Allan, 1980) that serves to assess the contingency between causes and outcomes (see the previous section) can be considered as a normative model for human causal learning. The  $\Delta P$  index is computed as the difference between two conditional probabilities, the probability of the outcome in the presence of the cause, and the probability of the outcome in the absence of the cause. Hence, the rule compares the frequency with which the outcome appears when the cause is present to a scenario where the cause is absent. Many experiments have shown that people and other animals' behaviors are sensitive to variations in contingency (Allan & Jenkins, 1983; Rescorla, 1968). However, as we already noted, there are also cases in which people do not show such sensitivity, and tend to show, instead, a causality bias (Matute et al., 2015, 2019). As the  $\Delta P$  rule describes a normative inference, it cannot account for these deviations from the norm and biases in general.

A second model that also belongs to this category is the power theory of probabilistic contrast, or power PC (Cheng, 1997). This theory is based on  $\Delta P$ , but it incorporates the notion of interactive causes. When we observe the outcome in the absence of the target cause (i.e., type "c" events in Table 7.1), we need to assume that an additional, hidden causal factor is responsible for these occurrences. For illustration, imagine that you are assessing the effect of your pressing of a button on the operation of an elevator. If you observe that the elevator moves without you pressing the button (i.e., a type "c" event), you must assume that an additional cause, such as a person on another floor, is producing the movement. Thus, when one makes a causal inference, the causal influence of the target cause must be isolated from that of any other cause operating in the background. Like the previous model, power PC is intended to be a normative description of how optimal causal inferences should be performed. However, it can predict outcome-density biases in a limited set of situations (Buehner et al., 2003). The problem with Power PC is that the participants' judgments are not always affected by factors that should influence power, hence many experimental results cannot be predicted by this theory.

Another approach is that of algorithmic models that describe rules of information processing that are not normative, thus allowing the prediction of biases and illusions. One example is weighted models, which describe causal estimation rules that include "weights" to make some pieces of information more important than others when making a judgment. For instance, the weighted version of  $\Delta P$  incorporates four weight parameters to represent the importance of each type of observation in Table 7.1 (a, b, c, and d). This endows the rule with enough flexibility to account for both accurate (i.e., when the weights are all similar to each other) and biased estimations of causality (when the weights differ).

Certain attempts to estimate these psychological weights in a typical experimental scenario (Wasserman et al., 1990) suggest that, when people make their causal judgments, type "a" events are the most important ones (i.e., they impact more strongly on the perception of causality), whereas type "d" events are given lesser weight (see also Chapter 6). This result is compatible with the causal biases exhibited by experimental participants: If

type “a” events are given much more importance when judging causality, then the bias should be stronger when these trials are frequent (i.e., high cause- and outcome-density situations).

Finally, a different family of theories is that of associative models, which were initially developed to describe associative learning in non-human animals. These models do not aim to describe processes from a normative view. They propose that the mental representations of the cause and the outcome can be connected by means of associations that are built on the basis of experience. The strength of these associations can change thanks to iterative, simple algorithmic rules based on the correction of the prediction error. That is, the algorithm makes a prediction of the outcome on each trial which is based on the strength of the association, and then this strength (and hence, the prediction), is corrected on the basis of the actual outcome.

Perhaps the most celebrated of associative models is the Rescorla-Wagner model (Rescorla & Wagner, 1972). The model can readily predict instances of both accurate and biased estimations of causality (Matute et al., 2019). An interesting prediction of this model is that both cause- and outcome-density biases appear only at the beginning of the training session, before they progressively wear off (i.e., they are called “pre-asymptotic” biases; see Figure 7.2).

The predictions made by associative models, and in particular the pre-asymptotic nature of causality biases, have been studied by researchers. However, the evidence remains mixed and inconclusive. For instance, while certain experiments showed a pre-asymptotic bias that is compatible with the model’s predictions in Figure 7.2 (Murphy et al., 2011), other experiments increased the training length and found no evidence of a reduction in the bias (Barberia et al., 2019).

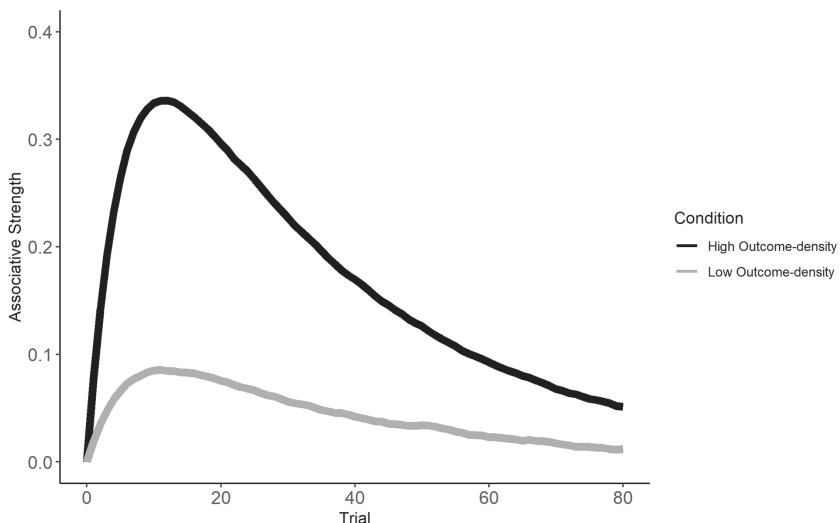


Figure 7.2 Simulations of the Rescorla-Wagner model.

Note: These simulations illustrate the outcome-density bias in a null contingency setting. When the outcome appears frequently (high outcome-density condition), the associative strength of the target cause is larger than when the outcome appears less frequently (low outcome-density condition). However, the bias is pre-asymptotic and tends to disappear as training progresses.

Thus, to sum up, several theories have been proposed to understand causal biases. All of them show promise in some respects, but also limitations that need to be further investigated to account for all the empirical results reported in the literature. Additionally, these results also suggest that the bias is complex and subject to additional factors, such as, for instance, contexts and expectations, which future researchers will need to incorporate within a theoretical framework.

## Individual differences

Causality and contingency estimation have been usually interpreted as the result of general learning/reasoning processes that are present in all individuals. We have already noted that causal illusions, or causality biases, appear readily in situations in which either the cause or the outcome appears frequently. Nevertheless, there are individual differences in the vulnerability to the bias, or in the way the bias is expressed. We will review some of these differences next.

The first step to correctly estimate a causal relation is that the information received by the participant needs to be of high quality (i.e., accurate, complete, and representative). Basically, to form a causal judgment people need an accurate representation of each of the events presented in Table 7.1 (a, b, c, and d events). In many situations, people are merely exposed to these different pieces of information, but there are other circumstances in which they are active agents who obtain this evidence (e.g., the active procedure described above). In these cases, individual differences in the strategy used for collecting the information may easily bias their causal beliefs.

Imagine, for example, that you are offered a vaccine against a dangerous disease. Imagine also that the vaccine has been developed recently, so you have never heard about it. Since you are afraid about its potential side-effects, you resort to the Internet to obtain information that may help you make an informed decision. In this situation, it is likely that you start your search by introducing the name of the vaccine on the search engine. This situation will allow you to mostly retrieve type “a” and “b” events, that is, pieces of information about the vaccine and its effects. However, this strategy will probably underrepresent type “c” and “d” information (crucial to accurately assess the relation between the vaccine and its potential side-effects). This in turn will increase the probability of experiencing causality biases due to a cause-density effect as described above (note that other strategies may also produce an unbalanced and unrepresentative sample of information).

In line with the previous example, recent research has reported that people use different strategies when searching for causal information, and that these strategies may actually modulate causal estimations producing a causality bias when type “a” events are frequent (Moreno-Fernández & Matute, 2020).

More generally, certain factors could affect people’s behavior so that, when faced with a task that uses the active procedure in a contingency learning task, people differ in the amount of cause-present trials that they observe. Consequently, those people who observe more cause-present (a and b) trials will exhibit a cause-density bias.

Thus, motivation, previous knowledge, expectations, and even mood (Blanco et al., 2011; Blanco & Matute, 2019; Yarritu et al., 2014) have been considered as potential sources of individual differences influencing the amount and type of trials that people collect and consider. For example, Blanco et al. (2012) conducted a study

on the illusion of control. Consistent with previous reports by Alloy and Abramson (1979), they found that college students with higher scores in depressive symptomatology (dysphoria) were more accurate in their causal judgments of a null contingency setting than those with lower scores (an effect known as depressive realism; Moore & Fresco, 2012; Msetfi et al., 2005). Then, these researchers showed that the reason why dysphoric participants were less biased is because they tended to be more passive, thus they introduced the potential cause less frequently, and thus they received more exposure to trials in which the potential cause was absent. By contrast, non-dysphoric participants tended to act and introduce the potential cause in most trials, and therefore they saw proportionally more coincidences (type “a” trials), which in turn increased their illusion of causality.

Nevertheless, differences in mood and in information-sampling strategy are not the only source of individual differences in the vulnerability to causal biases. Another one is known as jumping to conclusions. This trait, which is defined as the tendency to make conclusions based on scarce data, is connected to delusion formation and maintenance (Garety et al., 2013; Garety & Freeman, 2013). Moreno-Fernández et al. (2021) assessed the relation between the tendency to jump to conclusions and causal illusions. Their results showed that participants with higher tendency to jump to conclusions were also those with higher causal judgments on a null contingency task (like the ones described at the beginning of this chapter). That is, those participants who developed higher illusions of causality were also those individuals with a higher tendency to jump to conclusions. The authors suggested that this trait may bias our judgments of causality through a mechanism based on the differential salience of each type of evidence, following the proposal of weighted models of the illusion of causality that was described in a previous section (see Balzan et al., 2012; Speechley et al., 2010). Thus, if I expect two events to be related (e.g., I believe a vaccine is linked to severe side-effects), then the confirming evidence (e.g., type “a” events) will become more distinctive and salient, producing a biased overestimation of the actual relation between the two events that may facilitate a premature stop of data collection.

This differential salience mechanism may explain individual differences in situations in which people can decide about the amount of evidence that they are willing to collect, but it may also explain some individual differences in situations in which neither the amount nor the type of information that people are exposed to can be controlled. As we will detail in the next section, individual differences have also been reported in passive procedures when motivational factors may play a role (Blanco et al., 2018). Thus, attitudes and personal preferences have been found to modulate and bias our judgments of causality, making us more vulnerable to experience illusions of causality when information is congruent with our preferences and inclinations.

## Applications

Interesting applications of the causality bias can be found in many areas of our life. Below we describe some of them, but before you read about potential applications and think about additional examples, perhaps you might benefit from preparing and running one of these experiments yourself in order to make sure you understand how this bias works (see Text box 7.1).

**Text box 7.1 A classroom demonstration of causality bias**

As a classroom demonstration of causality bias, we suggest a modified version of the procedure used in Moreno-Fernández et al. (2017) in which the outcome-density effect is assessed.

**Method*****Participants***

Assuming  $\alpha$  and  $\beta$  error probabilities to be 0.05 and 0.20, respectively, this experiment needs 40 participants to optimally test for a large effect size (Cohen's  $d = 0.80$ ), or 100 participants to test for a medium effect size (Cohen's  $d = 0.50$ ) using a one-tailed between-groups  $t$ -test.

***Materials***

This experiment uses the passive contingency learning procedure in which the potential cause and the potential outcome are either presented or not in each trial. The sequence of trials can be presented to participants by a variety of means (e.g., as slides in a computer presentation, as pages in a booklet ...). Materials and instructions to build the training and test trials for each group can be found in the Appendix.

***Design***

This experiment uses a null contingency design ( $\Delta P = 0$ ) with only one between group factor (outcome density) manipulated in two levels: high vs. low (see Table 7.2 for additional details). The dependent variable is the causal judgment collected in a 0–100 scale.

**Procedure**

The experiment can be conducted in one session. At the beginning of the session participants are randomly assigned to one of two groups (High or Low outcome density) and are asked to observe the corresponding version of the experiment. Both versions ask participants to assess the effectiveness of an experimental fertilizer that may promote plants' growth. To achieve this goal, a series of 48 trials in which either the fertilizer is used or not, and either the plants grow or not (i.e., the four types of events in Table 7.1), are sequentially presented. After observing all trials, participants are asked to rate the effectiveness of the product by giving a value between 0 and 100. Note that, for both groups, the probability that the plants will grow does not depend on the fertilizer being used. That is,  $P(O|C) = P(O|\text{no}C)$ , which means that the contingency is null and the fertilizer is, in principle, not effective. However, there are more “a” and “c” trials in the High outcome-density group than in the Low outcome-density group, which means that we are manipulating the probability of the outcome.

Table 7.2 Contingency matrixes for (a) High and (b) Low outcome-density groups

	<i>Outcome present</i>	<i>Outcome absent</i>	
<i>(a) High outcome density</i>			
Cause present	a 20	b 4	$P(O   C) = .83$
Cause absent	c 20	d 4	$P(O   \text{no}C) = .83$
<i>(b) Low outcome density</i>			
Cause present	a 4	b 20	$P(O   C) = .17$
Cause absent	c 4	d 20	$P(O   \text{no}C) = .17$

*Note:* Panels show trial frequencies in the High and Low outcome-density conditions.

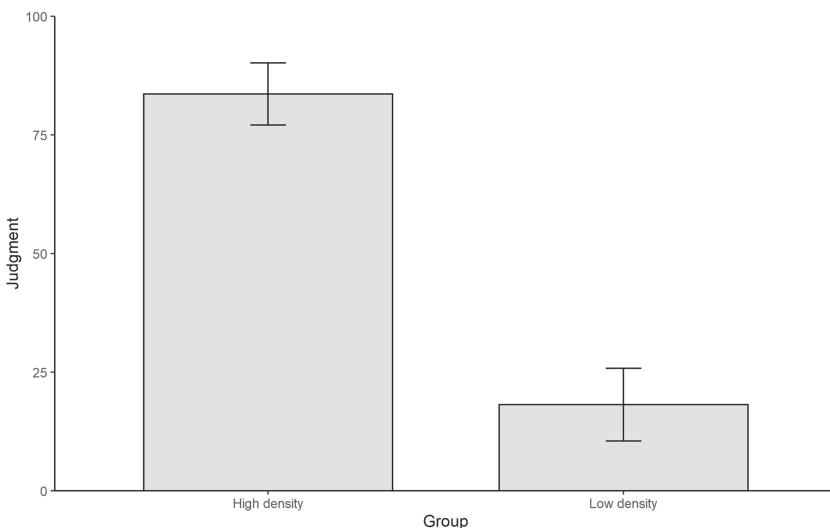
### Expected results

The results of the causality judgments can be visually described by using a bar chart like the example displayed in Figure 7.3. Although the fertilizer is equally ineffective in both groups (the plants' growth is non-contingent with its use), comparing the causal judgments of these two groups usually yields a significant difference known as the outcome-density effect: Participants in the high outcome-density group usually overestimate the effectiveness of the product to a higher extent than participants in the low outcome-density condition. The difference can be tested for statistical significance by using a one-tailed between-groups *t*-test.

### Pseudoscience

Pseudoscience is any belief or practice that pretends to be scientific, but it is not supported by evidence. One of the best examples is alternative medicine. Note that the word alternative is used to refer to all those treatments that have not been approved (because they are not supported by scientific evidence). Similarly, alternative psychological therapies are also examples of pseudoscience. In all these cases, people follow a given treatment and they are convinced that it works for them (Lilienfeld et al., 2014). They often rely on coincidences, either observed in others or experienced by themselves: "You should take this pill: I took it the other day and I feel much better now." However, because we know that people are susceptible to the causality bias, we have to be skeptical about our intuitions (and those of other people) concerning treatments that have not been approved.

Developing the illusion that there is a causal relationship is very common, and the only way to know for sure whether that treatment is really working is by submitting it to scientific scrutiny. If the scientific community and the official medical agencies tell us that a given treatment has not proven to be effective, then we should be aware that they are performing all the calculations (see Table 7.1), so even though our intuitions and those of our friends can sometimes suggest otherwise, it is very possible that we are



*Figure 7.3* Averaged causality judgments in the High and Low outcome-density groups.

Note: Error bars represent 95% confidence intervals for the means. These are simulated data created only for illustration purposes.

suffering an illusion. Research shows that they are often at the heart of pseudosciences; therefore, reducing causal biases can be an important strategy to minimize the impact of pseudosciences in our society (Chow et al., 2019; MacFarlane et al., 2018; Matute et al., 2011; Torres et al., 2020).

### *Ideology*

As we will show below, causal biases can be affected by our ideology and political attitudes. As an example, imagine this experiment. Rather than presenting fictitious patients taking or not taking a certain pill (the potential cause), and then observing whether or not they felt better as the potential effect, imagine that we change the cover story to represent a political scenario. We will now have a government of a fictitious country implementing (or not) a given policy as the potential cause at various times (trials) and then observing in each trial whether several positive indicators of wellbeing increased or not. Also, imagine the data in Table 7.1 with this scenario, so that we do program the trials in which the policy is applied (potential cause) and the cases in which the improvement (effect) occurs, to be independent of each other. That is, the probability of improvement is the same regardless of whether the government does or does not apply the policy. Therefore, nonbiased observers should conclude that the policy was not causally related to the positive indicators that occur in 75% of the occasions (as shown in Table 7.1).

However, and as we already showed with respect to health, the participants in these political-like experiments also show a bias of causality. In this case, they assume that the applied policy is the cause of the improvement when it occurs (Blanco et al., 2018). Moreover, the bias of causality in these cases is stronger when participants are told that

the government is of the same party as the one they prefer. That is, left-wing participants showed a stronger causality bias when the political party was leftist; and right-wing participants showed stronger causality bias when rightist political parties were the ones applying the policies. This research was conducted both in Spain and the UK, and the results were very similar in both cases. They suggest that our causality judgments become even more biased as a function of our attitudes and preferences. In this particular experiment, this was shown with a political preference, but quite possibly this effect could also be demonstrated with other types of strong preferences.

As an example, do you think you could develop an unbiased causal judgment with respect to something in which your preferred football team were involved?

### ***Debiasing***

The bias of causality is a pervasive phenomenon that can prompt wrong decisions in relation to health, politics, and many other important areas of our life. Thus, we should not finish this chapter without discussing some strategies that have been developed to reduce this bias.

Interventions that reduce the causality bias have basically taught people two things. First, that their causality detection abilities are far from perfect, so that, like other cognitive illusions such as those of color, size, or many others, causal illusions are also very common. That is, an important step is to show people that they are fallible, pushing them to find a solution. And second, that if they want to make accurate judgments of causality, they need to apply the basics of experimental control. At the end, the goal of these interventions is to show people how to use basic control conditions (i.e., asking for what happens when the potential cause is not present), and how to be skeptical and ask for any additional information they may need, whenever the control conditions seem nonexistent or simply not convincing to them. By doing so, it has been shown that both adolescents and adults do improve their ability to detect causal relationships accurately and do reduce their causal biases (Barberia et al., 2013, 2018). Our final recommendation is thus to teach, as early as possible in school, the importance of managing the principles of experimental control as basic debiasing tools. And doing so, not only for those who would like to become scientists, but for all citizens interested in reducing causal biases and making more informed and less biased decisions in their daily life.

### **Summary**

- A bias or illusion of causality is said to occur when people believe that there is a causal relationship between events that are actually independent of each other.
- There are two general procedures used to investigate the causality bias. One is the passive procedure, in which participants are mere observers of different pieces of information; the other one is the active procedure, in which the potential cause depends on the participants' behavior.
- Among the many factors affecting the causality bias, probably the most important ones are the probability of occurrence of the outcome and the probability of occurrence of the potential cue. The higher these probabilities, the higher the bias.
- Normative models such as  $\Delta P$  and Power PC focus on the computations that should be performed to estimate causality if humans were rational (and accurate) causality detectors. Non-normative models, such as weighted  $\Delta P$  and the associative

Rescorla-Wagner model, try to explain how causality judgments become biased under some circumstances.

- Motivation, previous knowledge, expectations, and even mood can affect the causality bias.
- Causal biases can be observed in any area of our life. Whenever our intuition suggests the existence of a causal relationship, we need to be aware that this intuition might be biased.

## **Further reading**

Chapters 6 and 8 in this book describe very interesting phenomena that are strongly related to the causality bias. Basic references to start with, and which are common to those two chapters and the current one, are those by Allan (1980) as well as Alloy and Abramsom (1979). Research on the applications of causality biases to the problem of pseudosciences is flourishing today. Some examples can be found in Chow et al. (2019), Lilienfeld et al. (2014), MacFarlane et al. (2018), and Torres et al., (2020). The article by Yarritu et al. (2014) can shed some light on the relationship between the illusion of control (see Chapter 8) and the illusion of causality. Research on debiasing strategies to reduce the causality bias can be found in Barberia et al. (2013, 2018). Further reviews of the causality bias are provided in Matute et al. (2015, 2019).

## **Acknowledgment**

The cited studies of the authors have been supported by grants PSI2016-78818-R, PSI2017-83196-R, and RTI2018-096700-J-I00 from the Agencia Española de Investigación and by grant IT955–16 from the Basque Government.

## **References**

- Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgement tasks. *Bulletin of the Psychonomic Society*, 15, 147–149.
- Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation*, 14(4), 381–405.
- Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General*, 108(4), 441–485.
- Balzan, R. P., Delfabbro, P., Galletly, C., & Woodward, T. (2012). Confirmation biases across the psychosis continuum: The contribution of hypersalient evidence-hypothesis matches. *British Journal of Clinical Psychology*, 52(1), 53–69.
- Barberia, I., Blanco, F., Cubillas, C. P., & Matute, H. (2013). Implementation and assessment of an intervention to debias adolescents against causal illusions. *PLoS ONE*, 8(8), e71303.
- Barberia, I., Tubau, E., Matute, H., & Rodríguez-Ferreiro, J. (2018). A short educational intervention diminishes causal illusions and specific paranormal beliefs in undergraduates. *PLoS ONE*, 13(1), e0191907.
- Barberia, I., Vadillo, M. A., & Rodríguez-Ferreiro, J. (2019). Persistence of causal illusions after extensive training. *Frontiers in Psychology*, 10(24), 1–9.
- Blanco, F., Gómez-Fortes, B., & Matute, H. (2018). Causal illusions in the service of political attitudes in Spain and the UK. *Frontiers in Psychology*, 9, 1033.
- Blanco, F., & Matute, H. (2019). Base-rate expectations modulate the causal illusion. *PLoS ONE*, 14(3), 1–25.
- Blanco, F., Matute, H., & Vadillo, M. A. (2011). Making the uncontrollable seem controllable: The role of action in the illusion of control. *Quarterly Journal of Experimental Psychology*, 64(7), 1290–1304.

- Blanco, F., Matute, H., & Vadillo, M. A. (2012). Mediating role of activity level in the depressive realism effect. *PLoS ONE*, 7, e46203.
- Blanco, F., Matute, H., & Vadillo, M. A. (2013). Interactive effects of the probability of the cue and the probability of the outcome on the overestimation of null contingency. *Learning & Behavior*, 41(4), 333–340.
- Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6), 1119–1140.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104(2), 367–405.
- Chow, J.Y. L., Colagiuri, B., & Livesey, E. J. (2019). Bridging the divide between causal illusions in the laboratory and the real world: The effects of outcome density with a variable continuous outcome. *Cognitive Research: Principles and Implications*, 4(1), 1–15.
- Freckleton, I. (2012). Death by homeopathy: Issues for civil, criminal and coronial law and for health service policy. *Journal of Law and Medicine*, 19(3), 454–478.
- Garety, P.A., & Freeman, D. (2013). The past and future of delusions research: From the inexplicable to the treatable. *British Journal of Psychiatry*, 203(5), 327–333.
- Garety, P. A., Joyce, E., Jolley, S., Emsley, R., Waller, H., Kuipers, E., & Freeman, D. (2013). Neuropsychological functioning and jumping to conclusions in delusions. *Schizophrenia Research*, 150, 570–574.
- Hannah, S. D., & Beneteau, J. L. (2009). Just tell me what to do: Bringing back experimenter control in active contingency tasks with the command-performance procedure and finding cue density effects along the way. *Canadian Journal of Experimental Psychology*, 63(1), 59–73.
- Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32(2), 311–328.
- Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2014). Why ineffective psychotherapies appear to work: A taxonomy of causes of spurious therapeutic effectiveness. *Perspectives on Psychological Science*, 9, 355–387.
- Lim, A., Cranswick, N., & South, M. (2010). Adverse events associated with the use of complementary and alternative medicine in children. *Archives of Disease in Childhood*, 96(3), 297–300.
- Matute, H. (1996). Illusion of control: Detecting response-outcome independence in analytic but not in naturalistic conditions. *Psychological Science*, 7, 289–293.
- Matute, H., Blanco, F., & Díaz-Lago, M. (2019). Learning mechanisms underlying accurate and biased contingency judgments. *Journal of Experimental Psychology: Animal Learning and Cognition*, 45(4), 373–389.
- Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A., & Barbería, I. (2015). Illusions of causality: How they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology*, 6, 888.
- Matute, H., Yarritu, I., & Vadillo, M. A. (2011). Illusions of causality at the heart of pseudoscience. *British Journal of Psychology*, 102, 392–405.
- MacFarlane, D., Hurlstone, M. J., & Ecker, U. K. H. (2018). Reducing demand for ineffective health remedies: Overcoming the illusion of causality. *Psychology & Health*, 33, 1472–1489.
- Moore, M. T., & Fresco, D. M. (2012). Depressive realism: A meta-analytic review. *Clinical Psychology Review*, 32(6), 496–509.
- Moreno-Fernández, M.M., Blanco, F., & Matute, H. (2017). Causal illusions in children when the outcome is frequent. *PLoS ONE* 12(9): e0184707.
- Moreno-Fernández, M.M., Blanco, F., & Matute, H. (2021). The tendency to stop collecting information is linked to illusions of causality. *Scientific Reports*, 11, 3942.
- Moreno-Fernández, M. M., & Matute, H. (2020). Biased sampling and causal estimation of health-related information: Laboratory-based experimental research. *Journal of Medical Internet Research*, 22(7), e17502.
- Msetfi, R. M., Murphy, R. A., Simpson, J., & Kornbrot, D. E. (2005). Depressive realism and outcome density bias in contingency judgments: The effect of the context and intertrial interval. *Journal of Experimental Psychology: General*, 134(1), 10–22.

- Murphy, R. A., Schmeer, S., Vallée-Tourangeau, F., Mondragón, E., & Hilton, D. (2011). Making the illusory correlation effect appear and then disappear: The effects of increased learning. *Quarterly Journal of Experimental Psychology*, 64(1), 24–40.
- Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review*, 14(4), 577–596.
- Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, 66(I), 1–5.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.
- Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 21, pp. 229–261). San Diego, CA: Academic Press.
- Speechley, W., Whitman, J. C., & Woodward, T. S. (2010). The contribution of hypersalience to the “jumping to conclusions” bias associated with delusions in schizophrenia. *Journal of Psychiatry and Neuroscience*, 35(1), 7–17.
- Torres, M., Barberia, I., & Rodríguez-Ferreiro, J. (2020). Causal illusion as a cognitive substrate of pseudoscientific beliefs. *British Journal of Psychology*, 111(4), 840–852.
- US Food and Drug Administration (2017). FDA proposes new, risk-based enforcement priorities to protect consumers from potentially harmful, unproven homeopathic drugs [News Release]. Silver Spring, MD: Author Retrieved from [www.fda.gov/newsevents/newsroom/pressannouncements/ucm589243.htm](http://www.fda.gov/newsevents/newsroom/pressannouncements/ucm589243.htm)
- Vadillo, M. A., Musca, S. C., Blanco, F., & Matute, H. (2011). Contrasting cue-density effects in causal and prediction judgments. *Psychonomic Bulletin & Review*, 18, 110–115.
- Wasserman, E. A. (1990). Detecting response-outcome relations: Toward an understanding of the causal texture of the environment. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 27–82). San Diego, CA: Academic Press.
- Wasserman, E. A., Dorner, W. W., & Kao, S. (1990). Contributions of specific cell information to judgments of interevent contingency. *Cognition*, 16(3), 509–521.
- Willett, C. L., & Rottman, B. M. (2021). The accuracy of causal learning over long timeframes: An ecological momentary experiment approach. *Cognitive Science*, 45, e12985.
- Yarritu, I., Matute, H., & Vadillo, M. A. (2014). Illusion of control: The role of personal involvement. *Experimental Psychology*, 61, 38–47.

## APPENDIX

### Materials for a causality bias experiment

The materials can be downloaded from <https://osf.io/9et48/>. The experiment can be conducted by using a computer or by making a paper booklet (each trial is a page), whatever is more convenient for you.

*The instructions* are presented on the first page or the first screen. You can use the provided image, or create your own presentation by using these instructions as a base:

Welcome to the Garden game!

In this garden we have a new fertilizer that might improve strawberries' growth, but we still need to test it to be sure. You are going to see the tests that the garden employee is conducting. When you finish, you will have to answer some questions.

The training trials present information about the state of the potential cause (present/absent) and the effect (present/absent). The four possible combinations are presented as four separate images named a, b, c, and d following the notation presented in Table 7.1. You can prepare your own stimuli, just ensure that the state of the cause and effect on each type of trial is as follows:

- a. Cause present/effect present
- b. Cause present/effect absent
- c. Cause absent/effect present
- d. Cause absent /effect absent

The test is presented in one single screen or page after the training trials. You can use the image *Test* to prepare this last screen or card, or you can create your own by using these instructions:

The garden employee has finished testing.

Now please rate how effective the product is to make the strawberries grow by using a number from 0 (not effective at all) to 100 (totally effective).

## **Building the experiment**

If you have decided to use a computer, we recommend using a slide show presentation program to prepare the task and asking participants to run the presentation in full screen mode. Alternatively, you can build a paper booklet. The important point is that you need to show all the training trials plus the final test trial on a separate screen or page. Follow the next steps to create the high and low outcome-density versions of the task:

1. Copy and paste the image *Instructions* as the first image of the presentation.
2. Copy and paste the images corresponding to the four types of training trials as many times as indicated in Table 7.2.
3. Randomize the order in which training trials are presented.
4. Finally, copy and paste the final test trial.

## 8 Illusions of control

*Suzanne C. Thompson*

One of the enduring themes of psychological theory and research is that human beings are motivated to believe that they have control over the events of their lives (Bandura et al., 2003; Rodin, 1986; White, 1959). Extensive research has demonstrated that perceived control and a sense of self-efficacy are 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 an 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 8.1). 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 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.

### **Text Box 8.1 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 here 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 or the passenger of the car (see Q1 and Q2 in Table 8.1). 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.

*Table 8.1 Questions used to demonstrate illusory control over driving*

- 
- Q1. Compared to other drivers, how likely do you think you are of being involved in an automobile accident when you are driving?
- Q2. Compared to other drivers, how likely do you think you are of being 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 which 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 which 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?
- 

Source: Adapted from McKenna (1993).

Note: Response scale from -5 (much less likely) through 0 (average) to 5 (much more likely).

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).

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 21), 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 rating 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 versus 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 et al., 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 classtime 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 classtime. The analysis is then done with reinforcement level (25% versus 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?

## **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.

### ***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 tasks with skill-associated features lead people to overestimate personal control (see Ayeroff & Abelson, 1976; Dunn & Wilson, 1990; Langer, 1975, for representative studies). However, recent research has found that at least one skill-related factor, making a choice, is not usually associated with illusory control. Klusowski et al. (2021) conducted 17 experiments involving a total of over 10,000 participants who were randomly assigned to a choice or no-choice condition. Other independent variables

such as illusory versus actual control were also manipulated in some of the studies. Only rarely did choice cause an illusion of control and those instances could be attributed to pre-existing beliefs rather than illusory control. The researchers concluded that there is no evidence that choice causes an illusion of control.

This set of studies casts doubt on the idea that the study design used by Langer and others in choice manipulation studies was measuring illusory control. Comprehensive research similar to Klusowski et al. (2021) that manipulates other skill-related factors and that measures, rather than infers, control judgments is needed to determine if skill-factors are or are not a contributor to illusory control.

### ***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 versus failure. An emphasis on success enhances illusions of control whereas failure emphasis undermines control illusions. 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.

### ***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 et al. (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 et al., 2009) examined the desire to avoid an aversive events 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 versus 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 et al. (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.

A more desirable outcome may lead to increased illusory control by increasing efforts to get the outcome. Benvenutia et al. (2018) manipulated the reward in a noncontingent-reinforcement task using points for winning that could be used to pay for photocopying (high reward) or did not have monetary value (low reward). Those in the high-reward condition responded more on the task and judged their control as higher. Thus across a variety of desirable outcomes, higher desire or need for the outcome increases illusory control.

### ***Mood***

A number of studies found that illusions of control are higher when people are in a positive mood. For instance, Alloy et al. (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 temporary more negative mood showed lower illusions of control.

### ***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).

### ***Power***

The effect of power on illusory control has also been examined. A series of experiments manipulated power through either assigned roles or priming past experience in a high or low power position (Fast et al., 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. Even obviously useless power can be desirable. Participants in the Sloof and von Siemens (2017) study were willing to pay more to be the decision-maker for an outcome that was explicitly described as determined by chance. The researchers concluded that being in the authority position has intrinsic value.

### ***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 could explain why the presence of skill-based elements such as choice, familiarity, involvement, and competition lead people to overestimate their control on chance-based tasks. However, 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. 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).

In the Benvenutia et al. (2018) study described earlier, more valuable points led to higher rates of responding and higher illusions of control. In this case, the stronger motive to get the valued points led to more responses which would strengthen the connection between action and outcomes, thereby increasing the perception of control. In contrast, these results are not easily explained by the skills-chance confusion theory as the manipulation of point value would not change the perception of skill involved in the task.

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 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. In addition, as described above, Kluosowski et al. (2021) found no support for the idea that choice, one factor that purportedly increases the confusion of chance and skill, leads to illusions of control.

### **Implications of illusions of control**

Most of the research in this area investigated 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.

Changes in illusory control levels also have effects on emotions. In Kaufman et al. (2019) study, illusions of control were manipulated through a light-onset-frequency task. Increasing the frequency of the desired outcome increased illusory control and reduced negative affect but did not change positive affect. Decreasing outcome frequency reduced illusory control and positive affect but did not change negative affect. Gaining or losing control, even if illusory, has emotional implications in complex but expected directions in that gaining illusory control is positive (i.e., reduces negative affect) and losing illusory control is negative (i.e., reduces positive affect).

Another effect of control illusions on persistence in the face of obstacles was examined in studies of the effects of illusory control in two circumstances that increase the difficulty of achieving success – diminishing returns and competition with a stronger opponent (Studer et al., 2020). First, illusions of control were manipulated through a light-onset task, then persistence was measured in a situation that either had a progressively decreasing reward or a more successful competitor. Those in the induced illusions of control condition persisted significantly longer in both of these difficult situations.

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.

Thus illusory control is associated with the positive outcomes of more positive emotions, less discouragement and more persistence in the face of adversity, lower depression, and an inclination to action. 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. Chapters 18 and 21).

### ***Negative consequences of control illusions***

In contrast to these positive finding, there is evidence that overestimating one's control has a number of costs and disadvantages in areas as diverse as childcare, driving behavior, protective health behaviors, financial investment, and problem gambling. Donovan et al. (1990) investigated the influence of illusory control on performance demands associated with childcare. 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 et al., 1999).

Fenton-O'Creevy et al. (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 prompts 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 et al., 2014; Toneatto et al., 1997). Furthermore, gamblers who exhibit higher illusory control also interpret their losses in a way that protects their perception of control. Cowley et al. (2015) found that high-illusory-control gamblers focused on their highest win to evaluate a losing gambling session. In contrast, gamblers with low illusory control used a more valid indicator of success in their evaluation, the final outcome of the entire session.

There are a number of characteristics of gambling situations that can help foster an illusion of control, including familiarity, and desirability of the outcome. The gambling industry helps enhance the perception of control through advertising slogans ("You can't win if you don't play"), the design of games that incorporate more behavioral involvement such as multiline slot machine games that allow gamblers to choose the number of paylines (Harrigan et al., 2014), and video-lottery terminal-stopping devices (Ladouceur & Sevigny, 2005). These designs give the illusion of control by emphasizing winning and strengthening the connection between action and outcome, 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 actual 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, Moritz et al. (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 et al. (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, functional magnetic resonance imagery (fMRI) scans were used to examine the correlates of illusory control. Two unexpected, but interesting, results were found. One result was cortico-striatal 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 and is supported by the control heuristic proposition that illusory control is based on both intention and connection. Saliency of outcomes on choice trials is likely to increase the perception of connection between one’s actions and the outcome. When perceived reinforcement rates are higher, the connection between one’s choice and desired outcome appears stronger and feeds into more illusory control.

Neural response at the reveal of the outcome was also found in a neural imaging study that examined neural responses in the cortico-striatal network associated with the brain reward system (Lorenz et al., 2015). Adolescents played a slot-machine game that involved deciding when to stop the spinning wheels. Although no control over getting a winning screen was possible, about a quarter of participants judged that they had some control. This group had strong striatal activity during the reward anticipation phase, compared to the group who reported no control. Thus control, even if illusory, can have inherent reward value. These reinforcing properties of illusions of control help explain why games of chance are appealing despite the long-term likelihood of loss.

## 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.
- Bandura, A., Caprara, G. V., Barbaranelli, C., Gerbino, G., & Pastorelli, C. (2003). Role of affective self-regulatory efficacy in diverse spheres of psychosocial functioning. *Child Development*, *74*, 769–782.
- Benvenuti, M. F. L., de Toledo, T. F. N., Simoes, R. A. G., & Bizarro, L. (2018). Comparing illusion of control and superstitious behavior: Rate of responding influences judgments of control. *Learning and Motivation*, *64*, 27–33.
- 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.
- Cowley, E., Briley, D. A., & Farrell, C. (2015). How do gamblers maintain an illusion of control? *Journal of Business Research*, *68*, 2181–2188.
- 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.
- 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.
- Kaufman, M., Goetz, T., Lipnevich, A. A., & Pekrun, R. (2019). Do positive illusions of control foster happiness? *Emotion*, 19, 1014–1022.
- Klusowski, J., Small, D. A., & Simmons, J. P. (2021). Does choice cause an illusion of control? *Psychological Science*, 32, 1–14.
- 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.
- Lorenz, R. C., Gleich, T., Kuhn, S., Pohland, L., Pelz, P., Wustenberg, T., Raufelder, D., Heinz, A., & Beck, A. (2015). Subjective illusion of control modulates striatal reward anticipation in adolescence. *NeuroImage*, 117, 250–257.
- 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.
- Sloof, R., & von Siemens, F. A. (2017). Illusion of control and the pursuit of authority. *Experimental Economics*, 20, 556–573.
- Studer, B., Geniole, S. N., Becker, M. L., Eisenegger, C., & Knecht, S. (2020). Inducing illusory control ensures persistence when rewards fade and other outperform us. *Psychonomic Bulletin & Review*, 27, 809–818.
- 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., & Tsanis, 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.

## 9 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 focused, 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, 2019). In the context of this book, it is also worth noting that Peter Wason was an early leader in emphasizing 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, 2019).

### **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 behavior, 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 which classified triples of three whole numbers. An example which 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 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, 2016; 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 behavior as fixated and obsessive, viewing the verbal formulations as rationalizations 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 it 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 if you turn over the 3 and find an A or find a B it would 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 focused on trying to find the disconfirming card that has an A and *not* a 3.

Of course, not everyone gets the task wrong and under some experimental conditions, people show apparent insight into the need to falsify the statement. A model based on the idea of degrees of insight was first proposed by Johnson-Laird and Wason (1970): they distinguished between no insight (A or A and 3 chosen) partial insight (A, 3, and not 3), and full insight (A and not 3). With partial insight, people see the need to falsify the statement but still try to verify it as well. A recent meta-analysis of large numbers of selection-task experiments showed an apparently good fit to this model using multinomial modeling (Ragni et al., 2018). This study analyzed the data from 228 selection task experiments in which the relevant combination of card selections were reported and compared the insight model to the inference-guessing model of Klauer et al. (2007), claiming a better fit for the former, a conclusion challenged by Kellen and Klauer (2020).

The account of the selection task in terms of verification and falsification was actually challenged very early on by some experiments of my own with an alternative account in terms of “matching bias”. This led to a collaboration with Peter Wason and the development of the dual-process theory of reasoning which had much influence on later work in the field. I tell these stories next.

### ***Matching bias***

Whilst plausible, the confirmation bias account was actually abandoned by Wason 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 favor 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 pre-conscious 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 et al., 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. In line with the insight model, they offer explanations in terms of either verifying or falsifying the statement, consistent with their choices. But in doing so 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 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 favored 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, rationalization, 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 already mentioned collaboration with myself (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. It seems implausible that adding a negative gives people insight into falsification which is absent when the statement is affirmative. Hence, we suggested instead that matching bias reflected an unconscious Type 1 process and that conscious Type 2 thinking was used to provide a *rationalization* of an unconsciously cued choice.

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 rationalizing 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 (see Evans, 2019). The general view is that people do not intervene as much as they should and are by nature “cognitive misers” (Kahneman, 2011; Stanovich, 2011, 2018). Hence, people may routinely seek to justify an intuitive answer but do so in a cursory manner and be 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 et al., 2011). In such cases, people spend little time thinking about their answers. A later paper showed that matching bias has precisely this quality (Thompson et al., 2013). That is, matching cards *feel right* and are often chosen quickly as a result.

Dual-process theory, which has multiple origins, has become a huge field of research in cognitive and social psychology and is also quite controversial, with several major critiques published in recent years (e.g. Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011). Keith Stanovich and I have presented a robust defense to these criticisms (Evans & Stanovich, 2013).

## 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 10). 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	Drinking	22 years	16 years
Beer	coke	of age	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, if 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 fulfill a condition (be 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.

### **Text box 9.1 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 neighbors 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 analyze this is with a series of  $2 \times 2$  Chi square 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 Square will show us whether this difference is significant. Repeat the test for the other three cases.

An example is Condition A in the classroom demonstration shown in Text box 9.1. This was adapted from an experiment reported by Manktelow and Over (1991). The italicized 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. This is not a rule for 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 italicized 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 £100 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 (e.g., 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 which are not normally considered the logical choices. They nevertheless appear to make appropriate choices given the goals which 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 particularly, 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 being 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), 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 et al., 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 criticized, 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 et al., 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 10). 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 favor 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 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 9.1). 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 center-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 program which is focused on individual differences in reasoning and decision-making. He has written several books on the topic (e.g. Stanovich, 2009, 2011; Stanovich et al., 2016) in which 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. He insists that conventional IQ tests are inadequate for assessing rational thought and suggests that they should be supplemented by tests of rational thinking dispositions. He has argued that we are “cognitive misers” who often fail to apply our intelligence or have failed to acquire the necessary knowledge for rational reasoning (Stanovich, 2018). His views contrast with those of authors who

have 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 program of work on judgment 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.
- Wason also contributed significantly to early work on dual-process theories of human reasoning, which has become a major field of study to the present day.
- 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 which 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). *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ: Erlbaum.
- 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. (2016). Reasoning, biases and dual processes: The lasting impact of Wason (1960). *Quarterly Journal of Experimental Psychology*, 69, 2076–2092.
- Evans, J. St. B. T. (2019). Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, 25, 383–415.
- 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, UK: Psychology Press.
- Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford: Oxford University Press.
- 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.
- 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.
- Johnson-Laird, P. N., & Wason, P. C. (1970). Insight into a logical relation. *Quarterly Journal of Experimental Psychology*, 22, 49–61.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus & Giroux.
- Kellen, D., & Klauer, K. C. (2020). Theories of the Wason selection task: A critical assessment of boundaries and benchmarks. *Computational Brain & Behavior*, 3, 341–353.
- Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, 4, 533–550.
- Klauer, K. C., Stahl, C., & Erdfelder, E. (2007). The abstract selection task: New data and an almost comprehensive model. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 33(4), 680–703.
- Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review*, 94, 211–228.
- Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberative judgements are based on common principles. *Psychological Review*, 118, 97–109.
- 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.
- Ragni, M., Kola, I., & Johnson-Laird, P. N. (2018). On selecting evidence to test hypotheses: A theory of selection tasks. *Psychological Bulletin*, 144, 779–796.
- 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. (2009). *What intelligence tests miss*. London: Yale University Press.
- Stanovich, K. E. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.
- Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, 24, 423–444.
- Stanovich, K. E., West, C., & Toplak, M. E. (2016). *The Rationality Quotient: Towards a test of rational thinking*. Cambridge, MA: MIT 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, 431–452.
- Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, 63, 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.

# 10 Belief bias in deductive reasoning

*Jonathan St. B. T. Evans, Linden J. Ball, and Valerie A. Thompson*

In Chapter 9, work was described on the Wason selection task which has been intensively studied by reasoning researchers and provided evidence of a number of biases. In this chapter, we 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 contextualize 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 we 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 we 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 a 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) that 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 they may realize that there are other possibilities. When they try to integrate the information in the second statement, they ought to reject this model as 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 9).

The study of reasoning biases is, however, of considerable psychological interest in its own right. In this chapter, we 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. We 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 we 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 we 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 et al. (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 with the premises, 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 9. They suggest very superficial processing by most of the participants. They also suggest that 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 10.1). Some of the conclusions are logically valid deductions and some are not and some have believable conclusions and some do not. A well-known study by Evans et al. (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

*Table 10.1* 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%

Invalid-Unbelievable (IU). Examples of these are shown in Table 10.1 together with the rates at which people accepted them as valid arguments over three experiments.

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 Ball & Thompson, 2018; Evans et al., 1993; Klauer et al., 2000; Newstead et al., 1992).

### **Text box 10.1 Classroom demonstration of belief bias in syllogistic reasoning**

#### **Participants**

The effects are typically quite large and have been demonstrated with small samples. We 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 10.1. 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 10.1 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 randomized order.

## Procedure

A booklet with eight pages is given 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 is 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:

$$\begin{aligned}\text{Logic Index: } & (\text{VB} + \text{VU}) - (\text{IB} + \text{IU}) \\ \text{Belief Index: } & (\text{VB} + \text{IB}) - (\text{VU} + \text{IU}) \\ \text{Interaction Index: } & (\text{VU} + \text{IB}) - (\text{VB} + \text{IU})\end{aligned}$$

The first two are fairly obvious. The Logic Index is the number of valid conclusions accepted minus the number of invalid conclusions 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  $(\text{IB} - \text{IU}) - (\text{VB} - \text{VU})$  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.

In the literature, a lot of interest has been focused 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 conclusions are 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 which 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 et al., 1989). The Selective Scrutiny Model is supported by evidence that, when people are forced to respond rapidly, the belief bias increases and the belief–logic interaction disappears (Evans & Curtis-Holmes, 2005), presumably because they do not have time to search for counterexamples to the unbelievable conclusion.

### **Dual processes and belief bias**

Dual-process theories offer a similar explanation. Recall that this theory posits that people have recourse to two qualitatively different processes: Fast, autonomous (Type 1) processes and slower, working-memory-demanding Type 2 processes (see Chapter 9). On this view, we can suppose that there are two kinds of belief bias: A Type 1 belief bias is a response bias which should favor 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 (Ball et al., 2006; Thompson et al., 2011). Other authors, however, have interpreted response latencies as consistent with the Selective Scrutiny Model if individual differences in reasoning ability are taken into account when examining participants' latency profiles (Stupple et al., 2011). A slightly different account is the Selective Processing Model (Evans et al., 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.

Claims have also been made (e.g., Dube et al., 2010) that belief-bias effects in syllogistic reasoning, including the interaction, can be accounted for as a pure response bias when analyzed using the methods of signal detection theory. 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 et al., 2013). Indeed, Trippas et al. (2014) provide compelling support based on analyses using signal detection theory in favor of a dual-process view of belief-bias effects. In their studies, Trippas et al. presented participants with *pairs* of syllogisms and asked them to choose which of the two conclusions was valid. The crucial twist in these studies was that for some trials the believability of the two presented conclusions was *equated* (i.e., they were both unbelievable or were both believable), which controlled for the response-bias component of belief bias. For these trials, however, there was still evidence for the influence of Type 2 motivated reasoning in determining decisions. In related work, Trippas et al. (2015) showed that it is the *disposition* to think analytically that is the best predictor of motivated reasoning effects in engendering Type 2 belief bias rather than cognitive ability (e.g., IQ or other measures of cognitive capacity).

Although the findings reported by Trippas et al. (2014, 2015) seem to provide convincing support for a dual-process account of belief bias, the issue continues to be contentious (e.g., see Stephens et al., 2019, for a recent defense of a pure response-bias account). Overall, however, our view remains that the balance of evidence supports the involvement of response biases (Type 1 processes) *and* motivated reasoning (Type 2 processes) in determining belief-bias effects in syllogistic reasoning. This view is also supported by neural-imaging research on belief-bias effects with syllogisms, which has provided good evidence for Type 1 or Type 2 responses being associated with activation in distinct brain regions (e.g., Luo et al., 2013; Tsujii & Watanabee, 2009). Moreover, this view has gained further credence by the proliferation of recent studies of belief bias in syllogistic reasoning that have drawn on dual-process ideas, including studies that have revealed some intriguing findings about the nature of Type 1 and Type 2 processing. We now turn to consider some of these recent findings as they have provoked some important new developments in our understanding of the involvement of dual processes in deductive reasoning.

### *Logical intuitions*

One finding that has sparked considerable interest is the extreme sensitivity that people seem to display to syllogisms in which the logical status of the conclusion conflicts with its belief status. Examining Table 10.1 shows that there are two problems of this type, termed “logic/belief conflict” problems: the Valid-Unbelievable item and the Invalid-Believable item. People’s sensitivity to such problems compared to “no-conflict” problems is revealed in multiple ways, including increased response latencies (e.g., De Neys & Glumicic, 2008; Stupple et al., 2011), heightened autonomic arousal as determined by galvanic skin conductance measures (De Neys et al., 2010), and reduced confidence in responses (De Neys et al., 2011; Thompson et al., 2011). Moreover, this conflict-detection sensitivity is even found in reasoners who respond primarily in accordance with Type 1 beliefs and who do not engage in any motivated Type 2 reasoning (De Neys, 2012, 2014; Stupple et al., 2011).

This latter finding is particularly curious: How can it be that those who are not engaging in slower, Type 2 reasoning still seem to have some awareness of the logical status of the conclusion in terms of its validity or invalidity? In other words, Type 1 response bias should always be completed before an attempt at Type 2 reasoning is initiated, so Type 2 processes should not be able to interfere with Type 1 processes. It is not just the case that the problems that have been presented in such studies have simple and readily discernible logical forms; evidence for people’s sensitivity to the logical status of conclusions with belief/logic conflict problems has also been obtained with moderately complex syllogisms (Trippas et al., 2017) and even with highly complex syllogisms that are responded to with very short response deadlines (Newman et al., 2017), attesting to the robustness of this rapidly occurring conflict-detection effect.

An equally curious finding established by Handley et al. (2011), albeit not with syllogisms but with conditional reasoning problems (discussed below), is that if people are explicitly instructed to respond to conclusions purely in terms of whether they are believable or unbelievable (a reversal of the typical belief-bias paradigm), then the logical status of conclusions is now observed to disrupt effective belief-based responding on logic/belief conflict items (see also Trippas et al., 2017). This finding is important as it suggests that the logical status of a conclusion is somehow determined fast enough to interfere with a normally rapid and straightforward Type 1 belief-based response.

How might the observations that logic/belief conflicts disrupt both logic-based and belief-based responding be explained from a dual-process perspective? A dominant account is that people’s fast, Type 1 processing can be driven by “logical intuitions” (De Neys, 2012, 2014; De Neys & Pennycook, 2019), with the idea here being that people can quickly apprehend the logical status of an argument’s conclusion using implicit, intuitive processes (see Morsanyi & Handley, 2012, for the earliest evidence of people’s apparent intuitive detection of the logic of syllogisms; for further evidence of this effect see Trippas et al., 2016). More recent findings, however, suggest that the concept of logical intuitions may be something of a misnomer. Rather than people possessing and applying a fast intuitive logic, it seems more psychologically plausible to assume that,

when responding either under logic or belief instructions, people are sensitive to fairly superficial structural cues in the presented problems that *correlate* with the logical status of conclusions (cf. Klauer & Singmann, 2013). The point is that deductive problems have highly formal structures that are likely to trigger a “feeling of rightness” (Ackerman & Thompson, 2017) regarding the validity or invalidity of a given conclusion, thereby influencing responding.

This latter point is not to deny the influence on responding that can arise from the surface features of problems that are correlated with the logical status of conclusion, but is rather to clarify that the basis of the observed effects may reside in available cues in the problem rather than in some kind of deep-seated logical understanding. That said, emerging evidence does suggest that it is people of higher cognitive ability who are best able to pick up on these cues. For example, Thompson et al. (2018) have shown that when reasoners are asked to respond on the basis of belief, then the interference that is caused by the logical status of conclusions is most marked in high-ability reasoners. This suggests that for this group of reasoners a logic-based response is their default response rather than a belief-based one. Lower ability people, on the other hand, show the reverse pattern, where conclusion believability interferes with judgments of validity (Thompson et al., 2018).

A related finding comes from the use of the “two-response paradigm”, in which people are required to give a fast intuitive response and are subsequently given time to deliberate and generate a revised response, if they wish (Thompson et al., 2011). This paradigm has revealed that those who get the answer to deductive reasoning problems correct do so at Time 1, rather than Time 2 (Bago & De Neys, 2017; Thompson et al., 2011). Particularly pertinent to the present discussion, however, is evidence that this tendency is most marked for high IQ people (Raoelison et al., 2020). This finding reinforces the fact that such individuals are more likely to “intuit” the correct answer, again presumably through their ability to pick up on structural cues in presented problems that align with their logical status.

In sum, the emerging picture with respect to syllogistic reasoning is of an intricate interplay between more rapid Type 1 processes and slower Type 2 processes that can bias responses to presented problems in predictable ways under both standard logic instructions and pragmatic instructions to determine the believability of conclusions. Such evidence paves the way toward the development of the next generation of dual-process theories of deduction which will have both the rigor and flexibility to explain a wide range of reasoning phenomena. Having demonstrated that belief bias is a complicated but nevertheless robust effect in the context of syllogistic reasoning, we 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. We will illustrate this first with letter-number rules, similar to those used for the abstract version of the Wason selection

task (Chapter 9). 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 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), which is consistent with the overall pattern of responding and it is likely that several cognitive biases could operate on particular inferences. As with the matching bias (Chapter 9), 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 9.

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, we 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* which 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 9). For example, it must be true that “If Boris Johnson is president of the USA, then he will declare war on Russia”, simply because Boris Johnson 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 favored 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 we 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 behavior of the Democrat 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  $P(\text{if } p \text{ then } q) = P(q | p)$ , often referred to as the Equation. There is now considerable evidence that people’s actual beliefs about conditional statements do conform to the Equation (e.g., Over et al., 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 – which 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 et al., 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 which 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 9) 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:* 17-year-olds should have the legal right to drink alcoholic beverages.

*Dale's premise or justification for belief:* 17-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:* 17-year-olds are three times more likely to be involved in an automobile accident while under the influence of alcohol than 19-year-olds (assume statement factually correct).

*Dale's rebuttal to critics' counter-argument:* 17-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 critic's 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 favored a conclusion with which they agreed. However, the effect was much stronger in older than 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.

Indeed, as described above, recent evidence suggests that high-IQ people may form a default response that is based on logic; the logical structure of arguments interferes with their ability to make judgments based on beliefs (Thompson et al., 2018). The accumulating evidence that we discussed about the role of "logical intuitions" in reasoning has challenged modern researchers, and it is not yet clear how to reconcile these findings with the broad base of findings that support the Selective Processing Model. As well, 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, that is, by weighing degrees of belief against the strength of the evidence or the argument they have been presented. The Thompson and Evans (2012) paper suggested that they do, but there is much to be learned about how well theories derived from formal paradigms can explain more everyday forms of inferences.

## 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.
- There have been numerous challenges to this perspective. Although we believe that the broad base of evidence supports this interpretation, it is less clear how to reconcile the dual-process view with the emergent evidence on logical intuitions.
- 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 deductive reasoning is provided by Ball and Thompson (2018). For a less technical and broader treatment of the role of knowledge and belief in reasoning, see Evans (2007, Chapter 4).

## References

- Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, 21(8), 607–617.
- Bago, B., & De Neys, W. (2017). Fast logic? Examining the time course assumption of dual process theory. *Cognition*, 158, 90–109.
- Ball, L. J., Phillips, P., Wade, C. N., & Quayle, J. D. (2006). Effects of belief and logic on syllogistic reasoning: Eye-movement evidence for selective processing models. *Experimental Psychology*, 53, 77–86.
- Ball, L. J., & Thompson, V. A. (2018). Belief bias and reasoning. In L. J. Ball & V. A. Thompson (Eds.), *The Routledge international handbook of thinking and reasoning* (pp. 16–36). Abingdon, Oxon: Routledge.
- Byrne, R. M. J. (1991). Can valid inferences be suppressed? *Cognition*, 39, 71–78.
- De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, 7, 28–38.
- De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning*, 20, 169–187.
- De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE*, 6, e15954, 1–10.
- De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, 106(3), 1248–1299.
- De Neys, W., Moyens, E., & Vansteenwegen, D. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, Affective, and Behavioral Neuroscience*, 10(2), 208–216.
- De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science*, 28(5), 503–509.
- 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.
- Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: A test of the default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 28–43.
- Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove, UK, and 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.
- Klauer, K. C., & Singmann, H. (2013). Does logic feel good? Testing for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 1265–1273.
- Luo, J., Liu, X., Stupple, E. J., Zhang, E., Xiao, X., Jia, L., et al. (2013). Cognitive control in belief-laden reasoning during conclusion processing: An ERP study. *International Journal of Psychology*, 48(3), 224–231.
- Manktelow, K. I. (2012). *Thinking and reasoning*. Hove, UK: Psychology Press.
- Morsanyi, K., & Handley, S. J. (2012). Logic feels so good – I like it! Evidence for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 596–616.
- Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(7), 1154–1170.
- 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.
- Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than analytic thinking. *Cognition*, 104, 381, 1–14.
- Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, 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.

- Stephens, R. G., Dunn, J. C., & Hayes, B. K. (2019). Belief bias is response bias: Evidence from a two-step signal detection model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(2), 320–332.
- 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.
- Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. St. B. T. (2018). Do smart people have better intuitions? *Journal of Experimental Psychology: General*, 147, 945–961.
- Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, 63, 107–140.
- 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.
- Trippas, D., Handley, S. J., Verde, M. F., & Morsanyi, K. (2016). Logic brightens my day: Evidence for implicit sensitivity to logical validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42, 1448–1457.
- Trippas, D., Pennycook, G., Verde, M. F., & Handley, S. J. (2015). Better but still biased: Analytic cognitive style and belief bias. *Thinking & Reasoning*, 21(4), 431–445.
- Trippas, D., Thompson, V. A., & Handley, S. J. (2017). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. *Memory & Cognition*, 45(4), 539–552.
- Trippas, D., Verde, M. F., & Handley, S. J. (2014). Using forced choice to test belief bias in syllogistic reasoning. *Cognition*, 133(3), 586–600.
- Tsuji, T., & Watanabe, S. (2009). Neural correlates of dual-task effect on belief-bias syllogistic reasoning: A near-infrared spectroscopy study. *Brain Research*, 1287, 118–125.

## **Part II**

# **Judgment**



Taylor & Francis  
Taylor & Francis Group  
<http://taylorandfrancis.com>

# 11 Availability

*Anine Riege and 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 their own contribution so that the sum exceeds 100%. People normally overestimate their own contribution to the joint product of a group. Ross and Sicoly (1979) found this effect in naturally occurring discussion groups, basketball players, groups assembled in the laboratory, and married couples. Why do individuals overestimate their contribution to a joint product? One explanation would be that they are motivated to see themselves in a positive light. It is, however, possible that cognitive processes alone, without 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 relevant information. 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 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, which takes about the same time and effort. Even if he remembers the paella, his memories of his own efforts expended on 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 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 12 in this volume) and anchoring and adjustment (see Chapter 13). Text box 11.1 provides a definition of availability.

### **Text box 11.1 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 with 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 these authors used the term availability differently from the way Tversky and Kahneman (1973) used it. In this chapter, we use the term “availability” as a general heuristic relying on ease or amount of recall, and the term *ease of recall* when discussing the specific mechanisms behind the availability heuristic. Later researchers used the term *retrieval fluency*, clarifying that ease of processing is the mechanism underlying the availability heuristic (e.g., Benjamin et al., 1998; Hertwig et al., 2008).

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, we first describe two of the experiments from the classical paper of Tversky and Kahneman (1973). We 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 discussing some studies that explore mental mechanisms behind availability.

## **Two basic experiments**

We start with summarizing two classical studies from Tversky and Kahneman’s (1973) seminal paper 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 11.2 for a classroom demonstration).

#### *Method*

Participants were presented with a tape-recorded list of 39 names, at a rate of two 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 even though non-famous names outnumbered less famous names. If 19 famous women and 20 less famous men were 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 were presented, participants were expected to judge that more men were shown.

### *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; only 13 recalled fewer famous names than less famous names. A sign test revealed that this difference was significant. Among the 99 participants who compared the frequency of men and women in the list, 80 erroneously believed that the gender consisting of the more famous names occurred more frequently. Again, a sign test revealed that this difference was significant. The authors concluded that the participants used the availability heuristic because they recalled more famous names and they judged famous names as occurring more frequently on the list.

### **Text box 11.2 Classroom demonstration of Experiment 1**

This is an easy experiment that always worked as a classroom demonstration; even with small sample sizes of 20 to 30 students, the result was numerically in the right direction. The design of the experiment can be simplified. Here is the recipe: Compile a list of nine famous women and ten less famous men, or vice versa. Update the list after a few years because some less famous people rise to stardom and some famous ones sink into oblivion. Present the list for about one minute – enough to read all names once. Then ask the participants whether there were more men or women in the list. The independent variable is fame, the dependent variable the estimated prevalence of female versus male names. Simply count how many participants chose the gender with the famous and the non-famous names, respectively, and assess the outcome by a sign test.

The idea behind the other part of the experiment – 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 and can be omitted in the classroom demonstration.

**Experiment 2: The letter-frequency experiment**

Another classical experiment instructs participants to judge letter frequency (Tversky & Kahneman, 1973, Exp. 3). A classroom adaptation of this experiment can be found in Text box 11.3.

*Method*

Participants of this study were given the following instructions (Tversky & Kahneman, 1973, pp. 211–212):

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, in our example 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 whether participants judged these letters to appear more frequently in the first position even though 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, the 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 the presented letters, and 47 judged the third position to be more likely for the letters. The authors employed a sign test and found a 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, even though each letter was more frequent in the third position.

**Text box 11.3 Classroom demonstration of Experiment 2**

For a classroom demonstration, you may choose an uneven number of consonants that in your language is more frequent in the third than in the first position. Ask the respondents to indicate for each of these letters – shown one by one – whether it is more frequent in the first position or in the third position. The independent variable is the objective letter position (first versus third; this manipulation was not

included in the original experiments), the dependent variable is the judgment of the perceived letter position (first versus third). Simply count how many respondents chose the first and how many chose the third position for the majority of letters. The original experiment should replicate that it is easier to recall letters at the first position and participants therefore overestimate the number of letters in the first position, compared to the third position. More participants may judge that there are more letters in the first position rather than the third position, but this result is unlikely to be significant with a sign test in small classes of less than 30.

We now turn to an overview on research on the availability heuristic that is partitioned into two sections: First, we review early research that applied the concept of availability before turning to more recent discussions of mechanisms underlying the availability heuristic. For a discussion of the difference between availability and representativeness, see Text box 11.4.

#### **Text box 11.4 What is the difference between availability and representativeness?**

Heuristics have been criticized for being unspecified processes where one heuristic can be said to explain a biased answer as easily as another heuristic (Gigerenzer, 1998). It is not easy to distinguish between the two heuristics because we cannot directly observe the underlying cognitive processes. The availability heuristic relies on the ease with which relevant instances of a class come to mind; the representativeness heuristic relies on the similarity of an instance to its class or category (Chapter 12 in this volume). In order to explore the difference between the use of the two heuristics, Braga et al. (2018) conducted a study in which they presented letter strings that either appeared to be a randomly generated string (e.g., HHTHTTHT) or to include a streak at the end (e.g., HHTTTTTT). In our example, people will typically predict that after six Ts in a row, an H will appear. This is the so-called gambler's fallacy that relies on the representativeness heuristic because the streak is not representative for a randomly generated letter string. This change from the last letter in the string to the predicted letter (from T to H) cannot be explained by the availability heuristic because the most available letter – the one that comes to mind most easily – is the last letter in the sequence, especially after a streak. In our example, participants using the representativeness heuristic would predict H, especially after seeing six Ts but participants using the availability heuristic would choose T, again especially after the streak of Ts.

Braga et al. (2018) exploited one characteristic difference between the two heuristics. As the representativeness heuristic is more abstract and includes more complex cognitive processes than the availability heuristic, they expected and found that participants use the representativeness heuristic only when they have sufficient time. In this case, participants switched the letter after a streak, committing the gambler's fallacy. However, under time pressure, participants more probably predicted that the last letter would appear again than when they had enough time; they used the availability heuristic. This result supported the notion that the two heuristics can be distinguished and that their underlying cognitive processes differ.

## **Applications of the availability heuristic**

The judgments in the examples above (letter frequency) are frequency judgments. Some studies show that people's ability to recall how often events occur is surprisingly good and differs from other memory processes in being largely automatic (e.g., Zacks & Hasher, 2002). Individuals usually have to make a deliberate effort to remember newly introduced people's names or to get what they need from the supermarket. However, for recalling frequencies, deliberate effort does not increase performance, nor is there much effect of training that typically improves performance on other memory tasks. So why are people's frequency judgments sometimes wrong?

To answer this question, we review first how biased encoding and retrieval influence availability of information, then vividness as a basis of availability of information, and 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, often 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 crimes 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 et al. (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.

Despite claims that we are rather good at tracking frequencies, we make mistakes in a systematic manner because of biased encoding and retrieval. Another way to think about this is Hogarth et al.'s (2015) concept of kind versus wicked environments. In kind environments, implicitly processed information leads to valid inferences, and thus correct judgments. Wicked environments, on the other hand, are environments where samples of experience are not representative, which in turn leads to incorrect judgments. Though people may process the data they see appropriately, they lack the metacognitive ability to correct for biases in the environment. For example, the choice of stimuli used in the original letter-experiment by Tversky and Kahneman (1973) might reveal the effects of a wicked environment because the letters K, L, N, R, and V differ from most other

consonants which occur more frequently in the first position. By contrast, the letters K, L, N, R, and V occur more frequently in the third position (Gigerenzer & Brighton, 2009). Indeed, when Sedlmeier et al. (1998) conducted the same experiment with German letters that were more representative of the distribution of letters in the German language, they did not obtain biased estimates. This does not invalidate the availability heuristic, but its use may be more likely to bias estimates in wicked environments. We reviewed evidence of biased retrieval of frequencies. Another source of availability is 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: 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. Alternatively, 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 et al. (1980) showed 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. For example, the pallid prosecution version was: "On his way out the door, Sanders [the defendant] staggered against a serving table, knocking a bowl to the floor." The vivid version read: "On his way out the door, Sanders staggered against a serving table, knocking a bowl of guacamole dip to the floor and splattering guacamole on the white shag carpet." The participants remembered more evidence of the prosecution and gave higher judgments of the defendant's 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 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 the 20 years after the US 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 probably retrieve more instances and 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 they have higher chances than their classmates to experience positive events and lower chances to experience negative events (cf. Chapter 8 on the illusion of control). Of course, the average 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 who overestimated their share of the housework (Ross & Sicoly, 1979). Individuals see their own actions that increase the probabilities of positive outcomes and decrease their probabilities of negative outcomes, 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.

### ***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.

Availability has been a very popular theoretical framework to explain different phenomena. Part of this appeal, some critics 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).

### **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 women were in the list. 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. Note that there is an inherent ambiguity 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 original experiments 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 participants 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 recall twelve assertive behaviors should judge themselves as being more assertive than those who recall 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? A recent meta-analysis found that the impact of feelings of ease of recall on judgment has a medium effect size. Moreover, the findings suggest that variables other than ease of retrieval influence the judgments (Weingarten & Hutchinson, 2018). Ease of retrieval includes ease of recall and ease of other kinds of memory retrieval but, mostly, ease of recall is manipulated.

Several variables have been found to affect the role of ease of recall, among them the diagnosticity of the recall experience; its representativeness towards the target; its relevance for the judgment; the malleability of the judgment; and processing motivation and opportunity (for reviews, see Greifeneder et al., 2011; Schwarz, 1998, 2004; Weingarten & Hutchinson, 2018). We 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 set-up, a participant in the same experiment listens to music, a meditational piece at half speed. He is told that this music

facilitates the recall of self-assertive 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 to be undiagnostic as information for judging 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. A naïve theory about the source of the feeling which is perceived as unrelated to the judgment renders the feeling non-diagnostic.

However, naïve theories can also serve as an explanation for the meaning of the feelings which in some circumstances influence the judgment. 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' naïve 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; and another group 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: 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.

### ***Representativeness of the feelings towards the retrieval target***

You are more likely to rely on your feelings in judgments when you think that the feelings originate from the target and reflect its essential characteristics (Weingarten & Hutchinson, 2018). One example is that you are more likely to display an ease of retrieval effect when making judgments about yourself versus others, or your ingroup versus an outgroup. For example, participants were asked to recall either two or six instances of when they had been creative. Participants who experienced greater ease of recall (two

instances) judged ingroup members to be more creative than outgroup members (Woltin et al., 2014).

### ***The relevance of the feeling for the judgment at hand***

The use of feelings as information becomes more likely if they are seen as relevant for the judgment, and their use depends on both dispositional characteristics and contextual factors (see Greifeneder et al., 2011). In a study exploring dispositional characteristics, 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, lay people but not experts were influenced by ease of recall (Ofir, 2000). Similarly, people's belief in the power of intuition increases the reliance on feelings (Keller & Bless, 2009), and powerful individuals rely 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). An example of a contextual factor is mood. People are more likely to rely on their recall experiences and to make fast evaluations when they are in a positive mood (Ruder & Bless, 2003).

### ***The malleability of the judgment***

Another moderator is the malleability of the judgment. Individuals with strong preexisting attitudes towards the topic are less likely to be influenced by feelings. Let us assume that the recall experience is both informative for the target stimulus and relevant for the judgment at hand, but people might still 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 et al. (1999) found that 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.

### ***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, your personal experiences may affect your judgment. If there is no history of heart disease in your family, it feels less relevant and you probably base your judgment of vulnerability on ease of recall. However, if heart disease has occurred in your family, it might feel more relevant to you and make you more motivated to process this information systematically, for example, by paying attention to the actual number of risk-increasing behaviors you are able to list. Rothman and Schwarz

(1998) explored the consequences of processing motivation by asking participants to list either three or eight behaviors that increased or decreased the risk of heart disease, where about half of the participants had a family history of heart disease, the others had not. 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-increasing behaviors or *eight* rather than *three* examples of risk-decreasing behaviors. This pattern reversed for participants with a family history of heart disease who instead relied on the amount of information they retrieved. These participants judged themselves to be more vulnerable and thought that there was a higher need to change their behavior if they had to recall *eight* rather than *three* examples of risk-increasing behaviors or *three* rather than *eight* examples of risk-decreasing 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.

Individuals sometimes do not lack processing motivation but the opportunity to process information deeply. Processing opportunities 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 recall as information to make their judgments when they were under cognitive load and did not have sufficient processing opportunities.

### **From availability to retrieval fluency**

The concept of availability has become very popular and went far beyond estimates of frequencies. The experimental paradigm introduced by Schwarz and colleagues (1991) revealed that people rely on ease of retrieval. Several variables have been found to increase the use of ease of recall such as the feelings' representativeness towards the target; its relevance for the judgment; or the opportunity and motivation to process the judgment. On the other hand, if the feelings can be attributed to some other source (and thus deemed non-diagnostic) or if the judgment lacks malleability, people use amount of information. Tversky and Kahneman (1973) defined availability as the ease with which relevant instances come to mind, which is compatible with findings by Schwarz and colleagues and the more recent concept of retrieval fluency. Nowadays, retrieval fluency is one of several types of processing fluency, which is the experienced ease with which a mental operation is performed (see Reber & Greifeneder, 2017). In other words, the concept of ease of retrieval has been broadened and – together with other cognitive processes, such as perceptual fluency and encoding fluency – been subsumed under the umbrella term *processing fluency*. For a classification of different types of fluency, see Alter and Oppenheimer (2009). We will now focus on some recent research on retrieval fluency.

Hertwig et al. (2008) tested the use of retrieval fluency as a judgmental heuristic in its own right. Participants in their study had to choose which of two US cities, such as San Antonio or Portland, is largest. 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. However, recent research has shown that, although theoretically interesting, it seems that people rarely use the fluency heuristic in such decisions (Pohl et al., 2016).

However, retrieval fluency influences judgments relevant for learning at school, such as feelings of knowing, judgments of learning, and performance estimates (see Reber & Greifeneder, 2017, for a review). Benjamin et al. (1998) examined effects of retrieval fluency on predictions of learning (see Chapter 19 on biases in metacognitive judgments). Participants in their study had to answer general knowledge questions and to predict for each question whether they will later remember the answer. Participants predicted that they would remember answers when they could retrieve them 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 experiencing difficulties when retrieving an answer results in predictions of worse recall when recall in fact is better. Another series of experiments showed that learners use retrieval fluency not only in prospective performance judgments but also in retrospective performance estimates (Reber et al., 2006). In other words, retrieval fluency is not only involved in predicting future performance but also in judging past performance. Similarly, participants had more confidence in solutions to a problem that were easy to retrieve; participants merely inferred from the ease with which they could retrieve a problem's solution that this solution must be true (Ackerman & Zalmanov, 2012).

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 fancy caviar in the last two years, most of them probably are able to count the frequency of this event. This means that availability – or retrieval fluency – 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) found by the manipulation of music as the alleged source of fluency that retrieval fluency is related to judged frequency (for similar findings, see Wänke et al., 1995), 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 stimulus events and demonstrated an influence of this manipulation on frequency judgments. To sum up, when estimating high frequencies, the evidence suggests that people use the availability heuristic often but not always. However, there is good evidence for the use of retrieval fluency as a cue to inferences, for example in the domain of learning.

## Conclusions

We 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 of the recall experience, representative of the target, relevant for the judgment, the judgment was malleable, and they had both the opportunity and motivation to process the judgment.

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 resolved, 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, vividness of information, and the inability to adopt the perspective of another person.
- Availability affects frequency estimates and various 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 feelings' representativeness towards the target; the relevance of the experience for the judgment; malleability of the judgment; and processing motivation and opportunities.
- Research on retrieval fluency explored effects of feelings relevant to learning at school.

## **Further reading**

The classical piece on this topic is the article by Tversky and Kahneman (1973) that has been cited over 12,000 times to date (source: Google Scholar). We recommend reading some elegant studies into availability, for example, by Lichtenstein et al. (1978), Ross and Sicoly (1979), and Schwarz et al. (1991). Weingarten and Hutchinson (2018) offer a meta-analysis of ease of retrieval effects. For early applications of the availability heuristic, see Nisbett and Ross (1980), for more recent studies, Braga et al. (2018) and Hertwig et al. (2008).

## **References**

- Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency–confidence association in problem solving. *Psychonomic Bulletin & Review*, 19, 1187–1192.
- 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.
- 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.

- Braga, J. N., Ferreira, M. B., Sherman, S. J., Mata, A., Jacinto, S., & Ferreira, M. (2018). What's next? Disentangling availability from representativeness using binary decision tasks. *Journal of Experimental Social Psychology*, 76, 307–319.
- Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 1539–1553.
- Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology*, 8(2), 195–204.
- Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1), 107–143.
- 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.
- Hogarth, R. M., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science*, 24(5), 379–385.
- Keller, J., & Bless, H. (2009). Predicting future affective states: How ease of retrieval and faith in intuition moderate the impact of activated content. *European Journal of Social Psychology*, 39(3), 467–476.
- 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.
- Pohl, R. F., Erdsfelder, E., Michalkiewicz, M., Castela, M., & Hilbig, B. E. (2016). The limited use of the fluency heuristic: Converging evidence across different procedures. *Memory & Cognition*, 44(7), 1114–1126.
- Reber, R., & Greifeneder, R. (2017). Processing fluency in education: How metacognitive feelings shape learning, belief formation, and affect. *Educational Psychologist*, 52, 84–103.
- 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.
- 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.

- 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.
- 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.
- Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5, 207–232.
- Tulving, E. (1983). *Elements in episodic memory*: Oxford: Oxford University 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.
- Weingarten, E., & Hutchinson, J. W. (2018). Does ease mediate the ease-of-retrieval effect? A meta-analysis. *Psychological Bulletin*, 144(3), 227–283.
- 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.
- Woltin, K.-A., Corneille, O., & Yzerbyt, V. Y. (2014). Retrieving autobiographical memories influences judgments about others. *Personality and Social Psychology Bulletin*, 40(4), 526–539.
- Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *Etc. Frequency processing and cognition* (pp. 21–36). Oxford: Oxford University Press.

# 12 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,  $p(\text{ticket No. 1}) = p(\text{ticket No. 50}) = p(\text{ticket No. 63}) = .01$ . 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 cell-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 2 and 3 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 12.1. When Kahneman and Tversky (1974) asked people to compare alternatives (a) and (b) in the example in the text box, 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 12.1 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 with 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. 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 12.2 (from Kahneman & Tversky, 1972, 1974). 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 large samples are much less likely to be biased than small samples. 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 12.2 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% of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50%, sometimes lower. For a period of one year, each hospital recorded the days on which more than 60% 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% 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 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  $p(\text{Data} | H_0)$ , 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  $p(H_0 | \text{Data})$ , 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 3). 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 12.3 (adapted from a study by Ayton & Wright, 1987). Following an “inverted” variant of the representativeness heuristic, we would believe that Alice in Text box 12.3 used a coin, because her sequence looks like a typical random sequence, whereas Susan probably cheated, because her sequence looks arranged.

#### **Text box 12.3 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 that can occur by chance, but also what kind of sequences that 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 12.4. 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 12.4 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 12.2 and 12.4 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 (Lem et al., 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 3 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 2. For instance, people thought that it was likely that Björn 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 this passage: "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, 2.277–278, 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 also 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 (1974) 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 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 on 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 18 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 et al., 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. In the latter case, 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 et al., 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, and 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 (X could be the typical income of college professors).
- M is a class and X is an instance of that class (X is regarded to be a “representative” American writer).
- M is a class and X is a subset of M (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.

(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 are not universal*

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 a 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 et al., 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 always 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 11) is more applicable to frequentistic probabilities (Jones et al., 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 irrational***

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 problems and base-rate problems have similarly been given pragmatic interpretations that make “conjunction errors” and “base-rate neglect” less fallacious than they originally appeared (see Chapters 2 and 3).

A drawback of 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 knowledge of probabilistic phenomena than the experimenters. For instance, Hahn and Warren (2009) noted that in a short random binary sequence, some strings of specific outcomes are actually less likely to be observed than others. However, more recent evidence shows that participants are less sensitive to variations of objective probabilities than to variations in representativeness for such sequences (Reimers et al., 2018).

### *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 2. Base-rate neglect, as discussed in Chapter 3, 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, for this player, 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 is unavoidable, but I will quickly convince myself that, despite the similarity, the probability of Putin himself or his twin brother riding the local bus is essentially zero. Similarly, statistical knowledge and logical arguments may convince me that large biases occur less frequently in big than in small 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”, 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 & Shul, 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. A distinction between Type 1 and Type 2 processing has currently been preferred as more neutral labels (Stanovich & Toplak, 2012). Labels aside, we may just 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.

## A final demonstration

When we communicate probabilities in daily life, we commonly use words. We say “likely” more often than we say “70%”. Does representativeness dictate our usage of this word? Consider a distribution of estimates like the one displayed in Text box 12.5 (adapted from Teigen, Juanchich, & Løhre, 2021). Measures of central tendency (means, medians and modes) are commonly agreed to be the most *representative* values of a distribution (Bhattacharyya & Johnson, 1977, p. 27). If probabilities reflect representativeness, the central value of a distribution would be described as “likely”, even if it has a less than 50% chance to occur.

### Text box 12.5 Does “likely” and representative mean the same?

Large public construction projects (schools, roads, railways) are carefully planned, and yet surrounded by uncertainty. They undergo a quality assurance process by experts who estimate expected costs, resulting in a probability distribution like the one in Figure 12.1 for a highway project in the South of England (costs in millions of pounds).

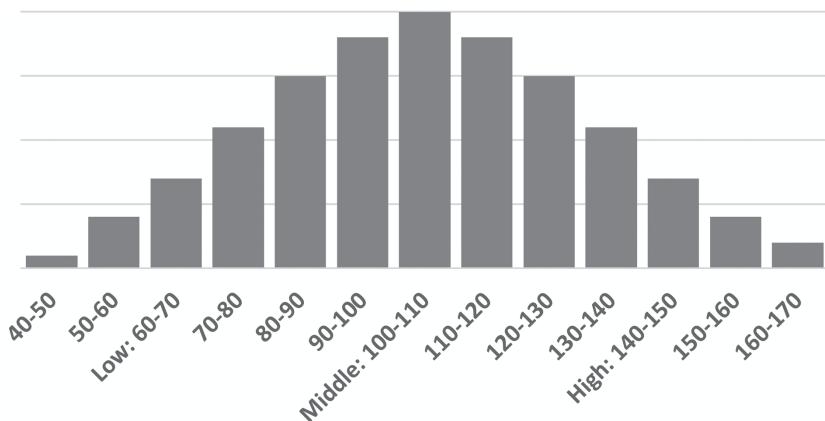


Figure 12.1 Probability distribution of costs.

How well do you agree with these statements? Select in each case the phrase that feels most right.

- (a) Costs between £100 and £110 million are likely/not likely (choose one)
- (b) Costs of more than £130 million are likely/not likely
- (c) Costs of less than £80 million are likely/not likely
- (d) Costs between £90 and £120 million are likely/not likely

Costs in statements (a)–(c) are approximately equally probable (15.0%, 18.0%, and 17.3%), according to the areas displayed in the graph, whereas the range in statement (d) corresponds to a probability of about 42%. If the term *likely* is reserved for probabilities larger than 50%, none of these statements describes “likely” costs. However, we predict that statements (a) and (d) will appear “likely”, by describing central and hence representative portions of the graph. Statements (b) and (c) are less representative, characterizing outcomes in the tails, and are accordingly “not likely”.<sup>1</sup>

### 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.

## Note

- 1 Since this chapter was written, we have actually conducted the experiment. Statements (a) and (d) were, as predicted, considered “likely” (by around 90%), whereas statements (b) and (c) were deemed “not likely” (by around 75%).

## 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 et al. (2012).

## References

- Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Meta-cognitive 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.
- Bhattacharyya, G. K., & Johnson, R. A. (1977). *Statistical concepts and methods*. New York: Wiley.
- 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.
- Hahn, U., & Warren, P. A. (2009). Perceptions of randomness: Why three heads are better than four. *Psychological Review*, 116(2), 454–461.
- 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.
- Reimers, S., Donkin, C., & Le Pelley, M. E. (2018). Perceptions of randomness in binary sequences: Normative, heuristic, or both? *Cognition*, 172, 11–25.
- 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? *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., & Toplak, M. E. (2012). Defining features versus incidental correlates of Type 1 and Type 2 processing. *Mind & Society*, 11, 3–13.
- 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., Juanchich, M., & Løhre, E. (2021). What is a “likely” amount? Representative (modal) values are considered likely even when their probabilities are low. Unpublished manuscript, University of Oslo.
- 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.

# 13 Anchoring effect

*Štěpán Bahník, Thomas Mussweiler, 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 and robust influences on human judgment, namely the anchoring effect (Tversky & Kahneman, 1974). As a legal judge you do not want to be directly influenced by a journalist's question. But if you were, you would be in good company. A study by Englich et al. (2006) showed that accomplished trial judges with an average of more than ten 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 eight months longer than judges who considered a low demand of one year. A difference of eight 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 attorney's counter-demand (Englich et al., 2005).

## The anchoring phenomenon

As was illustrated in the legal settings, human judgment is often influenced by biased values (for a classroom demonstration, see Text box 13.1). Such judgmental "anchoring" is defined as the assimilation of a judgment to a previously considered standard. For at least two reasons, anchoring is a remarkable influence on human judgment. First, anchoring

effects are pervasive and robust. Second, its underlying mechanisms are still a matter of lively debate even after many years of investigation.

### **Text box 13.1 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. 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 comparative 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.

Comparative anchoring questions and high (and low) anchor values may be:

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) AD?
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?

##### **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 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 (e.g., estimates of the mean temperature in the Antarctica; Mussweiler & Strack, 1999a) to the momentous (e.g., 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 et al., 2000; Northcraft & Neale, 1987), estimates of self-efficacy (Cervone & Peake, 1986), probability assessments (Plous, 1989), estimates of task duration (Lorko et al., 2019), valuation of products (Ariely et al., 2003; Yoon et al., 2019), legal judgments (Bystranowski et al., 2021; Englich & Mussweiler, 2001; Englich et al., 2005, 2006), and negotiations (Galinsky & Mussweiler, 2001).

Not only is the anchoring effect influential in a plethora of settings, its 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 in the direction of an unreasonably high anchor value of 140 years. Furthermore, anchoring may be in some cases uninfluenced by manipulations of accuracy motivation. Specifically, even a substantial financial incentive for being accurate did not appreciably reduce the anchoring effect (Enke et al., 2021). In addition, it has been demonstrated that anchoring is often largely unaffected by knowledge and expertise (Cheek et al., 2015; Englich & Mussweiler, 2001; Englich et al., 2006; Northcraft & Neale, 1987; but see Smith

et al., 2013). 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 with no legal knowledge. Furthermore, anchoring effects can persist even over fairly long periods of time. For example, anchoring effects were still noticeable eight weeks after the anchoring had occurred (Yoon & Fong, 2019). While the above examples show that anchoring is a robust phenomenon, its robustness depends on the mechanism that led to it in a given judgment.

### **Relevance**

Anchoring has not only been shown to be robust and pervasive in various domains, but it has also been suggested to play a role in a wide array of seemingly unrelated judgmental phenomena. For example, anchoring has been used to explain hindsight bias – the assimilation of a recollected estimate towards an observed outcome (Hawkins & Hastie, 1990; see also Chapter 27). The egocentricity of social judgments has also been attributed to an anchoring mechanism (Gilovich et al., 2000). Specifically, it was found that people may overestimate the extent to which they are noted by others, because they use their own rich experiences as an anchor when estimating the experience of others. 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 of 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 of the chapter). For example, first offers in negotiations may serve as anchors and thus influence the final 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). Like anchoring effects on criminal sentencing decisions, research in the civil context of damage awards has shown that the higher a plaintiff's request, the higher the awarded damage (Hastie et al., 1999; Malouff & Schutte, 1989; Marti & Wissler, 2000). In personal injury verdicts, the requested compensation was found to systematically influence 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 increase the awards (Hinsz & Indahl, 1995). While the applied research has clearly demonstrated anchoring in the laboratory, the effect seems to be more fragile in the field (Jung et al., 2016).

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 invoking an unspecific notion of anchoring. By itself, the anchoring notion does not 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 which does

not go beyond the terms “assimilation” and “contrast” (Strack, 1992). To be used as an explanatory concept, the underlying psychological mechanisms have to be sufficiently understood.

### **Paradigms**

Anchoring effects are typically examined in the standard paradigm introduced by Tversky and Kahneman (1974). In this paradigm, anchors are explicitly provided 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. 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 generated. 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, 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. As a result of the anchoring procedure, 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 ten-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 v. \$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 five seconds to estimate the result of a product that was either presented in an ascending sequence ( $1 \times 2 \times \dots \times 8$ ) or in a descending sequence ( $8 \times 7 \times \dots \times 1$ ). 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 (Epley & Gilovich, 2001).

The “sequential judgment paradigm” also relies on self-generated anchors. However, an anchor is self-generated by answering an unrelated question in this paradigm. 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). For example, in one experiment (Wilson et al., 1996) demonstrating such a “basic” anchoring effect, participants were first asked to copy either five pages of numbers ranging from 4,421 to 4,579 or five 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 five pages of high numbers estimated this number to be higher than those who had copied five pages of words. Thus, the irrelevant high anchor presented in the preceding task influenced the judgment. While some studies have suggested such basic anchoring effects, more recent studies (Brewer & Chapman, 2002; Klein et al., 2018; Röseler et al., 2021; Shanks et al., 2020) failed to replicate them, showing that subliminal anchoring is weaker and more fragile than other forms of anchoring.

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 their mere presence. Most of the anchoring research, however, first asks participants a comparative and then an absolute question following the standard paradigm introduced by Tversky and Kahneman (1974).

## Theoretical accounts

To date five main 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 the response scale.

### *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). The 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 may generate the duration of Earth's orbit as an anchor 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' orbit takes longer than Earth's (Epley & Gilovich, 2001). Therefore, 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 the 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 generating judgments that are higher versus lower than the anchor. That is, people should know the direction of adjustment from the anchor. Instead, 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 anchor values are unacceptable and self-generated rather than acceptable and externally provided. 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 externally provided (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 for externally provided anchors (Enke et al., 2021; Epley & Gilovich, 2005; Simmons et al., 2010).

Another explanation of insufficient adjustment argues that adjustment may be insufficient because people are averse to extreme corrections. Consistently, when predicting a stock price, participants were more likely to adjust from the current stock price by at least \$5 if the maximum allowable adjustment was "\$14 or more" than when it was "\$5 or more" (Lewis et al., 2019). The adjustment by \$5 or more was extreme in the latter case, but not when the maximum adjustment was \$14 or more. There is currently little research that would suggest under what circumstances extremeness aversion causes the anchoring effect and what are its limitations. Its contribution to anchoring effects is therefore not currently clear.

### ***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. Another study shows that conversational inferences may play some role in anchoring paradigms even if experimenters try to make the anchor uninformative. In particular, Frederick et al. (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 (Frederick & Mochon, 2012). Thus, although conversational inferences are potential determinants of anchoring in natural situations, they are not a necessary precondition.

### **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. A study by Oppenheimer et al. (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. 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 associations with the anchor. 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 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 eight weeks (Yoon & Fong, 2019) is clearly in conflict with this implication and further renders a purely numeric conceptualization of the standard anchoring paradigm unconvincing.

While numeric priming offers a parsimonious explanation for some outcomes (Critcher & Gilovich, 2008; Wilson et al., 1996; Wong & Kwong, 2000), these effects are not robust (Brewer & Chapman, 2002; Klein et al., 2018; Röseler et al., 2021; Shanks et al., 2020), and numeric priming thus does not seem to play a significant role in most anchoring effects (Newell & Shanks, 2014).

### **Selective accessibility**

The fourth theoretical account is the selective-accessibility model of anchoring. It claims that anchoring is the result of an increased accessibility of information consistent with an anchor (Mussweiler & Strack, 1999a, 1999b; Strack et al., 2016; Strack & Mussweiler, 1997; for a related account, see Chapman & Johnson, 1994, 1999). The model attempts to explain anchoring by linking it to two fundamental principles of social cognition research: (1) *hypothesis-consistent testing* and (2) *semantic priming*. More specifically, the model postulates that comparing the judgmental target to the anchor 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 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). To subsequently generate the requested numeric estimate, judges then rely primarily on easily accessible information (Higgins, 1996), so that their estimate is heavily influenced by the anchor-consistent knowledge that they had 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 role of hypothesis-consistent testing in anchoring has been supported by Mussweiler and Strack (1999a), who showed that absolute judgments in the standard anchoring paradigm are higher when the comparative question asks whether the target value is larger than the anchor than when it asks whether the target value is smaller than the anchor. The tested hypothesis is influenced by the wording of the comparative question and the hypothesis-consistent testing thus activates different information depending on the

wording. Another study showed that the anchoring effect is not affected by a prompt to consider similarities of the target and the anchor, but it is reduced when the prompt asks to consider differences (Chapman & Johnson, 1999). According to hypothesis-consistent testing, the default response to the comparative question is consideration of similarities to the anchor, which explains the results. The selective-accessibility model argues that hypothesis-consistent testing leads to activation of the information compatible with the anchor. Accordingly, when instructed to list features that came to mind in the standard anchoring paradigm, participants were more likely to list thoughts consistent with the anchor value (Mussweiler & Strack, 1999a).

Additional support comes from a study that tested a prediction derived from the selective-accessibility model (Bahník & Strack, 2016). In particular, the model implies that 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 which does not overlap with information normally used for making the absolute judgment of the average summer temperature and which 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 which 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).

Some studies (Frederick & Mochon, 2012; Mochon & Frederick, 2013) suggest that anchoring can occur even when the targets 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 a similar degree by a comparative question related to the wettest place in the US (Mount Waialeale) as to Death Valley (Frederick

& Mochon, 2012). 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. Accordingly, more recent evidence shows that anchoring is stronger when the targets of the comparative and absolute judgment questions are more similar (Sailors & Heyman, 2019). While some studies suggest that the selective-accessibility mechanism may not be necessary for producing anchoring in the standard paradigm, the evidence indicates that it is a potent theoretical account.

### ***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 had 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 an assimilation effect as a result of anchoring. For example, considering 10% of African states as the proportion in the United Nations results in the feeling that 45% 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 this 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 1,254 pounds. The anchor in a form of the estimated weight of a raccoon was therefore embedded in the subsequent judgment. Presumably, having answered 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 had 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 seven-point heaviness scale were not influenced by this judgment and estimated the weight of a giraffe to be on average 1,265 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 1,000 pounds than participants that did not estimate the weight of a wolf. Apparently, the same 1,000 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; Sailors & Heyman, 2019).

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. As the calorie estimates of a salad and cake are both lower than that of a cheesesteak, they should lead to a lower judgment of calories for a cheesesteak. This prediction follows from the scale distortion theory, which assumes that anchoring should be largely independent of the target of judgment. An anchor should also influence absolute judgments through scale distortion, but only if the scale is still distorted when making the judgment. Contrary to this prediction, Bahník (2021) showed that a comparative question in the standard anchoring paradigm influences an answer to the absolute judgment question even if the two questions are interposed by other judgments on the same scale. These judgments should have overridden any scale-distortion effect of the anchor by the time the absolute judgment is made. Contrary to the scale-distortion theory, anchoring can also occur under certain circumstances even if the anchor is on a different scale than the absolute judgment (Harris & Speekenbrink, 2016).

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 case 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. Selective accessibility seems to lie behind anchoring in case 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 case of little knowledge about the judgmental domain.

While the evidence suggests operation of different processes under different circumstances, there is little evidence that multiple 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, their interaction 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 numeric judgments toward a previously considered value 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, and may persist over long periods of time.
- There are different underlying mechanisms that may contribute to the generation of anchoring effects. Specifically, anchoring may result from insufficient adjustment, from the use of conversational inferences, from selective accessibility of information consistent with an anchor, or from the distortion of a response scale.

## Further reading

The main accounts of anchoring are explored in more detail in Strack et al. (2016), 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. (2021). Anchoring without scale distortion. *Judgment and Decision Making*, 16, 131–141.
- 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.
- Brewer, N. T., & Chapman, G. B. (2002). The fragile basic anchoring effect. *Journal of Behavioral Decision Making*, 15, 65–77.
- Bystranowski, P., Janik, B., Próchnicki, M., & Skórská, P. (2021). Anchoring effect in legal decision-making: A meta-analysis. *Law and Human Behavior*, 45, 1–23.
- 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 court-room. *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.
- Enke, B., Gneezy, U., Hall, B., Martin, D. C., Nelidov, V., Offerman, T., & van de Ven, J. (2021). *Cognitive biases: Mistakes or missing stakes?* (NBER Working Paper No. 28650). Cambridge, MA: National Bureau of Economic Research. [www.nber.org/papers/w28650](http://www.nber.org/papers/w28650)
- 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., Mochon, D., & Savary, J. (2014). *The role of inference in anchoring effects*. Working Paper. New Haven, CT: Yale University.
- Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General*, 141, 124–133.
- 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.
- Harris, A. J., & Speekenbrink, M. (2016). Semantic cross-scale numerical anchoring. *Judgment and Decision Making*, 11, 572–581.
- 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.
- 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: 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. (2016). Anchoring in payment: Evaluating a judgmental heuristic in field experimental settings. *Journal of Marketing Research*, 53, 354–368.
- 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: 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.
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., ... & Nosek, B. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1, 443–490.
- Lewis, J., Gaertig, C., & Simmons, J. P. (2019). Extremeness aversion is a cause of anchoring. *Psychological Science*, 30, 159–173.
- Lichtenstein, S., & Slovic, P. (1971). Reversal of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, 89, 46–55.
- Lorko, M., Servátka, M., & Zhang, L. (2019). Anchoring in project duration estimation. *Journal of Economic Behavior and Organization*, 162, 49–65.
- Malouff, J., & Schutte, N. S. (1989). Shaping juror attitudes: Effects of requesting different damage amounts in personal injury trials. *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., & 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.
- Röseler, L., Schütz, A., Blank, P. A., Dück, M., Fels, S., Kupfer, J., ... & Seida, C. (2021). Evidence against subliminal anchoring: Two close, highly powered, preregistered, and failed replication attempts. *Journal of Experimental Social Psychology*, 92, 104066.

- Sailors, J. J., & Heyman, J. E. (2019). Similarity, multiple estimations, and the anchoring effect. *Journal of General Psychology*, 146, 200–215.
- 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.
- Shanks, D. R., Barbieri-Hermite, P., & Vadillo, M. A. (2020). Do incidental environmental anchors bias consumers' price estimations? *Collabra: Psychology*, 6(1), 19.
- 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.
- Smith, A. R., Windschitl, P. D., & Bruchmann, K. (2013). Knowledge matters: Anchoring effects are moderated by knowledge level. *European Journal of Social Psychology*, 43, 97–108.
- 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.
- Strack, F., Bahník, Š., & Mussweiler, T. (2016). Anchoring: Accessibility as a cause of judgmental assimilation. *Current Opinion in Psychology*, 12, 67–70.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1130.
- 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.
- Yoon, S., & Fong, N. (2019). Uninformative anchors have persistent effects on valuation judgments. *Journal of Consumer Psychology*, 29, 391–410.
- Yoon, S., Fong, N. M., & Dimoka, A. (2019). The robustness of anchoring effects on preferential judgments. *Judgment and Decision Making*, 14, 470–487.
- Zhang, Y. C., & Schwarz, N. (2013). The power of precise numbers: A conversational logic analysis. *Journal of Experimental Social Psychology*, 49, 944–946.

# 14 Illusory truth effect

Lena Nadarevic

It is common knowledge that repeated studying leads to better memory than a single study episode. But repetition not only strengthens memory for information, it also increases the judged truth of information (Hasher et al., 1977). This *truth effect* emerges irrespective of whether the processed information is true or false. For this reason, the effect is also referred to as *illusory truth effect*. Further synonyms in the literature are *validity effect*, *reiteration effect*, and *truth-by-repetition effect*.

## Relevance of the truth effect

While most research on the truth effect has been conducted under controlled experimental conditions and with student samples, several studies show that the effect also replicates in field experiments (Boehm, 1994), with representative samples (Gigerenzer, 1984), and with naturalistic stimuli such as social media postings (Pennycook et al., 2018). First empirical work on the truth effect was published several decades ago (Hasher et al., 1977), yet interest in the effect has not diminished. To the contrary, interest in the truth effect is actually increasing, particularly with regard to its role in people's belief in misinformation and fake news (see Chapter 20). Indeed, recent studies showed that repetition enhances the perceived accuracy of true *and* fake news headlines (Pennycook et al., 2018; Smelter & Calvillo, 2020). Text box 14.1 provides further examples of the truth effect in everyday contexts.

### Text box 14.1 Real-world examples of the truth effect

- Coming across a news headline repeatedly on social media increases the perceived accuracy of the headline, even if it is discordant with one's own political view (Pennycook et al., 2018).
- Repetition also boosts people's belief in advertising claims (Hawkins & Hoch, 1992) and subjective opinions (Arkes et al., 1989).
- Repeated eyewitness testimony is perceived as more credible, even if the repeated testimony stems from a single source (Foster et al., 2012).
- Items presented in true/false tests and multiple-choice tests are judged more likely as true, if they are familiar from a previous test. This holds for correct items as well as distractors, at least in the absence of test feedback (Toppino & Luipersbeck, 1993).

## Experimental designs and measures

### Classical paradigm

The classical research paradigm on the truth effect is based on the first empirical truth-effect study by Hasher et al. (1977) and consists of at least two phases. In Phase 1, participants are presented with a list of statements. The typical stimuli used in truth-effect studies are trivia statements, half of which are factually true (e.g., *Manama is the capital of Bahrain*) and half of which are factually false (e.g., *Bolivia is the smallest landlocked country of South America*). Ideally, these statements are pretested and matched on plausibility. The participants' task in the first phase is to rate the truth of the presented statements on a Likert scale ranging from *definitely false* to *definitely true* or by providing binary true/false judgments. After several days or even weeks, participants take part in Phase 2. Again, the task is to judge the truth of several statements. This time, however, some of the statements are repeated statements from Phase 1 while the others are new.

In the classical design, there are two ways to investigate the truth effect (Dechêne et al., 2010). To test for a *within-items* truth effect, one exclusively focuses on the truth judgments for the repeated statements, that is, the statements presented in both judgment phases. A *within-items* truth effect is present if the truth judgments for these repeated statements (dark bars in Figure 14.1) are significantly higher in Phase 2 (i.e., at repeated exposure) compared to Phase 1 (i.e., at first exposure). In contrast, to test for a *between-items* truth effect, one exclusively focuses on the truth judgments in Phase 2. A *between-items* truth effect is present if the truth judgments for the repeated Phase 2 statements are significantly higher than the truth judgments for the new Phase 2 statements (see Figure 14.1).

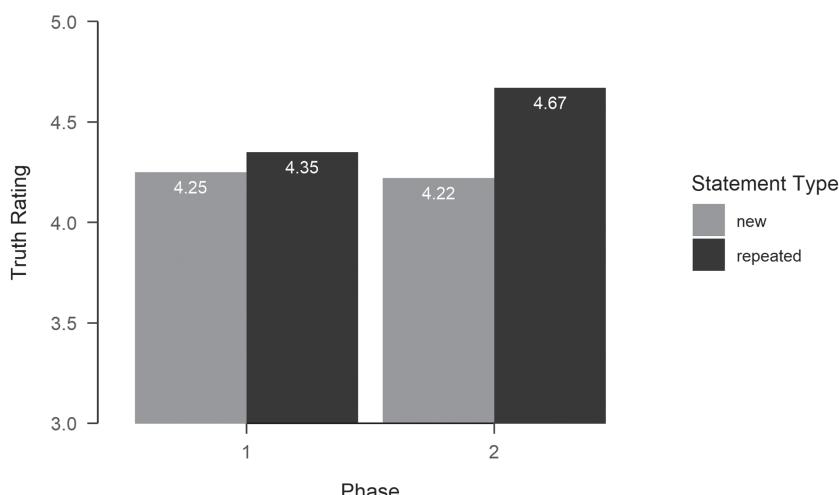


Figure 14.1 The truth effect in the classical paradigm.

Note: Mean truth ratings (1 = *definitely false*; 7 = *definitely true*) reported by Hasher et al. (1977). For ease and clarity, the graph covers only two judgment phases and the y-axis is truncated.

### **Exposure paradigm**

Because the truth effect does not replicate in the classical paradigm if the retention interval between the two judgment phases is only a few minutes instead of several days (Nadarevic & Erdfelder, 2014), many truth-effect studies use a slightly adapted procedure – the exposure paradigm. This paradigm is more economical because it allows researchers to examine the truth effect within a single experimental session. As in the classical paradigm, participants are exposed to a list of statements in the first phase of the experiment. In this *exposure phase*, however, participants do not judge the truth of the statements. Instead, they assign the statements to different semantic categories (Nadarevic & Erdfelder, 2014), judge the interestingness of the statements (Fazio et al., 2015), simply read the statements (Unkelbach & Rom, 2017), or engage in other processing tasks. The second phase, often referred to as the *test phase*, typically takes place directly after the exposure phase or shortly after. As in the classical paradigm, participants are asked to judge the truth of statements, some of which are repeated statements from the exposure phase while the others are new.

Because participants only provide truth judgments in the test phase, it is not possible to test for a within-items truth effect in the exposure paradigm. Therefore, it is even more important that researchers take great care in pretesting and counterbalancing their stimuli to ensure that a between-items truth effect cannot be attributed to factors other than statement repetition. Despite this drawback, the exposure paradigm has several advantages over the classical design. First, as noted above, it is much easier to implement (for a classroom demonstration, see Text box 14.2). Second, the exposure design is characterized by higher external validity than the classical design. People typically do not pay attention to accuracy when they read or listen to information. Thus, the exposure phase more closely resembles how participants process new information in everyday life.

### **Memory paradigm**

Finally, there are truth-effect studies in which the participants receive cues or information on the truth value of the presented statements in the first phase of the experiment – the *study phase*. Except for this credibility information in Phase 1, the procedure of this *memory design* is comparable to the other truth-effect paradigms, that is, participants provide truth judgments on repeated and new statements in a later test phase. In case of binary truth judgments, the following pattern of results indicates a truth effect in the memory design: First, the proportion of repeated, false statements judged as true is larger than the proportion of repeated, true statements judged as false. Second, the proportion of repeated, false statements judged as true is larger than the proportion of new statements judged as true. In the case of truth ratings, the second criterion (i.e., higher truth judgments for repeated, false statements compared to new statements) is typically adopted. Because the truth effect in the memory paradigm largely depends on people's memory for the truth-value information provided in the study phase, it is no longer a judgment effect in the strict sense. For this reason, this chapter focuses on the truth effect in the classical design and the exposure design.

**Text box 14.2 Classroom demonstration of the truth effect****Participants**

Ideally, there should be at least 27 participants in the class. This group size allows to detect a medium-sized truth effect (Cohen's  $d_z = 0.50$ ) with a power of at least .80, when analyzing the data with a one-tailed  $t$ -test ( $\alpha = .05$ ).

**Materials**

A minimum number of 20 statements is required. These statements are assigned to two stimulus sets matched on plausibility. Ideally, the sets should contain an equal number of true and false statements (two sample sets are provided in the Appendix).

**Procedure**

Students are randomly assigned to two groups (e.g., A and B). Group A receives a sheet with all Set-A statements and Group B a sheet with all Set-B statements listed in random order. The students' task is to rate the interestingness of the statements (e.g., on a six-point scale ranging from *not interesting at all* to *very interesting*). After this exposure phase, the first sheet is replaced by a second one, which contains the statements of both sets listed in random order. This second sheet is essentially the same for both groups. However, the correct group name (A or B) needs to be listed on the sheet in order to allow the later coding of repeated and new statements. This time, the students' task is to rate the truth of the presented statements (e.g., on a six-point rating scale ranging from *definitely false* to *definitely true*).

**Statistical analysis**

Depending on group membership, statements from one set are coded as repeated and statements from the respective other set as new. Mean truth judgments are then compared between the repeated and the new statements. Higher truth judgments for the repeated statements indicate the expected between-items truth effect. The statistical significance of this effect can be tested with a one-tailed  $t$ -test for dependent means.

**Theoretical accounts**

The following sections cover the most popular explanations of the truth effect.

**Familiarity account**

It has been proposed that the credibility-enhancing effect of repetition is mediated by familiarity (e.g., Arkes et al., 1991; Boehm, 1994). Support for this familiarity account comes from truth-effect studies in which participants judged not only the truth but also the familiarity of the presented statements. For instance, Boehm (1994) found that repetition enhanced familiarity, and familiarity enhanced judged truth. Moreover, the direct

effect of repetition on judged truth disappeared when controlling for familiarity. But how to explain the link between familiarity and judged truth?

Arkes et al. (1991) argued that when people experience a high familiarity with a given statement, they tend to misattribute this familiarity to a pre-experimental exposure to the statement. Moreover, the authors reasoned that such *source dissociations* increase judged truth, because people infer that they have heard or seen these statements from two independent sources (i.e., an external source and the experiment). Indeed, statements attributed to external sources receive considerably higher truth ratings than statements attributed to the experiment or judged to be new (Arkes et al., 1989). However, although source dissociations seem to contribute to the truth effect, the effect persists even when excluding truth judgments for statements attributed to external sources.

Alternatively, it has been proposed that people rely on familiarity when judging truth because it is an ecologically valid cue for truth (Reber & Unkelbach, 2010). In fact, if most people adhered to the maxim of quality (Grice, 1989) in their everyday communications, familiar true information should clearly outnumber familiar false information. Moreover, Unkelbach and Stahl (2009) reasoned that “there are countless possible false propositions about properties of the physical world, but only one true proposition” (p. 23). For this reason, the probability of repeatedly encountering a true statement (e.g., *Manama is the capital of Bahrain*) should be higher than encountering any related false statement (e.g., *Manama is the capital of Qatar*). Note, however, that this argument does not hold for false statements that are repeated and spread on purpose (e.g., advertisements, political propaganda, or fake news).

### ***Fluency account***

The term fluency denotes the subjective, metacognitive experience of processing ease. Under certain circumstances (which will be specified below), this experience elicits a feeling of familiarity (e.g., Whittlesea & Williams, 2000). Hence, if familiarity drives the truth effect, it should be possible to induce the effect by means of perceptual fluency manipulations instead of repetition. Indeed, Reber and Schwarz (1999) could replicate the truth effect by manipulating the readability of the presented statements. Statements that appeared in high-contrast colors on a white background, which made them easy to read, were characterized by higher truth ratings than statements that appeared in low-contrast colors on the white background, which made them more difficult to read. Other perceptual fluency manipulations produced similar results (e.g., Parks & Toth, 2006).

According to the *discrepancy-attribution hypothesis* (Whittlesea & Williams, 2000), fluency only elicits a feeling of familiarity if the stimulus is processed surprisingly fluently, that is, if there is a discrepancy between the expected and the experienced fluency. Several studies support this hypothesis (see Wänke & Hansen, 2015, for a review). For example, Dechêne et al. (2009) did not find a truth effect when manipulating statement repetition between subjects, with one group judging only repeated statements and another one judging only new statements in Phase 2. The authors reasoned that, unlike in mixed lists of new and repeated statements, people do not experience any fluency discrepancy in homogeneous lists. Although Garcia-Marques et al. (2019) did find a truth effect in a between-subjects design, the effect diminished significantly over the course of the statement list. This finding suggests that participants dynamically adjust their fluency expectations.

Even though perceptual fluency manipulations can induce illusory truth, perceptual truth effects are typically much smaller and less robust than the repetition-based

truth effect (e.g., Parks & Toth, 2006; Silva et al., 2016). Recently, Vogel et al. (2020) hypothesized that fluency effects depend on the match between the type of fluency (perceptual versus conceptual) and the judgment task (perception-related versus content-related). The authors tested this *fluency-specificity hypothesis* by orthogonally manipulating the perceptual fluency (by color contrast) and conceptual fluency (by content repetition) of statements. As predicted, conceptual fluency had a stronger effect than perceptual fluency on judgments of truth whereas the reverse pattern emerged for judgments of aesthetic pleasure. This finding is also in line with the assumption of Schwarz (2004) that people hold naïve beliefs about their metacognitive experiences. These naïve beliefs may influence whether or to what degree a certain experience is considered informative for the judgment at hand.

### ***Referential theory***

The *referential theory* of the truth effect (Unkelbach & Rom, 2017) proposes that truth judgments depend on the activation and coherence of localized networks in people's semantic memory. Specifically, the theory proposes that, when reading a statement, memory references get activated and linked. That is, building on the idea of spreading-activation models, existing links between the references are strengthened and new links start to form, which results in a localized information network. Thus, according to the theory, repeated statements are perceived as more likely true because they are characterized by more coherently linked references than novel statements. Conversely, incoherent links between references should increase the likelihood of judging the statement as "false". Accordingly, statements that contradict previously processed statements should be perceived as less likely true than new statements. In fact, such an *illusion of falseness* has been found in several studies, at least in case of a short retention interval between the initial statement presentation and its contradicting repetition (Garcia-Marques et al., 2015; Nadarevic et al., 2020; Silva et al., 2017; Unkelbach & Rom, 2017).

Importantly, Unkelbach and Rom (2017) also addressed the role of processing fluency in their referential theory. They argued that coherent statement processing produces a feeling of fluency. However, this fluency experience is considered as an output variable and not as the central mediator between repetition and perceived truth in their theory. Yet, because fluency typically accompanies perceptions of truth, the theory proposes that people learn to associate fluency with truth. This association can lead to illusions of truth "when fluency and truth are factually orthogonal, or when fluency is manipulated independent of repetition" (Unkelbach & Rom, 2017, p. 113).

### ***Integrative model***

Unlike competing theories concerning many other phenomena, the presented theories on the truth effect should not be considered as competitors. That is, the proposed explanatory approaches are not mutually exclusive, but rather complement each other. For example, fluency-based explanations supplement familiarity-based explanations by making clear predictions about the circumstances under which fluency is interpreted as familiarity. The referential theory, on the other hand, provides a process-based explanation for the source of conceptual fluency. Moreover, it can also account for the significantly smaller effect of perceptual fluency on judgments of truth. Given the high interrelatedness of the cognitive constructs that have been proposed to underlie the truth effect

(i.e., familiarity, fluency, semantic coherence), Unkelbach et al. (2019) proposed an integrative model. This model assumes that all of the above-mentioned constructs may directly or indirectly contribute to the truth effect. Possibly, future research will enable us to disentangle and to estimate the contribution of these different causes of the truth effect.

## Moderators of the truth effect

As outlined in the introduction, the truth effect is a very robust effect that occurs with different materials, in different contexts, and within different sample populations. There are only very few studies that have identified boundary conditions of the effect. The truth effect does not replicate with extremely implausible statements (e.g., *the earth is a perfect square*, Pennycook et al., 2018), with a short retention interval in the classical paradigm (Brashier et al., 2020; Calvillo & Smelter, 2020; Nadarevic & Erdfelder, 2014), and with a homogeneous list of repeated statements (Dechêne et al., 2009; but see Garcia-Marques et al., 2019). With those exceptions aside, the truth effect has been found in more than 100 studies (Fazio & Sherry, 2020). A meta-analysis of Dechêne et al. (2010) reported an average effect size of  $d = .49$ , 95% CI [.45, .55] for the between-items truth effect and  $d = .39$ , 95% CI [.32, .47] for the within-items truth effect. However, several variables have been identified that moderate the size of the effect. The following sections give an overview on different a) statement characteristics, b) context characteristics, and c) person characteristics that have been examined as potential moderators of the effect.

### **Statement characteristics**

As becomes evident from the examples presented in Text box 14.1, the truth effect has been observed for a variety of statement types. Moreover, the effect does not depend on the presentation time of statements and appears equally strong for visually and auditorily presented statements (Dechêne et al., 2010). In their meta-analysis, Dechêne et al. (2010) also compared the effect of verbatim statement repetition and gist repetition (i.e., reiteration of meaning, but not of wording) and found a larger between-items truth effect for verbatim repeated statements. For the within-items truth effect, in contrast, the number of studies with gist repetition was too small to perform a moderator analysis. Silva et al. (2017) conducted a direct test of verbatim versus gist repetition on the between-items truth effect. Their experiments did not reveal any significant differences between the two. These findings underline the importance of conceptual rather than perceptual fluency for the truth effect.

Some findings suggest that the truth effect only replicates if the presented statements stem from domains about which people are at least moderately knowledgeable (Arkes et al., 1989; Boehm, 1994). In contrast, Unkelbach and Rom (2017) observed a truth effect even for completely meaningless statements (e.g., *A ma is bigger than an omp*), although the effect was smaller than for meaningful statements. While Arkes et al. (1989) and Boehm (1994) investigated prior knowledge about the statement *domains*, they held the likelihood of prior knowledge about *individual* statements constant. In fact, early truth-effect studies had exclusively used statements for which knowledge about the correct truth status was very unlikely (so-called *difficult* statements). This choice of material rested on the widespread assumption that the truth effect could only occur for difficult statements. Fazio et al. (2015) were the first to put this assumption to the test.

At odds with the aforementioned assumption, their experiments disclosed that the truth effect also replicates for *easy* statements, that is, statements that are easily identifiable as true or false according to pretest norms. Similarly, participants even showed a truth effect for statements for which they demonstrated relevant knowledge in a later knowledge test. Furthermore, a simulation study by Fazio et al. (2019) suggests that, in principle, repetition boosts the perceived truth for easy and difficult statements alike. Empirically, however, ceiling effects counteract the truth effect for extremely plausible statements. Likewise, the midpoint of the truth judgments scale represents a ceiling for extremely implausible statements because people will refrain from judging these statements as true, despite an internal increase of perceived truth.

Nadarevic et al. (2018) investigated whether statement language moderates the truth effect. The authors presumed that the activation of semantic network references might be weaker when processing foreign-language statements as compared to native-language statements. This should lead to a smaller truth effect according to the referential theory. However, Nadarevic et al. (2018) did not observe any differences in the truth effect between a foreign-language and a native-language group, at least when the test phase followed the exposure phase in close succession. In contrast, after a two-week retention interval, the truth effect was significantly smaller in the foreign-language group. Overall, this pattern of results is incompatible with the authors' activation hypothesis. Instead, it indicates a faster decay of semantic networks in a foreign language. Although further research on foreign-language effects on truth judgments is still warranted, the findings by Nadarevic et al. (2018) show that contextual variables, such as the length of the retention interval, can play a crucial role in truth-effect studies.

### **Context characteristics**

In fact, the length of retention interval between initial statement exposure and repeated statement exposure varies strongly between different truth-effect studies. In their meta-analysis, Dechêne et al. (2010) did not find a significant effect of repetition delay on both, the between-items and the within-items truth effect. However, the results of later studies suggest that the influence of retention interval on the truth effect depends on the experimental paradigm. In the exposure paradigm, the truth effect tends to be larger when initial statement exposure and repeated exposure take place within the same experimental session than when they are separated by a one-week delay (e.g., Silva et al., 2017; Stump et al., 2021). Conversely, in the classical design the effect does not appear within one session but after a week (Nadarevic & Erdfelder, 2014).

Most studies on the truth effect have compared truth judgments for once-repeated statements and new statements. Few studies (e.g., Gigerenzer, 1984; Hasher et al., 1977; Pennycook et al., 2018) also implemented a second repetition and found a larger truth effect for twice-repeated than for once-repeated statements. However, the increase of the truth effect due to the second repetition was rather small (but see Pennycook et al., 2018). Going further, Hawkins et al. (2001) included up to four repetitions and DiFonzo et al. (2015) up to nine repetitions in their experiments. Both studies found a logarithmic relationship between number of repetitions and subjective truth ratings. But not all researchers observed an increase of the truth effect with further repetitions (see Arkes et al., 1991). Fazio et al. (2021) hypothesized that the size of the truth effect not only depends on the number of repetitions, but also on the spacing of repetitions. In line with prior studies, the authors found a logarithmic increase of the truth effect across repetitions. But, unlike

what was predicted, this pattern was not moderated by spacing (one day versus four days between each repetition).

Specific to the exposure paradigm it has also been studied whether the truth effect depends on the processing task in the exposure phase. For example, in a study by Hawkins and Hoch (1992), participants were either instructed to rehearse the statements in the exposure phase (*rote rehearsal task*), to judge the comprehensibility of the statements (*comprehension task*), or to count the number of a specific letter within a statement (*orthographic task*). The rote rehearsal group showed the largest truth effect followed by the comprehension group. In the orthographic group, in contrast, the effect was not significant. Based on this finding the authors concluded that “there appears to be a minimum level of processing that must occur for the truth effect to take place” (p. 222). Overall, the truth effect seems to be larger when the processing task increases memory for the content of the statements (see also Unkelbach & Rom, 2017).

What is less clear is the role of people’s processing capacity when judging truth. Two research papers on this topic obtained different results. Garcia-Marques et al. (2016) manipulated participants’ processing capacity in the test phase by means of cognitive load. They found a smaller truth effect in a low-load condition compared to a high-load condition, at least when task instructions emphasized accurate judgments. When the instructions emphasized intuitive judgments, in contrast, the truth effect was not affected by cognitive load. Nadarevic et al. (2021) investigated the truth effect under time-pressure conditions. Participants either had to provide very fast truth judgments in the test phase or could take as much time as they wanted for the truth judgments. Unlike cognitive load, time-pressure did not moderate the truth effect, and this null effect persisted with different task instructions and different response deadlines in the time-pressure group.

### **Person characteristics**

Typically, the truth effect has been studied with student samples of young adults. Thus, it is an interesting question whether the effect is stable across the lifespan. Studies that compared the effect between young and old adults produced mixed results. On a descriptive level, the meta-analysis of Dechêne et al. (2010) speaks in favor of a larger truth effect for older adults aged about 73 years ( $d = 0.64$ ) compared to younger adults aged about 23 years ( $d = 0.49$ ). However, this difference did not reach statistical significance. Surprisingly, there has been only a single study to date that has examined the truth effect in children (Fazio & Sherry, 2020). This study, which examined the truth effect in 5-year-olds, 10-year-olds, and a group of adults, showed that the effect is already present by the age of 5. What is more, the truth effect did not differ significantly between age groups. However, because Fazio and Sherry (2020) had not designed their study to detect effect-size differences between age groups, this null effect must be interpreted with caution.

A number of studies aimed at exploring whether individuals with certain personality traits are particularly susceptible to the truth effect. Arkes et al. (1991) and Boehm (1994) were the first to investigate interindividual differences in the truth effect. The authors tested a possible relationship between the truth effect and participants’ Need for Cognition (NfC, Cacioppo & Petty, 1982), but did not detect a relationship. Newman et al. (2020), in contrast, found tentative evidence that participants high in NfC show a larger truth effect than participants low in NfC. However, this effect was only evident when participants were not informed that they would see true *and* false statements in the exposure phase. De Keersmaecker et al. (2020) explored whether any of the following

constructs accounts for individual differences in the truth effect, but failed to find an association: cognitive ability (e.g., verbal intelligence), cognitive style (e.g., preference for intuition or deliberation, respectively), and Need for Cognitive Closure (NCC, Webster & Kruglanski, 1994). In contrast, a recent study by Stump et al. (2021) found a larger truth effect for people high in NCC than for people low in NCC. In one of their experiments, however, this relationship was evident only after a ten-minute retention interval but not after a one-week interval.

Taken together, procedural differences such as differences in task instructions (Newman et al., 2020) or differences in the length of retention interval (Stump et al., 2021) could account for the inconsistent findings on interindividual correlates of the truth effect. However, methodological differences could also play an important role. One general problem is that mean-difference scores of the truth effect ( $M_{repeated} - M_{new}$ ) are unreliable at the individual level (test-retest reliability:  $r \leq .12$ , Calio et al., 2021). Based on this finding, it is not particularly surprising that studies that have used mean-difference scores to examine correlates of the truth effect (e.g., De Keersmaecker et al., 2020) failed to find reliable associations. In contrast, it is more reasonable to make use of regression models or mixed linear models to test whether certain trait variables moderate the effect of statement repetition on judged truth (e.g., Newman et al., 2020; Stump et al., 2021). What is more, using a Bayesian modeling approach, Schnuerch et al. (2020) showed that people not only differ in the strength of the truth effect, but that some people even show a reversed truth effect where novel statements receive higher truth judgments than repeated ones. But it is still an open question how stable such reversals are and what causes them.

### The truth effect under naturalistic conditions

Even though the truth effect is an extremely robust phenomenon on the group level, one might ask whether it has any relevance under naturalistic conditions. In the meta-analysis by Dechêne et al. (2010), the size of the effect was moderate. But how strong is the influence of repetition on truth judgments in the real world? A study by Jalbert et al. (2020) suggests that the truth effect tends to be considerably larger in naturalistic settings than in most laboratory studies. This is because in truth-effect experiments, participants are usually “warned” before the exposure phase that they will see true and false statements. In many studies, participants are even informed about the actual base rates of true and false statements. In contrast, in the real world, there are typically no prior warnings that precede information processing. Without typical warning instructions, the size of the truth effect increased to  $d = 1.55$ , 95% CI [1.33, 1.76] in the study by Jalbert et al. (2020), which is three times larger than the effect size reported by Dechêne et al. (2010).

Most studies of the truth effect have used minimalistic designs to investigate the isolated effect of statement repetition on judgments of truth. Under these conditions, it is not particularly surprising that participants rely on metacognitive feelings (e.g., familiarity or fluency) when judging truth, because there are no other judgment cues available in the given context. But is the effect of repetition still present in a more complex environment (e.g., in a social-media context) that includes other cues? Nadarevic et al. (2020) investigated this question as follows: Participants were presented with repeated and new statements that mimicked social media news postings. Each statement appeared either with a source high in credibility (e.g., a real, trustworthy news source), a source low in credibility (e.g., a fabricated news source), or without any source information. In addition,

each statement was either accompanied by a thematically related but non-probative picture (as is often the case with social media news) or appeared without a picture. Under these conditions, Nadarevic et al. (2020) found significant effects of statement repetition and source information on judged truth, but no effect of non-probative pictures. Importantly, even though participants relied on two different cues for truth (repetition and source information), they used these cues in an additive fashion.

It is also important to note that the truth effect is not limited to individual statements or headlines, but also replicates for short (false) news stories. In a study by Polage (2012), participants provided higher plausibility and truth ratings to false news stories when they had been previously exposed to these stories (*exposure group*) than when they had not seen the stories before (*control group*). Moreover, participants in the exposure group rated themselves to be more knowledgeable about the covered topics and were more prone to source misattribution errors. That is, they were more likely to believe that they had previously heard the stories from a source outside of the experiment.

## Debiasing approaches

As the truth effect contributes to people's belief in misinformation, political propaganda, and false advertising promises, the question is how to eliminate or at least reduce this effect. Several debiasing approaches have been tested so far with different levels of success. The first debiasing study on the truth effect was conducted by Boehm (1994), who told one group of participants that they would have to justify their truth judgments to a group of peers. However, this *accountability manipulation* did not have any effect on the size of the truth effect. Nadarevic and Aßfalg (2017) tested whether it is possible to eliminate the truth effect by warning people about it. In their study, one group of participants received a warning before the test phase that informed them about the truth effect and instructed them not to show the effect. Even though the warning did not eliminate the truth effect, the effect was at least smaller for warned than for unwarned participants. A follow-up study by Calio et al. (2020) produced similar findings. Moreover, in a separate analysis of easy and difficult statements, the authors observed an elimination of the truth effect for easy statements in the warning group.

One debiasing approach that has been implemented by social-media platforms to combat misinformation is the tagging of misleading or false information with warning labels or corrections. Pennycook et al. (2018) tested the effectiveness of this intervention in a truth-effect experiment that involved fake and real news headlines. During the exposure phase, fake-news headlines appeared with a warning label ("Disputed by 3rd Party Fact-Checkers") in a warning group and without any label in a no-warning group. Even though the warning label reduced participants' willingness to share the tagged headlines on social media, the tagging did not prevent a truth effect for repeated fake-news headlines. That is, both, the warning group and the no-warning group showed a truth effect in a subsequent test phase, although the effect was somewhat smaller in the warning group.

Some studies speak to the crucial role of initial information encoding in the truth effect. For instance, in the exposure paradigm, the effect reduces drastically when participants are informed prior to information encoding that they will be exposed to true *and* false information (Jalbert et al., 2020). In the classical design, in which individuals rate the truthfulness of statements at the first encounter, the effect even disappears completely, at least in case of a short retention interval (Calvillo & Smelter, 2020; Nadarevic & Erdfelder, 2014).

Brashier et al. (2020) showed that focusing on the accuracy of information at exposure still prevented a truth effect two days later. However, this was only the case for statements for which the participants held prior knowledge.

Finally, research by Corneille et al. (2020) suggests that focusing on the fakeness rather than the truth of information may be an effective means to counter the truth effect for social-media misinformation. When the authors instructed participants to identify statements as fake news instead of judging their truth, participants even categorized repeated statements more likely as fake news than new statements. This finding suggests that in principle people are able to reinterpret familiarity (or fluency, respectively) in the context of social media. Hence, the mental reappraisal of metacognitive feelings could be an effective means to eliminate the truth effect. However, more applied studies on this approach are still pending.

## Conclusion

The truth effect is a very robust judgment bias. In contexts where true information clearly predominates (e.g., educational settings), the effect is quite adaptive, largely leading to correct truth judgments. However, in contexts where people are frequently exposed to misinformation (e.g., advertising, social media), the effect contributes to people's belief in falsehoods. Thus, in the latter context, individuals are well-advised to avoid processing information from untrustworthy or unknown sources as there is currently no intervention that reliably prevents a truth effect for misinformation. What is more, effects of misinformation on people's beliefs are difficult to correct once the false information has been encoded (Ecker et al., 2011). The authors of the *Debunking handbook 2020* (Lewandowsky et al., 2020) advise, among other things, that refutations are best presented multiple times to counter misinformation. Indeed, it makes perfect sense to fight the truth effect with its own weapon—repetition.

## Summary

- People judge repeatedly presented statements to be more likely true than new statements—a phenomenon referred to as (illusory) truth effect.
- A typical truth-effect experiment consists of two phases. In Phase 1, participants read or listen to statements which they either judge for truth (classical paradigm) or process in some other way (exposure paradigm). In Phase 2, participants are asked to judge the truth of new statements as well as of repeated statements from Phase 1.
- Higher truth judgments for the repeated Phase 2 statements than for the new Phase 2 statements indicate a between-items truth effect. The classical paradigm also allows to test for a within-items truth effect, that is, an increase in truth judgments in Phase 2 compared to Phase 1 for the repeated statements.
- Different assumptions have been made about which variables mediate the repetition effect on perceived truth, such as familiarity, fluency, and semantic coherence. Given the high relatedness of these constructs, an integrative model encompassing all of these variables currently seems most appropriate to explain the truth effect.
- Although the truth effect is a quite robust phenomenon, its effect size tends to vary depending on certain characteristics of the statements, the context, and the person. New findings suggest that the truth effect not only replicates under more naturalistic conditions, but is likely to be even larger under these conditions.

- To counteract the truth effect for false information, several debiasing approaches have been proposed and tested. Although none of these approaches has reliably prevented a truth effect, the suggested interventions could be quite effective when applied in combination.

## Further reading

Hasher et al.'s (1977) first article on the truth effect is a good read to get into the topic. The only meta-analysis of the truth effect so far is provided by Dechêne et al. (2010). It covers important findings on the effect published before the year 2010. A more timely theoretical review article stems from Unkelbach et al. (2019). Going beyond the truth effect, a review by Brashier and Marsh (2020) provides a systematic overview of various determinants of truth judgments.

## Acknowledgment

I am grateful to Martin Schnürch and Alina Kias for their valuable comments on an earlier version of this chapter.

## References

- Arkes, H. R., Boehm, L. E., & Xu, G. (1991). Determinants of judged validity. *Journal of Experimental Social Psychology*, 27(6), 576–605.
- Arkes, H. R., Hackett, C., & Boehm, L. E. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making*, 2(2), 81–94.
- Boehm, L. E. (1994). The validity effect: A search for mediating variables. *Personality and Social Psychology Bulletin*, 20(3), 285–293.
- Brashier, N. M., Eliseev, E. D., & Marsh, E. J. (2020). An initial accuracy focus prevents illusory truth. *Cognition*, 194, 104054.
- Brashier, N. M., & Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology*, 71, 499–515.
- Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42(1), 116–131.
- Calio, F., Nadarevic, L., & Musch, J. (2020). How explicit warnings reduce the truth effect: A multinomial modeling approach. *Acta Psychologica*, 211, Article 103185.
- Calio, F., Nadarevic, L., & Musch, J. (2021). Is the truth effect an individually stable phenomenon? An assessment of its test-retest stability [Manuscript in preparation]. Department of Psychology, University of Düsseldorf.
- Calvillo, D. P., & Smelter, T. J. (2020). An initial accuracy focus reduces the effect of prior exposure on perceived accuracy of news headlines. *Cognitive Research: Principles and Implications*, 5(1), 55.
- Corneille, O., Mierop, A., & Unkelbach, C. (2020). Repetition increases both the perceived truth and fakeness of information: An ecological account. *Cognition*, 205, 104470.
- De Keersmaecker, J., Dunning, D., Pennycook, G., Rand, D. G., Sanchez, C., Unkelbach, C., & Roets, A. (2020). Investigating the robustness of the illusory truth effect across individual differences in cognitive ability, need for cognitive closure, and cognitive style. *Personality and Social Psychology Bulletin*, 46(2), 204–215.
- 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(5), 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.

- DiFonzo, N., Beckstead, J. W., Stupak, N., & Walders, K. (2015). Validity judgments of rumors heard multiple times: The shape of the truth effect. *Social Influence*, 11(1), 22–39.
- Ecker, U. K. H., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, 18(3), 570–578.
- Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144(5), 993–1002.
- Fazio, L. K., Pillai, R. M., & Patel, D. (2021). The effects of repetition on belief in naturalistic settings. PsyArXiv. <https://doi.org/10.31234/osf.io/r85mw>
- Fazio, L. K., Rand, D. G., & Pennycook, G. (2019). Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin & Review*, 26(5), 1705–1710.
- Fazio, L. K., & Sherry, C. L. (2020). The effect of repetition on truth judgments across development. *Psychological Science*, 31(9), 1150–1160.
- Foster, J. L., Huthwaite, T., Yesberg, J. A., Garry, M., & Loftus, E. F. (2012). Repetition, not number of sources, increases both susceptibility to misinformation and confidence in the accuracy of eyewitnesses. *Acta Psychologica*, 139(2), 320–326.
- Garcia-Marques, T., Silva, R. R., & Mello, J. (2016). Judging the truth-value of a statement in and out of a deep processing context. *Social Cognition*, 24(1), 40–54.
- Garcia-Marques, T., Silva, R. R., Mello, J., & Hansen, J. (2019). Relative to what? Dynamic updating of fluency standards and between-participants illusions of truth. *Acta Psychologica*, 195, 71–79.
- Garcia-Marques, T., Silva, R. R., Reber, R., & Unkelbach, C. (2015). Hearing a statement now and believing the opposite later. *Journal of Experimental Social Psychology*, 56, 126–129.
- Gigerenzer, G. (1984). External validity of laboratory experiments: The frequency-validity relationship. *American Journal of Psychology*, 97, 185–195.
- Grice, H. P. (1989). Logic and conversation. In H. P. Grice (Ed.), *Studies in the way of words* (pp. 22–40). Cambridge, MA: Harvard University Press.
- Hasher, L., Goldstein, D., & Toppino, T. C. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, 16(1), 107–112.
- Hawkins, S. A., & Hoch, S. J. (1992). Low-involvement learning: Memory without evaluation. *Journal of Consumer Research*, 19(2), 212–225.
- Hawkins, S. A., Hoch, S. J., & Meyers-Levy, J. (2001). Low-involvement learning: Repetition and coherence in familiarity and belief. *Journal of Consumer Psychology*, 11(1), 1–11.
- Jalbert, M., Newman, E. J., & Schwarz, N. (2020). Only half of what I'll tell you is true: Expecting to encounter falsehoods reduces illusory truth. *Journal of Applied Research in Memory and Cognition*, 9(4), 602–613.
- Lewandowsky, S., Cook, J., Ecker, U. K. H., Albarraçín, D., Amazeen, M. A., Kendeou, P., Lombardi, D., Newman, E. J., Pennycook, G., Porter, E., Rand, D. G., Rapp, D. N., Reifler, J., Roozenbeek, J., Schmid, P., Seifert, C. M., Sinatra, G. M., Swire-Thompson, B., van der Linden, S., ... & Zaragoza, M. S. (2020). *The debunking handbook 2020*. <https://doi.org/10.17910/B7.1182>
- Nadarevic, L., & Aßfalg, A. (2017). Unveiling the truth: Warnings reduce the repetition-based truth effect. *Psychological Research*, 81(4), 814–826.
- Nadarevic, L., & Erdfelder, E. (2014). Initial judgment task and delay of the final validity-rating task moderate the truth effect. *Consciousness and Cognition*, 23, 74–84.
- Nadarevic, L., Plier, S., Thielmann, I., & Darancó, S. (2018). Foreign language reduces the longevity of the repetition-based truth effect. *Acta Psychologica*, 191, 149–159.
- Nadarevic, L., Reber, R., Helmecke, A. J., & Köse, D. (2020). Perceived truth of statements and simulated social media postings: An experimental investigation of source credibility, repeated exposure, and presentation format. *Cognitive Research: Principles and Implications*, 5(1), 56.
- Nadarevic, L., Schnuerch, M., & Stegemann, M. J. (2021). Judging fast and slow: The truth effect does not increase under time-pressure conditions. *Judgment and Decision Making*, 16(5), 1234–1266.
- Newman, E. J., Jalbert, M., Schwarz, N., & Ly, D. P. (2020). Truthiness, the illusory truth effect, and the role of need for cognition. *Consciousness and Cognition*, 78, 102866.

- Parks, C. M., & Toth, J. P. (2006). Fluency, familiarity, aging, and the illusion of truth. *Aging, Neuropsychology, and Cognition*, 13(2), 225–253.
- Pennycook, G., Cannon, T., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12), 1865–1880.
- Polage, D. C. (2012). Making up history: False memories of fake news stories. *Europe's Journal of Psychology*, 8(2), 245–250.
- Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, 8(3), 338–342.
- Reber, R., & Unkelbach, C. (2010). The epistemic status of processing fluency as source for judgments of truth. *Review of Philosophy and Psychology*, 1(4), 563–581.
- Schnuerch, M., Nadarevic, L., & Rouder, J. N. (2020). The truth revisited: Bayesian analysis of individual differences in the truth effect. *Psychonomic Bulletin & Review*, 28, 750–765.
- Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology*, 14(4), 332–348.
- Silva, R. R., Garcia-Marques, T., & Mello, J. (2016). The differential effects of fluency due to repetition and fluency due to color contrast on judgments of truth. *Psychological Research*, 80(5), 821–837.
- Silva, R. R., Garcia-Marques, T., & Reber, R. (2017). The informative value of type of repetition: Perceptual and conceptual fluency influences on judgments of truth. *Consciousness and Cognition*, 51, 53–67.
- Smelter, T. J., & Calvillo, D. P. (2020). Pictures and repeated exposure increase perceived accuracy of news headlines. *Applied Cognitive Psychology*, 34, 1061–1071.
- Stump, A., Rummel, J., & Voss, A. (2021). Is it all about the feeling? Affective and (meta-)cognitive mechanisms underlying the truth effect. *Psychological Research*. Advance online publication.
- Toppino, T. C., & Luipersbeck, S. M. (1993). Generality of the negative suggestion effect in objective tests. *Journal of Educational Research*, 86(6), 357–362.
- Unkelbach, C., Koch, A., Silva, R. R., & Garcia-Marques, T. (2019). Truth by repetition: Explanations and implications. *Current Directions in Psychological Science*, 28(3), 247–253.
- Unkelbach, C., & Rom, S. C. (2017). A referential theory of the repetition-induced truth effect. *Cognition*, 160, 110–126.
- Unkelbach, C., & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition*, 18(1), 22–38.
- Vogel, T., Silva, R. R., Thomas, A., & Wänke, M. (2020). Truth is in the mind, but beauty is in the eye: Fluency effects are moderated by a match between fluency source and judgment dimension. *Journal of Experimental Psychology: General*, 149(8), 1587–1596.
- Wänke, M., & Hansen, J. (2015). Relative processing fluency. *Current Directions in Psychological Science*, 24, 195–199.
- Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, 67(6), 1049–1062.
- Whittlesea, B. W. A., & Williams, L. D. (2000). The source of feelings of familiarity: The discrepancy-attribution hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(3), 547–565.

## APPENDIX

Trivia statements for a classroom demonstration taken from prior truth-effect studies (Nadarevic & Abfalg, 2017; Nadarevic & Erdfelder, 2014).

### Set A

- The area of skin between the eyebrows is called “glabella”. (true)
- Manama is the capital of Bahrain. (true)

- The game of dominoes is played with 28 tiles. (true)
- March and April are the hottest months in the Maldives. (true)
- Italy won the football world championship for the first time in 1934. (true)
- A dingo can survive longer without water than a camel. (false)
- Ultrasound is sound waves with frequencies below 16 Hertz. (false)
- The all-wheel drive was invented in England in 1938. (false)
- Sumatra is the fourth biggest island of the world. (false)
- The Montgolfier brothers created the first parachute. (false)

### Set B

- There are three American cities named “Santa Claus”. (true)
- Mongols flavor their tea with salt instead of sugar. (true)
- “Ligne” is a unit of length used to measure the diameter of buttons. (true)
- The Ad-Dahna desert is located in Saudi Arabia. (true)
- Canada has the longest coastline of any nation worldwide. (true)
- Hoverflies are the only insects that are able to fly backwards. (false)
- Bolivia is the smallest landlocked country of South America. (false)
- The kilt originated in Ireland and not in Scotland. (false)
- In the film *Pulp Fiction* all clocks are set to 4:10. (false)
- The world’s highest giant mammoth tree is called “King Arthur”. (false)

# 15 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 medium effect size. Subsequent investigations have confirmed this result (e.g., Gillebart et al., 2012; Inoue et al., 2018; Monahan et al., 2000; Seamon et al., 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 14 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*

MEE experiments have shown that repeated exposure to unfamiliar music leads to more positive ratings of this music (Aboukoumin, 2018; Moreland & Topolinski, 2010). Similar patterns emerge in real-world settings: The impact of repeated exposure on music sales is so strong that it is often illegal (and always unethical) for radio and internet 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 the 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 (Ruggieri & Boca, 2013; Yagi & Inoue, 2010). 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 important application of MEE principles and methods concerns intergroup behavior, with psychologists investigating 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. History is replete with examples of neighboring groups for whom decades of exposure have only heightened hostility (e.g., Israelis and Palestinians). Consistent with this pattern, Flores et al. (2018) found that mere exposure to photographs of transgender adults enhanced participants' attitudes regarding members of this group – but only for participants whose attitudes regarding transgender individuals were initially neutral or positive.

## ***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). 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, after which they report how much they like each stimulus. Most experimental studies of the MEE use a within-participants design, 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 between 10 and 20 exposures, then diminishing somewhat at higher exposure frequencies (Gillebart et al., 2012; Montoya et al., 2017).

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 one or two points on a nine-point scale (e.g., Bornstein et al., 1990; Caruso et al., 2013). Though this degree of attitude change may seem modest, 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 et al., 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., where 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 et al., 1987; Jones et al., 2011; Siegel et al., 2019). 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 (Kawakami & Yoshida, 2019; Murphy & Zajonc, 1993), and meta-analytic reviews of the mere-exposure literature (Bornstein, 1989, 1992; Montoya et al., 2017). 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, and Montoya et al., 2017, for reviews).

### *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 generally support this result (Bornstein, 1989), although Montoya et al. (2017) found that auditory stimuli produced significantly weaker exposure effects than did other types of stimuli, with a more pronounced downturn in ratings at higher exposure frequencies.

### *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 (Bornstein et al., 1990; Montoya et al., 2017). Two processes are involved. First, complex stimuli typically produce a more rapid increase in liking at lower exposure frequencies (i.e., one, two, and five exposures). Second, complex stimuli produce a less pronounced downturn in liking at higher exposure frequencies (i.e., ten 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. In most studies MEE researchers obtain an increase in liking ratings through 10 or 20 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 two separate meta-analyses (Bornstein, 1989; Montoya et al., 2017), both of which found that across different stimuli and rating dimensions, the strongest MEEs occurred following a maximum of about 10 stimulus exposures.

#### *Exposure sequence*

Significantly stronger MEEs are obtained when stimuli are presented in a heterogeneous (i.e., randomized) sequence than a homogeneous (i.e., massed) sequence during the familiarization phase of the study (Bornstein, 1989; Gillebart et al., 2012). Consistent with the results of individual experiments, meta-analytic comparisons indicated that while heterogeneous exposures produce a robust MEE ( $r = 0.30$ ), homogeneous exposures do not ( $r = -0.02$ ).

### *Exposure duration*

The relationship between stimulus exposure duration and magnitude of the exposure effect is complex. Bornstein's (1989) meta-analysis found that studies using 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. Montoya et al.'s (2017) meta-analysis found stronger exposure effects for stimuli presented for less than 15 milliseconds (ms) than for those presented at durations between 16 and 999 ms; at longer durations the magnitude of the MEE increased.

### *Delay between exposure and rating*

Seamon et al. (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; Montoya et al., 2017). In addition, Bornstein's (1989) results indicated 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 may actually inhibit 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). Consistent with these patterns, Montoya et al.'s (2017) meta-analysis indicated that stimuli presented for 15 ms or less (a conservative cutoff for "researcher defined" subliminality; see Montoya et al., 2017, p. 468) produced stronger MEEs than did stimuli with exposure durations between 16 and 999 ms.

The inverse relationship between stimulus recognition accuracy and magnitude of the MEE has been obtained in individual experiments as well. 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 et al., 2013).

Although converging results have been obtained in studies comparing MEEs for subliminal and supraliminal presentations of the same stimuli, and in investigations of patients whose neurological conditions attenuate conscious awareness of previously seen stimuli, these patterns must be qualified in two ways. First, some researchers have found stronger MEEs when stimulus presentations lead to increases in recognition memory for previously exposed stimuli (e.g., Newell & Shanks, 2007). In addition, some studies have found that the type of outcome measure employed is critical in assessing the impact of merely exposed stimuli, with subliminal stimuli leading to significant changes in implicit (but not explicit) attitudes, and supraliminal stimuli producing the reverse pattern (Kawakami & Yoshida, 2019; Smith et al., 2008).

### *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 Compton et al., 2002, for additional evidence that merely exposed stimuli are conceptually categorized during familiarization).

Along somewhat similar lines, Bornstein et al. (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. 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 of facial expressions during repeated exposures of photographs of faces, 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). Montoya et al. (2017) also concluded that mental imagery plays a key role in the exposure effect following their meta-analytic synthesis of research in this area, noting that “initial exposure produces a mental representation and subsequent exposures strengthen that representation” (p. 476); they went on to suggest that this interpretation of the MEE is consistent with “extensive evidence across different models of memory and with models for encoding information into long-term storage” (p. 476).

### *Individual differences*

Several individual difference variables have been examined in MEE studies, including need for approval, social anxiety, tolerance of ambiguity, evaluation apprehension, boredom-proneness, and sensation-seeking. For the most part, these variables had modest moderating effects, with three exceptions. Bornstein et al. (1990) found that boredom-prone participants produced significantly weaker MEEs than did non-boredom-prone participants. Kruglanski et al. (1996) found that high levels of evaluation apprehension undermined the MEE. Siegel et al. (2019) found stronger MEEs for pictures of angry faces in socially anxious than non-anxious participants.

### **Text box 15.1 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. 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. Six line-drawn visual illusions taken from Gregory (1968): the Hering illusion, Wundt’s converse of the Hering illusion, the Necker illusion, the Zöllner 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'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 nine-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). 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 questions, the experimenter presents 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 five seconds.

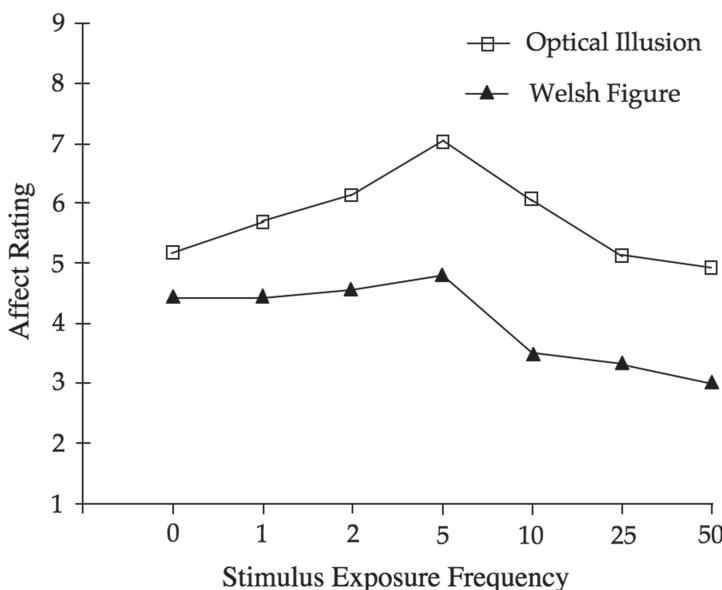
Immediately following stimulus presentations, participants are given the rating booklet, and asked to provide ratings of each stimulus. Participants circle the number on each rating scale corresponding to their rating of the stimulus pictured on that page.

### **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 15.1. 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 15.1* 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*, 58, 795. © 1990 by the American Psychological Association. Adapted with permission of the publisher.)

### 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 15.1).

#### 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 five seconds. Subsequent to the presentation phase, participants rated the seven stimuli from each set on two nine-point rating scales: like–dislike, and simple–complex, both rating scales anchored with the terms *Not at all* (1) and *Very* (9).

### **Results**

As Figure 15.1 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 five exposures then declined below baseline levels at higher exposure frequencies. Statistically, a  $2 \times 7$  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  $\times$  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 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 15.1).

Figure 15.2 summarizes the effects of stimulus type and exposure frequency on simple–complex ratings. As this figure 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) 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  $\times$  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 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 (Monahan et al., 2000).

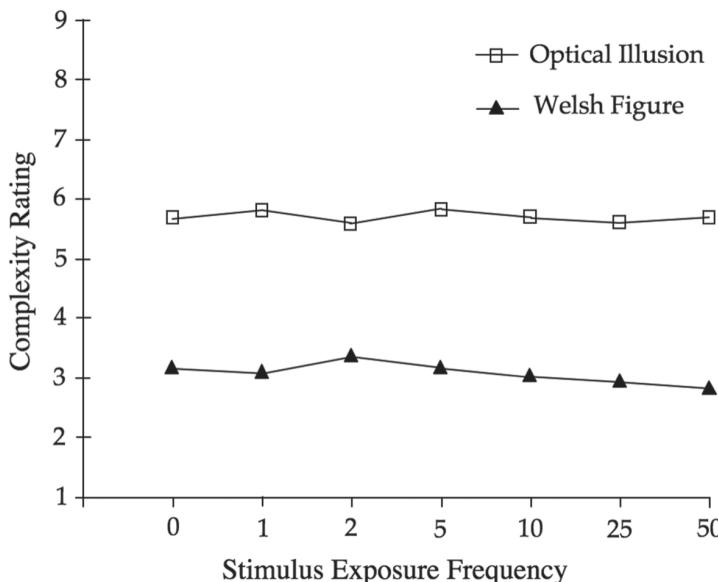


Figure 15.2 Effects of stimulus type and exposure frequency on complexity ratings of merely exposed stimuli (from “Boredom as a limiting condition on the mere exposure effect” by Robert F. Bornstein, Amy R. Kale, and Karen R. Cornell, 1990, *Journal of Personality and Social Psychology*, 58, 796. © 1990 by the American Psychological Association. Adapted with permission of the publisher.)

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 et al., 2000). In addition, two studies have used functional Magnetic Resonance Imaging (fMRI) to examine the neurological correlates of the exposure effect. In the first, Kongthong et al. (2014) found increased gamma activity in the parietal-occipital region in participants who showed strong MEEs. Ballard et al. (2017) found that magnitude of change in activation in the ventral tegmental area scaled with magnitude of self-reported preference change in an experiment wherein participants consumed a novel fluid over ten days, with liking ratings and fMRI data collected on days 1 and 10. These studies 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; Montoya et al., 2017; Moreland & Topolinski, 2010; Whittlesea & Price, 2001; Zajonc, 2001). Four of these models continue to be influential.

### ***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. 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 11 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.

### ***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.

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. Moreover, 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). Montoya et al. (2017) also concluded that the two-factor model accounts for a broad array of findings in the MEE literature.

### ***The Perceptual Fluency/Attributional Model***

Bornstein and D'Agostino's (1992, 1994) perceptual fluency/attributional (PF/A) model conceptualizes the MEE in terms of increased perceptual fluency (i.e., ease of perceptual processing) for repeatedly exposed stimuli. Consistent with the perspective of 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. The latter ("attributional") portion of the PF/A model is supported by findings which 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.

### ***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 self-generated 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. Some researchers favor an affect-based model of the MEE; others 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.

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 (Flores et al., 2018; 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 illusory truth effect (Chapter 14), 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 peoples' 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 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 et al., 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, Oeusooonthornwattana and Shanks (2010) obtained preliminary results extending the recognition heuristic from inferences to preferences.

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, gustatory), in a broad array of contexts, including both laboratory and field settings.
- MEEs have numerous real-world implications, helping explain voting behavior, advertising effects, preferences for different types of music and art, and attitudes toward people 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 (although these subliminal-supraliminal differences are in part a function of the type of measure used to assess preference and attitude change).
- 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. The formation of mental images of merely exposed stimuli plays a key role in the MEE.

## **Further reading**

Zajonc's (1968) monograph summarizes the history of the 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, as do Bornstein's (1989) meta-analysis of early research on the MEE, and Montoya et al.'s (2017) meta-analysis of more recent findings in this area. Kunst-Wilson and Zajonc's (1980) experiment has served as a model for most subliminal MEE studies during the past 35 years. 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 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

- Abakoumin, G. (2018). Mere exposure effects in the real world: Using natural experiment features from the Eurovision song contest. *Basic and Applied Social Psychology*, 40, 236–247.
- Ballard, I. C., Hennigan, K., & McClure, S. M. (2017). Mere exposure: Preference change for novel drinks reflected in human ventral tegmental area. *Journal of Cognitive Neuroscience*, 29, 793–804.
- 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.
- 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., 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., & Bornstein, R. F. (2006). Self-generated visual imagery alters the mere exposure effect. *Psychonomic Bulletin & Review*, 13, 1056–1060.
- Flores, A. R., Haider-Markel, D. P., Lewis, D. C., Miller, P. R., Tadlock, B. L., & Taylor, J. K. (2018). Challenged expectations: Mere exposure effects on attitudes about transgender people and rights. *Political Psychology*, 39, 197–216.
- Gillebaart, M., Forster, 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.
- 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.
- Inoue, K., Yagi, Y., & Sato, N. (2018). The mere exposure effect for visual image. *Memory & Cognition*, 46, 181–190.
- 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.
- Kawakami, N., & Yoshida, F. (2019). Subliminal versus supraliminal mere exposure effects: Comparing explicit and implicit attitudes. *Psychology of Consciousness: Theory, Research, and Practice*, 6, 279–291.
- Kongthong, N., Minami, T., & Nakuchi, 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.
- Montoya, R. M., Horton, R. S., Vevea, J. L., Citkowicz, M., & Lauber, E. A. (2017). A re-examination of the mere exposure effect: The influence of repeated exposure on recognition, familiarity, and liking. *Psychological Bulletin*, 143, 459–498.
- 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.
- Newell, B. R., & Shanks, D. R. (2007). Recognizing what you like: Examining the relation between the mere-exposure effect and recognition. *European Journal of Cognitive Psychology*, 19, 103–118.
- 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.
- 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.
- Siegel, P., Selvaggi, S., Sims, V., & Rinck, M. (2019). Social anxiety elicits an approach mere exposure effect for angry faces. *Psychology of Consciousness*, 7, 30–45.

- Smith, P. K., Dijksterhuis, A., & Chaiken, S. (2008). Subliminal exposure to faces and racial attitudes: Exposure to Whites makes Whites like Blacks less. *Journal of Experimental Social Psychology*, 44, 50–64.
- 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.
- Yagi, Y., & Inoue, K. (2010). The contribution of attention to the mere exposure effect for parts of advertising images. *Frontiers in Psychology*, 9, article 1635.
- 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 Cognitive*, 18, 223–251.

# 16 Halo effects

*Simon M. Laham and Joseph P. Forgas*

The halo effect is a general cognitive bias in impression formation. It can be defined as the tendency of judges to assume that if a person possesses some known good or bad characteristics, their other, unrelated, and unknown characteristics are likely to be of similar valence (i.e., good or bad, respectively). 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, in which a glowing circle or “halo” is depicted as floating above the heads of saints indicating their general goodness and elevated 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 but overestimate the correlation between attributes.

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 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 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 judge 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 “typical” philosopher, a middle-aged male, rather than an unconventional young female.

This study also explored whether more attentive processing reduces, and more superficial processing accentuates, the halo effect. Judges were induced into either a positive or a negative mood before forming impressions. It was expected and found that positive mood recruited a more heuristic, superficial, and simplified information-processing strategy that increased the size of 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](http://www.psychologyconcepts.com/halo-effect). A simple classroom demonstration is set out in Text box 16.1.

#### **Text box 16.1 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 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, with about 20–25 persons per condition to produce a reliable effect.

#### **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 manipulation, 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 4-year-old girl. She has beautiful blonde hair and lovely blue eyes.

OR

*Negative halo condition:* Manga is rather unattractive looking 4-year-old girl. She has greasy hair and close set eyes.

The other day when she was playing with a neighbor's little 4-year-old boy, she threw a stone at him which 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.

## **Overview of research on halo effects**

### ***History***

Halo biases were first discussed in any detail in the early 20th 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 inter-correlations 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 ingenuous 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 actively and subconsciously engage 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 best known halo effect is associated with physical attractiveness, first demonstrated by Dion et al. (1972). Dion and her colleagues presented participants with photos of young females who were previously rated as physically attractive, of average attractiveness, or unattractive. Participants who viewed these images were then asked to form impressions about, and rate these targets, on a variety of different characteristics such as their personality, expected future personal and occupation success, life satisfaction, and happiness. Perhaps the most remarkable aspect of this study was that participants were willing and able to perform the task at all. After all, how could a person reasonably judge the personality or intelligence of an individual based only on how they look on a photo? Results showed that, as expected, physically attractive targets were judged to have significantly more socially desirable characteristics on a variety of dimensions, including personality, expected future occupational success, and happiness (see Table 16.1).

*Table 16.1* The effects of a person's physical attractiveness on perceptions of other characteristics

<i>Rated characteristics</i>	<i>Physical appearance of target</i>		
	<i>Unattractive</i>	<i>Average</i>	<i>Attractive</i>
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

Source: Adapted from "What is Beautiful is Good", by K. K. Dion, E. Berscheid, and E. Walster, 1972, *Journal of Personality and Social Psychology*, 24, 288. © 1972 by the American Psychological Association.

*Note:* Higher numbers correspond to more positive judgments.

As Table 16.1 shows, even though attractive 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 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 reflects accumulated past individual and cultural experiences and 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", that is, what they already know and expect about the relationship between different personality traits. Subsequent research also confirmed that the occurrence of halo effects is partly dependent on the particular judgmental dimension studied (e.g.,Forgas et al., 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 et al., 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 16.2). 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!

*Table 16.2* 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
<b>Ratings of essay quality</b>				
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	
<b>Ratings of writer's overall ability</b>				
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	

Source: 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, 302. © 1974 by the American Psychological Association.

*Note:* Higher numbers indicate more positive evaluations on a scale from 1 to 10.

The physical attractiveness halo effect also appears to emerge fairly early in childhood. In a remarkable study, Dion (1973) found that pre-schoolers aged 3 to 6½ 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.

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 16.1 was designed to reproduce very similar results.

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) has shown when assessing judgments of parental competence. Indeed, the “what is beautiful is good” expectation may 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 16.1).

Physical attractiveness is an enduring personal quality, so it is perhaps not surprising that it can exert such a powerful halo effect on impressions in many areas. Would more fleeting and temporary attractiveness signals, such as a 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 16.2). Other and more diffuse “attractiveness” signals can also produce halo effects. For example, Nisbett and 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. In a recent demonstration of the effects of less stable cues on halo effects, McDonald and Ma (2015) showed that children inferred competence on the basis of the formality of a target’s clothes.

The robustness of physical attractiveness halo effects is further illustrated by research suggesting that halo effects may also be transferable to persons associated with the original target. Sigall and Landy (1973) reported that the physical attractiveness of a good-looking

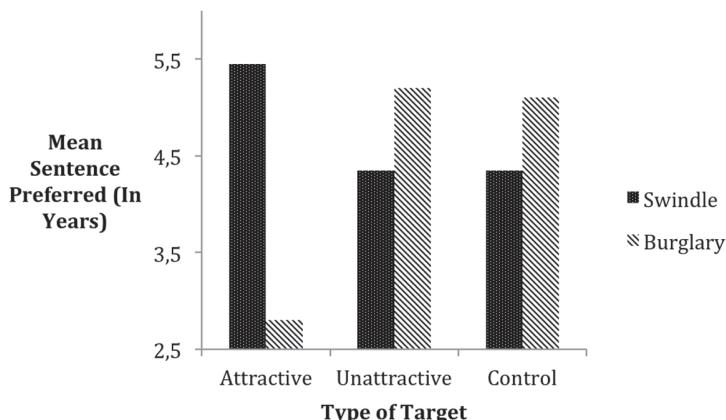


Figure 16.1 The effects of a female's physical attractiveness on sentences for crimes in which attractiveness did (swindle) or did not (burglary) play a role. Figure based on data from Sigall and Ostrove (1975).

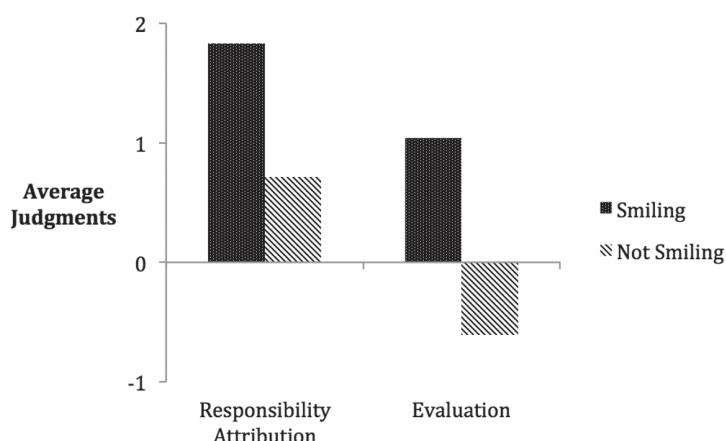


Figure 16.2 The effects of facial expression on responsibility attribution and the evaluation of a person who committed a transgression.

Note: Higher numbers correspond to less responsibility attributed, and more positive evaluation. Figure based on data from Forgas et al. (1983).

female partner exerted a positive influence on impressions about her male consort. Further, a recent study examining halo effects in the consumer domain illustrated a similar cross-target halo whereby reviewer attractiveness influenced brand perception (Ozanne et al., 2019).

#### *Other halo effects*

Several further intriguing demonstrations 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: Estimates of "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 16.1). 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 study by Laham et al. (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 malingeringer, a drug abuser, and suffering from a personality disorder (reported in Gross, 2015).

Recent applications of the halo effect in consumer psychology have examined so-called "health" and "green" halo effects (e.g., Besson et al., 2019; Sorqvist et al., 2015). In these cases, known attributes signaling the healthiness or eco-friendliness of targets (usually products) increase the likelihood of positive evaluations on other attribute dimensions. For example, labeling food "organic" increases perceptions of nutrition and safety, positive brand attitudes and trust (Ellison et al., 2016); labeling food as "eco-friendly" leads people to think that it tastes better (Sorqvist et al., 2015; see also Chapter 24 for further examples).

### ***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 judgment dimension (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, in the Sigall and Ostrove (1975) study mentioned earlier, an attractive woman was judged more leniently when committing a burglary, but

received a more severe sentence when judges thought that she actually used her good looks to commit a swindle.

Findings such as those from Sigall and Ostrove (1975) andForgas (2011) 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 emerge. However, more careful, systematic, and attentive processing of the available information may well reduce halo effects.

### ***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 in the rich array of information available (confirmation bias), and positive impressions may in turn lead to preferential treatment in a range of domains from interpersonal relations, to the workplace, the health and legal systems, and even for decision-making and consumer choices (self-fulfilling prophecy). 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 (Hamermesh & Biddle, 1994).

Of particular relevance to the judicial system are studies that show that attractive perpetrators, in comparison to less attractive ones, are judged less guilty and receive less severe punishments for cheating (Efran, 1974; Forgas et al., 1983). Within the health system, Martin et al. (1977) found that more attractive inpatients suffering from schizophrenia were rated as emotionally better adjusted than were less attractive fellow patients. Further, Napoleon et al. (1980) found that less attractive patients received more severe diagnoses and stayed longer 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 the mechanisms underlying 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, come from. 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 fallback strategy. Several recent experiments suggest that when a more elaborate, systematic, and analytic processing style is triggered (e.g., as a result of induced negative affect), halo effects tend to disappear (Forgas, 2011). More recent and direct evidence of the role of systematic processing in limiting halo effects comes from Wen et al. (2020). These authors found that, when analytic thinking was activated prior to an impression formation task, participants were less influenced in their judgments by central traits, thus demonstrating less susceptibility to the halo effect.

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 casts 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 11, 14, and 15 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”) seems 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 or attributes, rather than 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 suggests that at least some halo effects, as is the case with many other constructive biases, may have an evolutionary origin (e.g., Buss, 2019; Langlois et al., 2000; von Hippel, 2018). The case for evolutionary origins seems strongest for physical attractiveness halos. Physical appearance appears to function as a universally recognized evolutionary signal indicating social desirability and reproductive fitness (Jokela, 2009). Further, the fact that such attractiveness halo effects emerge in early childhood (Dion, 1972) is consistent with such an evolutionary explanation.

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 have achieved limited success in reducing their occurrence. For example, Wetzel et al. (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, manipulations designed to influence the kind of information-processing strategy adopted by judges may well be more effective. As noted above, rather direct attempts to increase systematic processing have been successful in reducing the halo effect (Wen et al., 2020). Less direct interventions aimed at manipulating processing style might also have effects.Forgas (2011) showed that inducing a negative mood (which is associated with a more analytical, systematic, and externally focused processing strategy) eliminated the halo effect.

### **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 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). For an evolutionary approach to understand how adaptive pressures shaped human thinking, see Buss (2019) and von Hippel (2018).

## **References**

- Asch, S. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, 41, 258–290.
- Besson, T., Lalot, F., Bochard, N., Flaudias, V., & Zerhouni, O. (2019). The calories underestimation of “organic” food: Exploring the impact of implicit evaluations. *Appetite*, 137.
- Buss, D. (2019). *Evolutionary psychology: The new science of the mind*. New York: Routledge.
- 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.
- Ellison, B., Duff, B. R. L., Wang, Z., & White, T. B. (2016). Putting the organic label in context: Examining the interactions between the organic label, product type, and retail outlet. *Food Quality and Preference*, 49.
- 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.
- Hamermesh, 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 teacher’s expectations. *Journal of Educational Psychology*, 65, 222–225.

- Jokela, M. (2009). Physical attractiveness and reproductive success in humans: Evidence from the late 20 century United States. *Evolution and Human Behavior*, 30, 342–350.
- 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., Friendmeyer, M. T., & Moore, J. E. (1977). Pretty patient–healthy patient? Study of physical attractiveness and psychopathology. *Journal of Clinical Psychology*, 33, 990–994.
- McDonald, K. P., & Ma, L. (2015). Dress nicer = know more? Young children's knowledge attribution and selective learning based on how others dress. *PLoS ONE*, 10(12), e0144424.
- 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.
- Ozanne, M., Liu, S. Q., & Mattila, A. S. (2019). Are attractive reviewers more persuasive? Examining the role of physical attractiveness in online reviews. *Journal of Consumer Marketing*, 36(6), 728–739.
- 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 juridical judgments. *Journal of Personality and Social Psychology*, 41, 410–414.
- Sörqvist, P., Haga, A., Langeborg, L., Holmgren, M., Wallinder, M., Nöstl, A., Seager, P. B., & Marsh, J. E. (2015). The green halo: Mechanisms and limits of the eco-label effect. *Food Quality and Preference*, 43, 1–9.
- Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 25–29.
- Von Hippel, W. (2018). *The social leap: The new evolutionary science of who we are, where we come from, and what makes us*. New York: Harper Collins.
- 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.
- Wen, W., Li, J., Georgiou, G. K., Huang, C., & Wang, L. (2020). Reducing the halo effect by stimulating analytic thinking. *Social Psychology*, 51(5), 334–340.
- 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.

# 17 Assumed similarity

*Isabel Thielmann and Benjamin E. Hilbig*

In everyday social interactions, people regularly need to judge what other people are like, even if these others are complete strangers. For example, one may want to buy a used car and thus seek to judge the seller's trustworthiness; one may be looking for a new roommate and thus seek to judge their orderliness and sociability; or one may see a new doctor and seek to judge their thoroughness. Arguably, being able to *accurately* judge other people is adaptive because correctly predicting their behaviors allows for according adjustment of one's own actions (Fiske, 1993; Zebrowitz & Montepare, 2006). Interestingly, a consistent observation from diverse research across the social sciences suggests that individuals to some extent rely on *their own* characteristics when judging others, in the sense of perceiving others – including random strangers – to be somewhat similar to them. This ubiquitous *assumed similarity* bias in person perception is the focus of this chapter.

## The phenomenon of assumed similarity

### **Definition**

Assumed similarity denotes the convergence between how judges (i.e., *perceivers*) see themselves and how they see other (i.e., *target*) individuals (Cronbach, 1955; see Figure 17.1 and Text box 17.1). This convergence can occur for any kind of trait, attitude, or behavior (broadly referred to as *characteristics* in what follows). Crucially, assumed similarity does not merely reflect a perceiver's accurate perception of actual similarity with the target (Kenny & Acitelli, 2001; Lee et al., 2009; Ready et al., 2000). Thus, some authors have explicitly defined assumed similarity as the congruence between a perceiver's self-view and their perceptions of others *over and above* actual similarity (e.g., Human & Biesanz, 2011) – as we also do here. This reflects the notion that assumed similarity is a judgment *bias* or similarity *illusion*, respectively.

### **Text box 17.1 Definition of relevant terms**

*Assumed similarity:* Convergence between a perceiver's self-view on a characteristic and their judgment of others on that same characteristic that cannot be accounted for by an accurate perception of actual similarity. For example, Paul may ascribe a similar level of trustworthiness to himself and a random stranger named Tina (i.e., convergence between A and B in Figure 17.1).

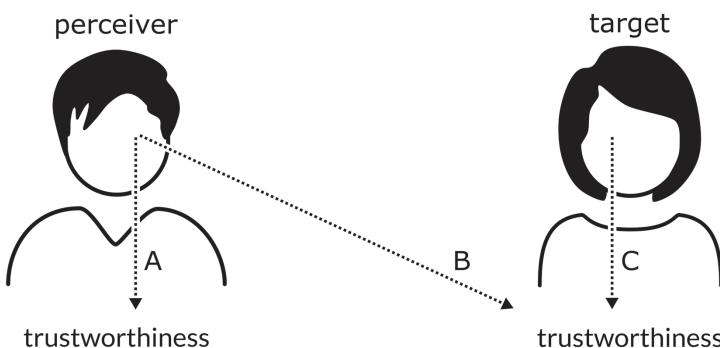


Figure 17.1 Graphical illustration of different person perception phenomena.

Note: Assumed similarity is represented by the convergence between A and B. Actual similarity is represented by the convergence between A and C. Accuracy (i.e., self-other agreement) is represented by the convergence between B and C.

*Actual similarity:* Convergence between a perceiver's level on a characteristic and a target's level on that same characteristic. For example, Paul and Tina may both have a high level on trustworthiness and thus actually be similar. As a proxy for actual similarity, research usually relies on the convergence between a perceiver's and a target's self-report on the same characteristic, that is, how Paul and Tina view themselves on trustworthiness (i.e., convergence between A and C in Figure 17.1).

*Accuracy (i.e., self-other agreement):* Convergence between a perceiver's perception of a target on a characteristic and the target's actual level on that characteristic – again, typically operationalized through the target's self-view. For example, Paul's judgment of Tina's trustworthiness may be accurate to the extent that it converges with how trustworthy Tina sees herself as a proxy for how trustworthy she is (i.e., convergence between B and C in Figure 17.1).

Besides assumed similarity, various other terms have been proposed and used in the literature to refer to the convergence between a perceiver's self-view and their judgment of others. Although all these terms differ with regard to the mechanisms suggested to produce the bias, operationally they all refer to the same phenomenon. Most prominently, assumed similarity is often called *social projection*, suggesting “projection ... as the process by which people come to believe that others are similar to them” (Krueger, 2007, p. 2). Thus, whereas the term assumed similarity implies a somewhat conscious process of “assuming” similarity with others, the term social projection implies a largely unconscious process. Similarly, the *self-based heuristic* is explicitly defined as “an unconscious, semiautomatic, and variable process that is relied upon when a rater is faced with a difficult rating task” (Ready et al., 2000, p. 222). Thus, in addition to proposing the process to be unconscious, the self-based heuristic further adds the premise that perceivers will particularly rely on their own characteristics when judging others if they have insufficient information about the target. The term *attributive projection* (Holmes, 1968), in turn, describes the situation where perceivers are consciously aware of their own characteristics

that are projected onto others, while consciousness of the projection process itself is not required. Another concept closely linked to assumed similarity is the *false consensus effect* which describes individuals' tendency to "see their own behavioral choices and judgments as relatively common and appropriate to existing circumstances" (Ross et al., 1977, p. 280). As indicated by its name, the false consensus effect implies a judgment "error" resulting from illegitimately assuming one's own characteristics as being widely (consensually) shared. Finally, *self-anchoring* specifically refers to judgments of targets from one's in-group, and it is defined as the "tendency to base in-group judgments on the self" (Cadinu & Rothbart, 1996, p. 661).

Taken together, many terms other than assumed similarity have been used to describe essentially the same phenomenon while (implicitly or explicitly) suggesting different underlying mechanisms. Here, we will exclusively use the term "assumed similarity" in an operational sense to refer only to the observed convergence between perceivers' self-views and their perception of others above and beyond actual similarity – without making any assumptions about the underlying psychological processes at this point.

### **Measures**

A common approach to demonstrate the phenomenon of assumed similarity is to ask perceivers to rate both themselves and a specific target person – which may be the same for all perceivers or differ between perceivers – on a certain characteristic and to compute the correlation between perceivers' self-ratings and their ratings of the target on that same characteristic. One may also ask perceivers to judge how they think the target will behave in a specific situation and correlate these *beliefs* about the target's behavior with perceivers' behavior in the same situation. Targets may, in turn, either be known to perceivers (e.g., friends, family) or unknown to them and, for example, be presented on photos (see Text box 17.2), videos, or via verbal description. However, especially if perceivers judge a target they know, actual similarity between perceiver and target needs to be accounted for statistically to isolate the bias aspect of assumed similarity, which is not an issue if assumed similarity is studied among random strangers (Thielmann et al., 2020a). Targets may also be representatives of a certain group (e.g., a typical student or the average participant in a study) which may, however, induce what is called *spurious similarity* (Paunonen & Hong, 2013). Spurious similarity denotes the convergence between perceivers' self-views and their perceptions of others based on a joint group membership. Specifically, when perceiver and target share a certain group membership (e.g., their gender, age, or profession), perceivers may simply base their judgment on this group membership, without any reference to the self. For example, if both Paul and Tina are students, Paul may judge Tina to be trustworthy based on the stereotype that students – or indeed simply in-group members – are trustworthy, rather than on his own (high) level of trustworthiness.

A straightforward way to diminish such influences of spurious similarity on assumed similarity is to ask perceivers to rate multiple targets each, which – preferably – differ with regard to certain group memberships (e.g., gender or age groups). This can, for example, be achieved in *round-robin designs* in which each perceiver in a group judges every other participant in the group. Thus, by design, participants are both perceivers and targets. The group may be randomly brought together in a lab session, and it is optimally composed of participants that are strangers to each other. Another method assessing judgments of multiple targets while dispensing with real-time social interaction is the *half-block design*

(Kenny, 1994) in which all perceivers rate the same standardized set of targets with regard to the same characteristics. For example, in the Online-Tool for Assessing Perceiver Effects (O-TAPE; Rau et al., 2021b), perceivers are presented with ten standardized Facebook profiles, which have been designed to include targets that cover both sexes equally, span a certain age range, and feature varying levels of attractiveness and expressiveness.

In general, besides counteracting influences of spurious similarity, having perceivers rate multiple (i.e.,  $\geq 3$ ) targets on the same characteristics has the advantage that the variance in judgments can be decomposed using the Social Relations Model (SRM; Kenny, 1994; Kenny & La Voie, 1984). The SRM differentiates between *perceiver effects* (i.e., how perceivers generally perceive others), *target effects* (i.e., how targets are generally perceived by others), and *relationship effects* (i.e., dyad-specific effects existing above and beyond the perceiver and target effects) and thereby allows for a particularly fine-grained analysis of other-perceptions. Using the SRM, assumed similarity is reflected in a positive relation between perceivers' self-views and their perceiver effects. Alternatively, one may simply aggregate across all of a perceiver's ratings of a characteristic and relate this aggregate measure to the perceivers' self-view of that characteristic, which has been shown to yield highly comparable assumed-similarity correlations as the SRM approach (De Vries, 2010).

Another distinction in the operationalization of assumed similarity is the *trait-based approach* versus *profile-based approach* (for an overview, see, e.g., Back & Nestler, 2016). Using a trait-based approach, assumed similarity refers to the correspondence between perceivers' self-reports on a given characteristic and their perceptions of the target(s) on that same characteristic. By implication, if multiple characteristics are under scrutiny in a study, assumed similarity is determined for each characteristic separately. Using a profile-based approach, by contrast, assumed similarity refers to the correspondence in the *pattern* of traits in perceivers' ratings of themselves and the targets (Human & Biesanz, 2011). Thus, assumed similarity refers to a set of traits and is defined by the similarity of profiles in self-views and other-perceptions (Cronbach & Gleser, 1953; Furr & Wood, 2013). In general, assumed similarity is consistently observed using both the trait-based and the profile-based approach. However, trait-specific differences in assumed similarity can only be detected using the former.

## **Text box 17.2 A classroom demonstration of assumed similarity**

### **Participants**

Assumed similarity correlations are approximately medium-sized on average. To be able to detect  $r = .30$  with sufficient power ( $1-\beta = .80$ ) and a conventional  $\alpha$  of .05 (one-tailed, given that assumed similarity implies a positive correlation), 67 participants are required. Because classes are rarely that large, we recommend interpretation of the effect size rather than whether the effect is statistically significant.

### **Materials**

Particularly robust assumed similarity effects are found for judgments of the personality trait Honesty-Humility from the HEXACO model of personality structure (Thielmann et al., 2020a). We will thus focus on this trait in the experiment.

Self-report and observer report forms of the HEXACO-60 questionnaire (Ashton & Lee, 2009) are freely available for academic use in various languages on <http://hexaco.org/hexaco-inventory> (see Appendix for English versions). As target stimuli, two random photos of physically attractive individuals (one female, one male) should be used. The selection of attractive targets is likely to increase assumed similarity (see below); the rating of strangers ought to counteract influences of actual similarity; and the rating of two targets of different sexes ought to counteract influences of spurious similarity.

### **Procedure**

First, participants provide self-ratings on the ten Honesty-Humility items on a scale from 1 = *strongly disagree* to 5 = *strongly agree*. Next, participants are presented with one of the two targets and are asked to rate the target using the same ten Honesty-Humility items (now in observer report form). This procedure is subsequently repeated for the second target. To avoid participants answering the same items three times in a row, assessment of self- and observer reports can be separated in time (e.g., by assessing self-reports at the beginning of the class and observer reports at the end, or by having one week in between).

### **Analysis**

First, all reverse-keyed items from the self- and observer reports must be recoded, such that high scores indicate high Honesty-Humility levels for all items. Next, for each participant, average scores across all items are computed, separately for the self-report and each of the two observer reports; then, the two observer report means can be aggregated. Assumed similarity is computed as the zero-order correlation between participants' self-report and the (aggregate) observer report. A positive correlation indicates assumed similarity.

### **Assumed similarity and accuracy**

As defined above, assumed similarity denotes a judgment bias that exists over and above accurate perceptions of actual similarity. One may thus argue that assumed similarity should diminish judgment accuracy to the extent that perceivers "overproject" their own characteristics onto others. Supporting this reasoning, it has been shown that, when examined across traits, assumed similarity is inversely related to self-other agreement (Beer & Watson, 2008; Human & Biesanz, 2012; Watson et al., 2000) – and thus accuracy (see Text box 17.1). That is, those traits that are typically judged with higher accuracy tend to yield weaker assumed similarity, and vice versa (but see Kenny & West, 2010). However, the picture changes when examining the relation between assumed similarity and accuracy across perceivers. Specifically, perceivers' ability to judge others accurately is largely independent of their tendency to assume similarity with others (Human & Biesanz, 2011, 2012). By implication, any one perceiver may accurately judge others despite being prone to assumed similarity.

## Moderators of assumed similarity

### *Characteristics of targets*

Assumed similarity has been shown for various types of relationships, including romantic partners (Beer et al., 2013; Liu et al., 2018b), co-workers (Cohen et al., 2013), roommates (Paunonen & Hong, 2013), and even strangers (Beer & Watson, 2008; Thielmann et al., 2020a). Nonetheless, assumed similarity typically increases with a higher relationship closeness (Kenny, 1994; Lee et al., 2009; Selfhout et al., 2009; but see Kenny & West, 2010). One may argue that this effect is due to perceivers actually sharing more characteristics with targets to whom they feel close. However, even when controlling for actual similarity, the positive effect of closeness on assumed similarity remains (Lee et al., 2009; Selfhout et al., 2009). Thus, higher actual similarity cannot fully account for stronger assumed similarity with increasing relationship closeness.

The observation that assumed similarity is more pronounced when perceivers judge close others also aligns more broadly with evidence showing that assumed similarity increases with increasing attraction towards the target. Interpersonal attraction denotes “a positive attitude or evaluation regarding a particular person” (Aron & Lewandowski, 2001, p. 7860) and it is often treated as equivalent to liking (see Chapter 16 on the halo effect, which describes the positive main effect of attraction on perceptions of others). In turn, irrespective of whether perceivers judge acquainted or unacquainted targets, assumed similarity is stronger – or indeed only apparent at all – when perceivers judge targets they do not dislike (Collisson & Howell, 2014; Davis, 2017; Human & Biesanz, 2011; Locke et al., 2012; Weller & Watson, 2009). Similarly, assumed similarity is more pronounced when perceivers judge politicians from a preferred rather than a non-preferred party (De Vries & van Prooijen, 2019), or when they judge interaction partners they are going to cooperate rather than compete with in the future (Riketta & Sacramento, 2008; Toma et al., 2010). Further, experimental manipulation of attraction via evaluative conditioning or emotional expressions of targets yielded more pronounced assumed similarity for judgments of positive targets, while revealing assumed *dissimilarity* for judgments of negative targets (Machunksky et al., 2014).

There are different explanations for why liking moderates assumed similarity. First, assuming similarity with likable targets may facilitate self-enhancement because it emphasizes one’s commonalities with valued others (Marks & Miller, 1987). Second, individuals are generally motivated to maintain cognitive balance (Heider, 1958), which may be achieved by ascribing one’s own characteristics to likable others. Specifically, perceivers – at least those holding a positive self-view – should achieve cognitive balance by seeing favorable characteristics as being shared by favorable individuals (i.e., themselves and likable others). Finally, assuming similarity with likable others may increase feelings of belongingness to and communion with valued others (Machunksky et al., 2014; Morrison & Matthes, 2011). Taken together, all these explanations suggest that assumed similarity may at least partly be attributable to a *motivation* of perceivers to assume similarity with specific others.

### *Characteristics of perceivers*

Besides characteristics of the target, assumed similarity is also influenced by characteristics of perceivers. Indeed, there are strong individual differences in assumed similarity (Human

& Biesanz, 2011). Evidence suggests stronger assumed similarity among perceivers with higher self-esteem (Human & Biesanz, 2011; Locke et al., 2012). Conversely, perceivers experiencing chronic negative affect tend to assume less similarity with others on positive characteristics (Lane & Gibbons, 2007; Moss et al., 2007), but more similarity on negative characteristics (i.e., negative emotions; Papp et al., 2010). This is compatible with the idea of assumed similarity as a manifestation of individuals striving for cognitive balance: Perceivers with a positive self-view see others as sharing their positive characteristics whereas perceivers with a negative self-view see others as sharing their negative characteristics.

Moreover, evidence suggests that perceivers with a higher communal orientation – the tendency to care for and cooperate with others as well as to behave morally (Abele & Wojciszke, 2014) – show stronger assumed similarity, but only for ratings of liked others, romantic partners, and in-group members (Locke et al., 2012). Likewise, women – who, on average, have a higher communal orientation (Moshagen et al., 2019; Moskowitz et al., 1994) – as well as individuals from collectivistic cultures tend to be more prone to assumed similarity than men and individuals from individualistic cultures (Locke et al., 2013; Ott-Holland et al., 2014). Overall, these findings are in line with the idea that assumed similarity may serve to increase feelings of belongingness with (valued) others.

### ***Characteristics to be judged***

Assumed similarity has been shown for a notable variety of characteristics, including political preferences (Locke et al., 2012), vocational interests (Holtrop et al., 2018), communication styles (Mathison, 1988), and state and trait affect (Thomas et al., 1997; Watson et al., 2000). Most prior research, however, has studied assumed similarity of basic personality traits, such as the Big Five (i.e., Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness; Goldberg, 1990; McCrae & Costa, 1987) or the HEXACO dimensions (e.g., Cohen et al., 2013; Lee et al., 2009; Srivastava et al., 2010; Thielmann et al., 2020a; Wood et al., 2010).<sup>1</sup> Crucially, according to a recent meta-analytic review, assumed similarity is only apparent for some of these basic personality dimensions, but not for others (Thielmann et al., 2020a). Specifically, among the Big Five, only Agreeableness and Openness to Experience yield reliable assumed similarity effects. Among the HEXACO dimensions, in turn, Honesty-Humility and Openness to Experience are the only dimensions showing consistent evidence for assumed similarity. Strikingly, exactly these dimensions are the ones that are most strongly linked to values (Anglim et al., 2017; Fischer & Boer, 2015), giving rise to the value account of assumed similarity as detailed below.

## **Theoretical accounts**

### ***Lack-of-information account***

One account of assumed similarity argues that perceivers will particularly rely on their own characteristics in judgments of others when they have insufficient information about the target available (Kenny & West, 2010; Ready et al., 2000; Watson et al., 2000). According to this idea, assumed similarity essentially reflects a “lack of information” effect. However, several empirical findings are in conflict with this account. First, as summarized above,

assumed similarity typically increases with increasing relationship closeness between perceiver and target (e.g., Lee et al., 2009; Selfhout et al., 2009) and thus, becomes more pronounced as perceivers have *more* rather than less information available about the target. In fact, even in the absence of any trait-relevant information about the target, assumed similarity is only apparent for *some* characteristics (Thielmann et al., 2020a), whereas the lack-of-information account dictates that assumed similarity should be ubiquitous in such situations. Second, assumed similarity is not stronger for characteristics that are less easily observed or judged and for which lack of information is thus more prevalent. For example, Big Five Neuroticism – a trait that is mostly expressed in one's thoughts and feelings and that is thus relatively difficult to observe from the outside – does not yield stronger assumed similarity effects than Big Five Extraversion (Thielmann et al., 2020a) – which is mostly expressed in one's observable behaviors (Connelly & Ones, 2010; Funder & Dobroth, 1987) and is thus relatively easy to observe. Taken together, the empirical picture is largely incompatible with the lack-of-information account.

### **Value account**

An alternative view of assumed similarity that intends to explain assumed similarity effects for certain traits in particular is provided by the value account (Lee et al., 2009). According to this account, a characteristic's link to values determines that characteristic's susceptibility to assumed similarity, in the sense that a stronger link to values should increase assumed similarity. Values denote “moral, social, or aesthetic principle(s) accepted by an individual or society as a guide to what is good, desirable, or important” (VandenBos, 2007), and they are most prominently captured by Schwartz's taxonomy of basic values (Schwartz, 1992; Schwartz et al., 2012). Lee and colleagues (2009) suggested that “because values are presumably an important part of one's identity, one would likely tend to ... assume one's own values to be shared by persons with whom one has a close social relationship” (p. 464). This proposition was supported by the findings that when perceivers judged close others (i) the two HEXACO dimensions exhibiting the strongest links to values (i.e., Honesty-Humility and Openness to Experience) were the only ones yielding reliable assumed similarity effects and (ii) a similar pattern of assumed similarity emerged for the two basic axes of values, that is, Self-Transcendence versus Self-Enhancement and Openness to Change versus Conservation.

Importantly, the value account does not only seem to provide a useful explanation of trait-specific assumed similarity for judgments of close others, but also for judgments of strangers. Across a set of studies in which perceivers judged the personality of strangers, assumed similarity was again only apparent for Honesty-Humility and Openness to Experience, but not for any other personality dimension. This suggests that individuals do not only want their values to be shared by others they know, but by others *in general* – which is, again, arguably attributable to the high importance of values for individuals' identity. However, effect sizes were descriptively weaker as compared to judgments of close others and only consistent for Honesty-Humility across studies, but not for Openness to Experience (Thielmann et al., 2020a). Beyond basic personality traits, in turn, evidence on assumed similarity of political preferences (Locke et al., 2012) – which are closely tied to values – also aligns with the value account. Finally, the value account can explain why assumed similarity is more pronounced for judgments of targets one feels close and/or attracted to: Perceivers will arguably consider it particularly important that favorable others share their values.

### **Global positivity**

Questioning the existence of “true” assumed similarity, some authors have argued that assumed similarity may be an artifact of individual differences in how positively or negatively perceivers see others in general (Wood et al., 2010). Specifically, perceivers differ with regard to how positively they judge others (Rau et al., 2021a), and these individual differences in *global positivity* are related to certain (benevolent) characteristics in perceivers (Rau et al., 2021b; Wood et al., 2010). For example, a highly agreeable perceiver may simply judge others as highly agreeable because the perceiver generally tends to see others more positively – and thus also as more agreeable – than a less agreeable perceiver. In line with this reasoning, Wood et al. (2010) found that perceivers’ self-reports on Big Five Agreeableness not only relate positively to their ratings of targets’ Agreeableness, but also to their ratings of other socially desirable characteristics in targets. The authors therefore concluded that “a single factor concerning how positively others are perceived is sufficient to capture most of the covariation in how individuals tend to see others across a broad range of traits” (Wood et al., 2010, p. 183).

However, other studies have shown that assumed similarity cannot be sufficiently accounted for by individual differences in global positivity of other-perceptions. Using a round-robin design, Srivastava and colleagues (2010) showed that assumed similarity among the Big Five remained meaningful even after statistically controlling for global positivity in perceiver effects. This finding was recently corroborated using a half-block design in which perceivers rated ten unknown targets (using the O-TAPE measure; Rau et al., 2021b) on the HEXACO dimensions (Thielmann et al., 2020b). Although assumed similarity correlations considerably decreased once accounting for global positivity in perceiver effects, they remained robust for Honesty-Humility and Openness to Experience. Moreover, using a profile-based approach, the authors found that assumed similarity increased with the judged traits’ value-relatedness, even after accounting for the traits’ social desirability (i.e., positivity). Finally, a meta-analytic review of round-robin studies revealed a null relation between the evaluativeness of characteristics (i.e., their valence or desirability) and assumed similarity (Kenny & West, 2010), meaning that more desirable characteristics did not yield stronger assumed similarity as would be implied by the global-positivity reasoning. Overall, the current state of research thus concurs with the notion that “assumed similarity reflects more than just broad evaluative tone” (Srivastava et al., 2010, p. 526).

### **Dynamic perception model**

The dynamic perception model (DPM; Hughes et al., 2020) draws on social feedback effects in interpersonal perception to account for assumed similarity. According to the DPM, assumed similarity may result from perceivers eliciting certain behaviors in targets that are then (accurately) perceived by the perceivers. Specifically, in social interactions, perceivers’ personality influences their behavior (i.e., *trait expression*), which likely elicits certain behaviors in targets (i.e., *behavior elicitation*). Especially in the affiliative (communal) domain, perceivers’ behavior should elicit concordant behaviors in targets (Kiesler, 1983). For example, friendly behavior usually encourages friendly responses (Markey et al., 2003; Sadler & Woody, 2003). These behaviors, in turn, provide certain cues that perceivers can use to form a judgment about the target (i.e., *circumscribed accuracy*). Thus, the DPM proposes that perceivers may not only perceive targets as similar to themselves due to

projecting their own (preeexisting) characteristics on the targets (termed *trait assumed similarity* in the DPM), but also due to accurately perceiving the behaviors elicited in targets in a specific interaction (i.e., *perceiver-elicited similarity*). In essence, this aligns with the crucial distinction between assumed similarity and accuracy, both of which may contribute to perceiving similarity with others. Although an empirical test of the DPM suggested trait assumed similarity to have a stronger effect on perceivers' judgments of targets than perceiver-elicited similarity, the latter may facilitate the development of trait assumed similarity over time. As Hughes et al. (2020) argue, "it is possible that trait assumed similarity is the cumulative product of many instances of interpersonally elicited and perceived similarity" (p. 14). By and large, this perspective provides a fruitful avenue for future research, which should also critically test whether the DPM can explain the apparent trait-specific assumed-similarity effects in line with the value account. For example, it is conceivable that especially the expression of one's values encourages concordant behaviors in others, thereby being one source through which value-related characteristics may become prone to assumed similarity.

## Practical relevance

As hinted in our opening paragraph, there are many real-life situations in which individuals need to judge others, including strangers, and in which these judgments about others' characteristics influence individuals' own behavior. We will elaborate here on a few prominent examples. First, perceptions of others' cooperativeness and trustworthiness are key drivers of prosocial behavior, such as cooperation (Balliet & Van Lange, 2013). Mutual cooperation is often required to maximize social welfare, whether it be in micro-level interactions with colleagues or friends or in regard to macro-level issues such as mitigating climate change and solving intergroup conflict (Thielmann et al., 2021). By implication, assuming similarity in these situations may have far-reaching consequences for the quality of interpersonal relationships and for the functioning of societies at large by influencing perceivers' own willingness to behave in a prosocial (e.g., caring, honest, environmentally friendly) manner. Second, assumed similarity may affect partner choice. Research suggests that individuals want to have partners who share their values (Liu et al., 2018a). In turn, the findings that assumed similarity is most prevalent (i) for value-related characteristics and (ii) when perceivers judge targets they feel attracted to suggest that individuals may overestimate the suitability of attractive targets as partners in particular. Third and finally, assumed similarity may distort interviewers' assessment of candidates during a job interview, and this may – again – particularly hold when interviewers feel attracted to a candidate (Mathison, 1988). Thus, the interviewers' characteristics may have severe consequences for the selection process.

## Summary

- Assumed similarity is apparent when perceivers' self-views converge with their judgments of others beyond actual similarity between perceiver and target.
- Assumed similarity affects judgments of well-acquainted others and strangers alike, albeit more strongly so for targets one feels close and/or attracted to.
- There are individual differences in assumed similarity. Perceivers high in self-esteem and communal orientation are more likely to see their positive characteristics as being shared by others.

- Assumed similarity is not apparent for all characteristics alike, but particularly for those that are more closely tied to values.
- The value account provides the most promising explanation for trait-specific assumed similarity so far, suggesting that people want to assume that others share their values in particular.

## Note

- 1 The HEXACO model (Ashton & Lee, 2007) is an extension and variation of the Big Five that adds Honesty-Humility as a sixth dimension and also incorporates some conceptual changes to Emotionality (the counterpart of Big Five Neuroticism) and Agreeableness (for details, see Ashton et al., 2014).

## Further reading

Early work on assumed similarity is provided by Cronbach (1955) and Holmes (1968). Robbins and Krueger (2005) offer a meta-analysis on differences in assumed similarity for judgments of in-group versus out-group members. A meta-analytic review on assumed similarity of personality traits (and related concepts) in round-robin designs is provided by Kenny and West (2010). For a recent meta-analytic summary of assumed similarity of basic personality traits (as well as a comparative test of theoretical accounts of assumed similarity), we refer interested readers to Thielmann et al. (2020a).

## References

- Abele, A. E., & Wojciszke, B. (2014). Communal and agentic content in social cognition: A dual perspective model. In M. P. Zanna & J. M. Olson (Eds.), *Advances in experimental social psychology* (Vol. 50, pp. 195–255). New York: Academic Press.
- Anglim, J., Knowles, E. R. V., Dunlop, P. D., & Marty, A. (2017). HEXACO personality and Schwartz's personal values: A facet-level analysis. *Journal of Research in Personality*, 68, 23–31.
- Aron, A., & Lewandowski, G. (2001). Psychology of interpersonal attraction. In N. J. Smelser & P. B. B. T. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 7860–7862). Amsterdam: Pergamon.
- Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, 11(2), 150–166.
- Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, 91(4), 340–345.
- Ashton, M. C., Lee, K., & DeVries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, 18(2), 139–152.
- Back, M. D., & Nestler, S. (2016). Accuracy of judging personality. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately* (pp. 98–124). Cambridge: Cambridge University Press.
- Balliet, D., & Van Lange, P. A. M. (2013). Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin*, 139(5), 1090–1112.
- Beer, A., & Watson, D. (2008). Personality judgment at zero acquaintance: Agreement, assumed similarity, and implicit simplicity. *Journal of Personality Assessment*, 90(3), 250–260.
- Beer, A., Watson, D., & McDade-Montez, E. (2013). Self-other agreement and assumed similarity in neuroticism, extraversion, and trait affect: Distinguishing the effects of form and content. *Assessment*, 20(6), 723–737.

- Cadinu, M. R., & Rothbart, M. (1996). Self-anchoring and differentiation processes in the minimal group setting. *Journal of Personality and Social Psychology, 70*(4), 661–677.
- Cohen, T. R., Panter, A. T., Turan, N., Morse, L., & Kim, Y. (2013). Agreement and similarity in self-other perceptions of moral character. *Journal of Research in Personality, 47*(6), 816–830.
- Collisson, B., & Howell, J. L. (2014). The liking-similarity effect: Perceptions of similarity as a function of liking. *Journal of Social Psychology, 154*(5), 384–400.
- Connnelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*(6), 1092–1122.
- Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin, 52*(3), 177–193.
- Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*(6), 456–473.
- Davis, M. H. (2017). Social projection to liked and disliked targets: The role of perceived similarity. *Journal of Experimental Social Psychology, 70*, 286–293.
- De Vries, R. E. (2010). Lots of target variance: An update of SRM using the HEXACO personality inventory. *European Journal of Personality, 24*(3), 169–188.
- De Vries, R. E., & van Prooijen, J.-W. (2019). Voters rating politicians' personality: Evaluative biases and assumed similarity on honesty-humility and openness to experience. *Personality and Individual Differences, 144*, 100–104.
- Fischer, R., & Boer, D. (2015). Motivational basis of personality traits: A meta-analysis of value-personality correlations. *Journal of Personality, 83*(5), 491–510.
- Fiske, S. T. (1993). Social cognition and social perception. *Annual Review of Psychology, 44*(1), 155–194.
- Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology, 52*(2), 409–418.
- Furr, R. M., & Wood, D. (2013). On the similarity between exchangeable profiles: A psychometric model, analytic strategy, and empirical illustration. *Journal of Research in Personality, 47*(3), 233–247.
- Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216–1229.
- Heider, F. (1958). *The psychology of interpersonal relations*. Chichester: John Wiley.
- Holmes, D. S. (1968). Dimensions of projection. *Psychological Bulletin, 69*(4), 248–268.
- Holtrop, D., Born, M. P., & De Vries, R. E. (2018). Perceptions of vocational interest: Self-and other-reports in student-parent dyads. *Journal of Career Assessment, 26*(2), 258–274.
- Hughes, B. T., Flournoy, J. C., & Srivastava, S. (2020). Is perceived similarity more than assumed similarity? An interpersonal path to seeing similarity between self and others. *Journal of Personality and Social Psychology, 121*(1), 184–200.
- Human, L. J., & Biesanz, J. C. (2011). Through the looking glass clearly: Accuracy and assumed similarity in well-adjusted individuals' first impressions. *Journal of Personality and Social Psychology, 100*(2), 349–364.
- Human, L. J., & Biesanz, J. C. (2012). Accuracy and assumed similarity in first impressions of personality: Differing associations at different levels of analysis. *Journal of Research in Personality, 46*(1), 106–110.
- Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford Press.
- Kenny, D. A., & Acitelli, L. K. (2001). Accuracy and bias in the perception of the partner in a close relationship. *Journal of Personality and Social Psychology, 80*(3), 439–448.
- Kenny, D. A., & La Voie, L. (1984). The social relations model. *Advances in Experimental Social Psychology, 18*, 141–182.
- Kenny, D. A., & West, T. V. (2010). Similarity and agreement in self-and other perception: A meta-analysis. *Personality and Social Psychology Review, 14*(2), 196–213.
- Kiesler, D. J. (1983). The 1982 Interpersonal Circle: A taxonomy for complementarity in human transactions. *Psychological Review, 90*(3), 185–214.

- Krueger, J. I. (2007). From social projection to social behaviour. *European Review of Social Psychology*, 18, 1–35.
- Lane, D. J., & Gibbons, F. X. (2007). Am I the typical student? Perceived similarity to student prototypes predicts success. *Personality and Social Psychology Bulletin*, 33(10), 1380–1391.
- Lee, K., Ashton, M. C., Pozzebon, J. A., Visser, B. A., Bourdage, J. S., & Ogunfowora, B. (2009). Similarity and assumed similarity in personality reports of well-acquainted persons. *Journal of Personality and Social Psychology*, 96(2), 460–472.
- Liu, J., Ludeke, S. G., Haubrich, J., Gondan, M., & Zettler, I. (2018a). Similar to and/or better than oneself? Singles' ideal partner personality descriptions. *European Journal of Personality*, 32(4), 443–458.
- Liu, J., Ludeke, S. G., & Zettler, I. (2018b). Assumed similarity in personality within intimate relationships. *Personal Relationships*, 25(3), 316–329.
- Locke, K. D., Craig, T., Baik, K.-D., & Gohil, K. (2012). Binds and bounds of communion: Effects of interpersonal values on assumed similarity of self and others. *Journal of Personality and Social Psychology*, 103(5), 879–897.
- Locke, K. D., Zheng, D., & Smith, J. (2013). Establishing commonality versus affirming distinctiveness: Patterns of personality judgments in China and the United States. *Social Psychological and Personality Science*, 5(4), 389–397.
- Machunsky, M., Toma, C., Yzerbyt, V., & Corneille, O. (2014). Social projection increases for positive targets: Ascertaining the effect and exploring its antecedents. *Personality and Social Psychology Bulletin*, 40(10), 1373–1388.
- Markey, P. M., Funder, D. C., & Ozer, D. J. (2003). Complementarity of interpersonal behaviors in dyadic interactions. *Personality and Social Psychology Bulletin*, 29(9), 1082–1090.
- Marks, G., & Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin*, 102(1), 72–90.
- Mathison, D. L. (1988). Assumed similarity in communication styles: Implications for personnel interviews. *Group & Organization Studies*, 13(1), 100–110.
- McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52(1), 81–90.
- Morrison, K. R., & Matthes, J. (2011). Socially motivated projection: Need to belong increases perceived opinion consensus on important issues. *European Journal of Social Psychology*, 41(6), 707–719.
- Moshagen, M., Thielmann, I., Hilbig, B. E., & Zettler, I. (2019). Meta-analytic investigations of the HEXACO Personality Inventory(-Revised): Reliability generalization, self-observer agreement, intercorrelations, and relations to demographic variables. *Zeitschrift für Psychologie*, 227(3), 186–194.
- Moskowitz, D. S., Suh, E. J., & Desaulniers, J. (1994). Situational influences on gender differences in agency and communion. *Journal of Personality and Social Psychology*, 66(4), 753–761.
- Moss, S. A., Garivaldis, F. J., & Toukhsati, S. R. (2007). The perceived similarity of other individuals: The contaminating effects of familiarity and neuroticism. *Personality and Individual Differences*, 43(2), 401–412.
- Ott-Holland, C. J., Huang, J. L., Ryan, A. M., Elizondo, F., & Wadlington, P. L. (2014). The effects of culture and gender on perceived self-other similarity in personality. *Journal of Research in Personality*, 53, 13–21.
- Papp, L. M., Kouros, C. D., & Cummings, E. M. (2010). Emotions in marital conflict interactions: Empathic accuracy, assumed similarity, and the moderating context of depressive symptoms. *Journal of Social and Personal Relationships*, 27(3), 367–387.
- Paunonen, S. V., & Hong, R. Y. (2013). The many faces of assumed similarity in perceptions of personality. *Journal of Research in Personality*, 47(6), 800–815.
- Rau, R., Carlson, E. N., Back, M. D., Barranti, M., Gebauer, J. E., Human, L. J., Leising, D., & Nestler, S. (2021a). What is the structure of perceiver effects? On the importance of global positivity and trait-specificity across personality domains and judgment contexts. *Journal of Personality and Social Psychology*, 120(3), 745–764.

- Rau, R., Nestler, W., Dufner, M., & Nestler, S. (2021b). Seeing the best or worst in others: A measure of generalized other-perceptions. *Assessment*, 28, 1897–1914.
- Ready, R. E., Clark, L. A., Watson, D., & Westerhouse, K. (2000). Self- and peer-related personality: Agreement, trait ratability, and the “self-based heuristic”. *Journal of Research in Personality*, 34(2), 208–224.
- Riketta, M., & Sacramento, C. A. (2008). “They cooperate with us, so they are like me”: Perceived intergroup relationship moderates projection from self to outgroups. *Group Processes & Intergroup Relations*, 11(1), 115–131.
- Robbins, J. M., & Krueger, J. I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and Social Psychology Review*, 9(1), 32–47.
- Ross, L., Greene, D., & House, P. (1977). The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3), 279–301.
- Sadler, P., & Woody, E. (2003). Is who you are who you’re talking to? Interpersonal style and complementarity in mixed-sex interactions. *Journal of Personality and Social Psychology*, 84(1), 80–96.
- Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 1–65). New York: Academic Press.
- Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.-E., Demirutku, K., Dirilen-Gumus, O., & Konty, M. (2012). Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103(4), 663–688.
- Selfhout, M., Denissen, J. J. A., Branje, S., & Meeus, W. (2009). In the eye of the beholder: Perceived, actual, and peer-rated similarity in personality, communication, and friendship intensity during the acquaintanceship process. *Journal of Personality and Social Psychology*, 96(6), 1152–1165.
- Srivastava, S., Guglielmo, S., & Beer, J. S. (2010). Perceiving others’ personalities: Examining the dimensionality, assumed similarity to the self, and stability of perceiver effects. *Journal of Personality and Social Psychology*, 98(3), 520–534.
- Thielmann, I., Böhm, R., Ott, M., & Hilbig, B. E. (2021). Economic games: An introduction and guide for research. *Collabra: Psychology*, 7(1), 19004.
- Thielmann, I., Hilbig, B. E., & Zettler, I. (2020a). Seeing me, seeing you: Testing competing accounts of assumed similarity in personality judgments. *Journal of Personality and Social Psychology*, 118(1), 172–198.
- Thielmann, I., Rau, R., & Locke, K. D. (2020b). Trait-specificity versus global positivity: A critical test of alternative sources of assumed similarity in personality judgments. Manuscript submitted for publication.
- Thomas, G., Fletcher, G. J. O., & Lange, C. (1997). On-line empathic accuracy in marital interaction. *Journal of Personality and Social Psychology*, 72(4), 839–850.
- Toma, C., Yzerbyt, V., & Corneille, O. (2010). Anticipated cooperation vs. competition moderates interpersonal projection. *Journal of Experimental Social Psychology*, 46(2), 375–381.
- VandenBos, G. R. (2007). *APA dictionary of psychology*. Washington, DC: American Psychological Association.
- Watson, D., Hubbard, B., & Wiese, D. (2000). Self-other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of Personality and Social Psychology*, 78(3), 546–558.
- Weller, J., & Watson, D. (2009). Friend or foe? Differential use of the self-based heuristic as a function of relationship satisfaction. *Journal of Personality*, 77(3), 731–760.
- Wood, D., Harms, P., & Vazire, S. (2010). Perceiver effects as projective tests: What your perceptions of others say about you. *Journal of Personality and Social Psychology*, 99(1), 174–190.
- Zebrowitz, L. A., & Montepare, J. (2006). The ecological approach to person perception: Evolutionary roots and contemporary offshoots. In M. Schaller, J. A. Simpson, & D. T. Kenrick (Eds.), *Evolution and social psychology* (pp. 81–113). New York: Psychosocial Press.

**APPENDIX**

Honesty-Humility items for self-report (reverse-keyed items are marked by \*):

- I wouldn't use flattery to get a raise or promotion at work, even if I thought it would succeed.
- If I knew that I could never get caught, I would be willing to steal a million dollars. (\*)  
Having a lot of money is not especially important to me.
- I think that I am entitled to more respect than the average person is. (\*)
- If I want something from someone, I will laugh at that person's worst jokes. (\*)
- I would never accept a bribe, even if it were very large.
- I would get a lot of pleasure from owning expensive luxury goods. (\*)
- I want people to know that I am an important person of high status. (\*)
- I wouldn't pretend to like someone just to get that person to do favors for me.
- I'd be tempted to use counterfeit money, if I were sure I could get away with it. (\*)

Honesty-Humility items for observer report (reverse-keyed items are marked by \*):

- He/she wouldn't use flattery to get a raise or promotion at work, even if he/she thought it would succeed.
- If he/she knew that he/she could never get caught, he/she would be willing to steal a million dollars. (\*)
- Having a lot of money is not especially important to him/her.
- He/she thinks that he/she is entitled to more respect than the average person is. (\*)
- If he/she wants something from someone, he/she will laugh at that person's worst jokes. (\*)
- He/she would never accept a bribe, even if it were very large.
- He/she would get a lot of pleasure from owning expensive luxury goods. (\*)
- He/she wants people to know that he/she is an important person of high status. (\*)
- He/she wouldn't pretend to like someone just to get that person to do favors for him/her.
- He/she'd be tempted to use counterfeit money, if he/she were sure he/she could get away with it. (\*)

# 18 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 that 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 that 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 et al., 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 and the third edition of this book). More generally, overconfidence occurs if our confidence related to our judgments, inferences, or predictions is 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: types, 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 a 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):

$$B = \frac{1}{N} \sum_{i=1}^N (r_i - c_i)^2 \quad (18.1)$$

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 (for various decompositions of the Brier score, see Murphy, 1972, and for a discussion of other measures, see Keren, 1991, and Lichtenstein et al., 1982).

The measure that is most frequently used in the literature is *over/underconfidence*, which is simply defined as the difference between percentage correct (across all items) and mean confidence (across all items), specifically:

$$\text{Over / underconfidence} = \bar{r} - \bar{c} \quad (18.2)$$

### ***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% of the work on this project.” These assessments (and the hidden “I did 90% 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 11 and 21). 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 US 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 US 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 8). 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 21).

### **Classroom demonstrations**

It is easy to demonstrate overconfidence in the classroom. Text box 18.1 describes an experiment that uses the calibration paradigm. Text box 18.2 provides instructions and ten 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 1,332 intervals provided by 140 students over several years, an average of only 5.04 instead of 9 hits, as should be.

### Text box 18.1 Calibration in the classroom

#### Method

Calibration is obtained in each of two item sets that Gigerenzer et al. (1991) referred to as the *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 & Hertwig, 2006, and Hoffrage, 2011).

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 18.1 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).

#### 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

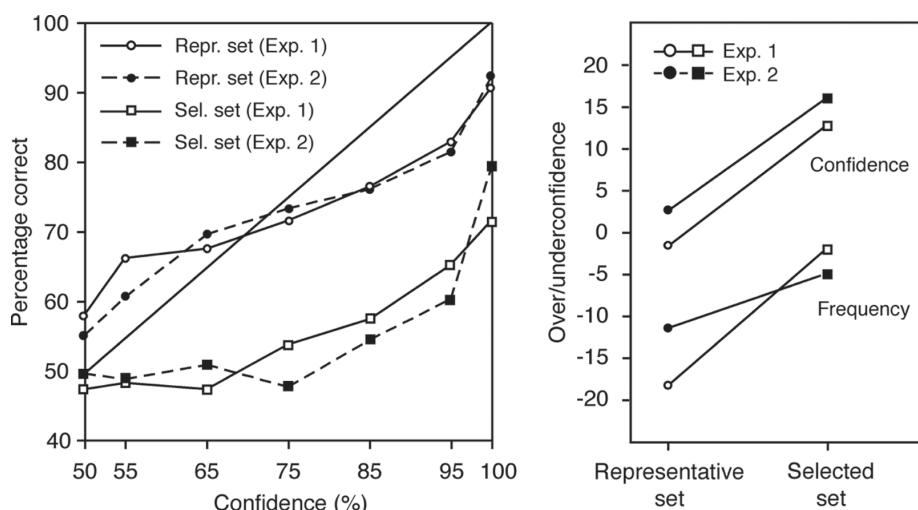


Figure 18.1 Left: Calibration curves for representative and selected item sets. Right: The graphs labeled *Confidence* depict mean confidence judgments minus mean percentage correct, and those labeled *Frequency* depict mean frequency estimates (which have been transformed to percentages to be comparable) minus percentage correct. Positive differences denote overconfidence, negative differences denote underconfidence. Data taken from Gigerenzer et al. (1991).

to capture the quite differentiated picture displayed in Figure 18.1. The theory of “probabilistic mental models” (Gigerenzer et al., 1991) provides such an approach and will be introduced below.

### Text box 18.2 Precision of numerical estimates in the classroom

The following instruction and ten-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<sup>2</sup>; 335 days.

## Typical findings

This section reviews some well-established results. 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 et al., 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 ( $\bar{r}$ ) and the over/underconfidence score ( $\bar{r} - \bar{c}$ ), 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 (Lichtenstein et al., 1982).

### ***The confidence-frequency effect***

In Figure 18.1 (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 et al. (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).

### ***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 et al.,

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 et al., 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 13), 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.

### ***Ecological models***

Independently, Gigerenzer et al. (1991) with their theory of probabilistic mental models (PMMs) and Juslin (1994) developed what were 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 et al. (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 18.1 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; see also Moore & Healy; 2008) 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” (Moore & Cain, 2007, 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  $p(\text{skilled} \mid \text{no accident in 3 past years}) > p(\text{unskilled} \mid \text{no accident in 3 past years})$ , 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, pp. 1592–3). 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.

### **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 27) 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 metacognitive 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 et al. (1999) put it: "In a world in which competence is hard to measure, confidence often wins the day" (p. 243). In this vein, Vitanova (2019) has shown, using a sample of 733 CEOs of US public companies, that power-induced overconfidence (i.e., overconfidence that is positively correlated with the amount of power held by a specific leader and that can hence be seen as endogenous) has a positive impact on overall firm performance, once this endogeneity is taken into account. Finally, Burks et al. (2013) have provided evidence suggesting that signaling can better account for overconfidence than rational beliefs or consumption.

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 it 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 find 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 et al. (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). In their analysis of firm data ranging from 1963 to 2001, Chuang and Lee (2006) provided evidence supporting the following four hypotheses related to overconfidence in financial markets: "Overconfident investors overreact to private information and underreact to public information" (p. 2492), "market gains (losses) make overconfident investors trade more (less) aggressively in subsequent periods" (p. 2494), "excessive trading of overconfident investors in securities contributes to observed excessive volatility" (p. 2499), and "overconfident investors underestimate risk and trade more in riskier securities" (p. 2500). For a detailed overview of studies that focus on overconfidence in finance and trading, see also 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. Following-up on Camerer and Lovallo's seminal study, Bernoster et al. (2018) found that overconfidence is indeed related to intended market entry, after all, only people with very high confidence in their ability to succeed decide to become entrepreneurs. However, these authors also report that overconfidence "does not place entrepreneurs in a particular position regarding entrepreneurial orientation" (p. 11), partly also because, for those who entered, competition and learning kick in (see also Chen et al., 2018). Based not only on theoretical but also on empirical grounds, Bernoster et al. urge us to treat overconfidence and optimism as separate constructs: Optimism drives individuals into entrepreneurship too, but in contrast to overconfidence, it is related to the market position entrepreneurs have, in particular with respect to proactiveness and risk taking.

Åstebro et al. (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). For a review of biases in entrepreneurship, with a focus on overconfidence but also with a discussion of other biases, see Zhang and Cueto (2017).

### ***Management and leadership***

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 et al. (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). Relatedly, in their analysis of 2,179 firms, 3,305 CEOs, and 13,120 firm years, Hribar and Yang (2016) found that overconfidence increases (1) the likelihood of issuing a forecast, (2) the amount of optimism in management forecasts, and (3) the precision of the forecast as indicated by narrower confidence intervals.

## 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 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 ruinous investments, lost court trials, 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 et al., 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).

## **Acknowledgment**

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.

## **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. *The Journal of Finance*, 55(2), 773–806.
- Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *The 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.
- Benoit, J. P., & Dubra, J. (2011). Apparent overconfidence. *Econometrica*, 79(5), 1591–1625.
- Bernoster, I., Rietveld, C. A., Thurik, A. R., & Torrès, O. (2018). Overconfidence, optimism and entrepreneurship. *Sustainability*, 10(7), 2233.
- 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.
- 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 signalling. *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 and Organization*, 61(3), 453–470.
- Chen, J. S., Croson, D. C., Elfenbein, D. W., & Posen, H. E. (2018). The impact of learning and overconfidence on entrepreneurial entry and exit. *Organization Science*, 29(6), 989–1009.
- Chuang, W. I., & Lee, B. S. (2006). An empirical evaluation of the overconfidence hypothesis. *Journal of Banking and Finance*, 30(9), 2489–2515.
- 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.
- Finkelstein, M. O., & Levin, B. (2001). *Statistics for lawyers*. New York: Springer Science & Business Media.
- 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. *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.
- 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 and New York: Cambridge University Press.
- Hribar, P., & Yang, H. (2016). CEO overconfidence and management forecasting. *Contemporary Accounting Research*, 33(1), 204–227.
- 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). Naïve 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: 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., & Winkler, R. L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *National Weather Digest*, 2, 2–9.
- Murphy, A. H. (1972). Scalar and vector partitions of the probability score (Part 1): Two-state situation. *Journal of Applied Meteorology*, 12, 595–600.
- 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.
- Scheinman, 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*. London: Allen Lane, Penguin Books.
- Vitanova, I. (2019). Nurturing overconfidence: The relationship between leader power, overconfidence and firm performance. *The Leadership Quarterly*, 101342.
- 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.
- Zhang, S. X., & Cueto, J. (2017). The study of bias in entrepreneurship. *Entrepreneurship Theory and Practice*, 41(3), 419–454.

# 19 Metacognitive illusions

Monika Undorf, Sofia Navarro-Báez, and Malte F. Zimdahl

People often reflect on their own learning, memory, and thinking. Imagine, for instance, students who are preparing for an exam. They engage in study activities such as reviewing lecture slides, rereading book chapters, writing summaries, or using flashcards. At the same time, they think about whether they understand the texts they are reading, consider how well they can remember definitions, concepts, and theories, and evaluate whether they have sufficiently learned the material to succeed in the exam. Thus, some of the students' cognitions are cognitive processes about cognitive processes, often termed *metacognition* (Flavell, 1971; Nelson & Narens, 1990).

Metacognition entails two components (Nelson & Narens, 1990). *Metacognitive monitoring* refers to thoughts, knowledge, and judgments about cognitive processes as well as to assessments of one's own cognitions. *Metacognitive control* refers to the use of this information for regulating cognition and behavior. In our example, the students engage in metacognitive monitoring when thinking about their understanding and learning, when reflecting on their learning progress, and when assessing their overall level of knowledge. They engage in metacognitive control when utilizing the output of their monitoring processes to self-regulate their learning. A student who feels that he does not progress well might try a different study strategy or take a break, whereas a student who thinks that she has sufficiently mastered the material might stop studying altogether.

Decades of research have demonstrated that accurate metacognition is critical for good performance on various cognitive tasks. For instance, a recent meta-analysis showed that accurate metacognition positively predicts academic performance in adults, adolescents, and children even when controlling for intelligence (Ohtani & Hisasaka, 2018). Thus, much can be gained from accurate metacognition. At the same time, there is a real danger that metacognitive illusions undermine cognitive performance.

Metacognitive illusions are defined as reliable and systematic dissociations between people's metacognitions and cognitions. Thus, unlike optical or cognitive illusions, illusory metacognitions do not deviate from some external "reality" (see Introduction) but from the cognitions they are supposed to assess. Apart from this difference, however, cognitive and metacognitive illusions share their defining features. In particular, illusory metacognitions occur involuntary, clash with people's conviction that they know their own minds, are difficult to avoid, and have attracted a great deal of interest from researchers and practitioners (e.g., Dunlosky & Metcalfe, 2009; Koriat, 2007; Undorf, 2020).

In this chapter, we present an overview of metacognitive illusions and describe how to explore metacognitive illusions in the classroom. We then discuss theoretical accounts for metacognitive illusions, point out real-life consequences of metacognitive illusions, and describe research on the mending of metacognitive illusions.

### Text box 19.1 Metacognitive illusions in real life

- 1) An employee who is faced with a logical problem works on it alone rather than together with her colleagues, because she believes that individual workers achieve at least as good reasoning performance as work teams. This metacognitive belief is held by people from various countries including managers and psychologists who study reasoning. However, it is in stark contrast to evidence that group discussions considerably improve reasoning performance (Mercier et al., 2015).
- 2) A responsible citizen aims to learn more about the positions of parliamentary candidates regarding various political topics. Because she considers herself well informed about the most relevant and most discussed issues, she focuses on topics that have not been much in the news lately. Research suggests that she falls prey to a metacognitive illusion: People overestimate their understanding of political topics they feel everyone should know about (Gaviria & Corredor, 2021).

## The assessment of metacognition

### *Metacognitive judgments*

Researchers assess metacognitive monitoring by asking people to judge their own cognitions. Numerous metacognitive judgments have been obtained (e.g., Dunlosky & Metcalfe, 2009). Among the most popular metacognitive judgments are the following: *Judgments of learning (JOLs)* are made during learning and predict one's chances of remembering recently studied information on a later test. *Feeling of knowing (FOK) judgments* predict the likelihood that currently unrecallable information is nevertheless available in memory and will be remembered in the future. *Metacomprehension judgments* are people's judgments about how well they have comprehended and learned text materials. People's *judgments of solvability* assess whether they can solve specific problems. *Confidence judgments* assess people's confidence in the correctness of their responses to cognitive tasks.

### *Accuracy of metacognitive judgments*

Metacognitive judgments are accurate if they correspond well with cognitive performance. Because high correspondence may take different forms, three aspects of metacognitive accuracy need to be distinguished. The first aspect is *calibration* (or *absolute accuracy*; e.g., Koriat, 2007). Metacognitive judgments are well calibrated if their level is similar to the level of actual performance. If a person's mean confidence judgment in her responses is 60% and she gives correct answers to six out of ten items, then absolute accuracy is excellent. Alternatively, metacognitive judgments can be *overconfident* or *underconfident*. Overconfident metacognitive judgments significantly exceed cognitive performance (see Chapter 18), whereas underconfident metacognitive judgments significantly underestimate cognitive performance. A widely used measure of calibration is the signed difference between a person's mean metacognitive judgments and the person's mean performance, referred to as *bias* in studies on metacognition (e.g., Dunlosky & Metcalfe, 2009) and as *over-/underconfidence* in Chapter 18. A positive bias reflects overconfidence,

a negative bias reflects underconfidence, and a bias near zero reflects good calibration. Calibration is relevant for metacognitive control processes at the task level. For instance, well-calibrated JOLs allow learners to devote adequate time and effort to a learning task as a whole. In contrast, overconfident learners prematurely terminate studying and receive disappointing test grades, whereas underconfident learners spend too much time and effort (Koriat, 2007).

The second aspect of metacognitive accuracy is *resolution* (or *relative accuracy*; e.g., Koriat, 2007). Resolution refers to the extent to which metacognitive judgments distinguish between correct and incorrect responses. If a person displays high confidence in all correct responses and low confidence in all incorrect responses, then relative accuracy is excellent. A widely used measure of resolution is the within-person Goodman-Kruskal gamma correlation between judgments and performance (Nelson, 1984; for criticism and alternatives, see, e.g., Bröder & Undorf, 2019). Gamma can range from  $-1$  to  $1$ , with higher values indicating higher resolution than lower values. Resolution is relevant to metacognitive control processes at the item level. High resolution of JOLs helps learners to selectively allocate study time and effort to material that is not yet well learned. In contrast, low resolution of JOLs may result in learners allocating study time and effort to material they have already mastered (Koriat, 2007). It is important to note that calibration and resolution are largely independent. Metacognitive judgments can be grossly overconfident but nevertheless exhibit high resolution. Also, metacognitive judgments can be well calibrated but still have poor resolution.

Finally, metacognitive judgments can be accurate in that they *track differences in actual performance* across conditions. This means that metacognitive judgments and actual performance show the same pattern across conditions. If a person is more confident in her responses on a music quiz than an astronomy quiz and indeed achieves better performance on the music quiz, then metacognitive judgments accurately track performance differences across quizzes. Notably, this does not necessarily imply that the person's confidence judgments are well-calibrated (e.g., confidence might exceed performance on both quizzes) or reveal high resolution (e.g., confidence might be similar for correct and incorrect responses in one or both quizzes). Accurate tracking of performance differences across conditions is relevant for metacognitive control processes at an intermediate level. Recognizing that a certain subject, topic, or type of problem is particularly difficult allows learners to allocate more study time and effort to this subject, topic, or problem type than to others.

Because the remainder of this chapter focuses on metacognitive illusions, it must be emphasized that metacognitive judgments are often quite accurate in terms of calibration and resolution and usually track performance differences across conditions (e.g., Koriat, 2007). Thus, as is the case with cognitive illusions (see Chapter 1), metacognitive illusions are the exception rather than the rule. Also, some metacognitive illusions were found in rather artificial experimental situations (see, e.g., Undorf, 2020). Finally, metacognitive illusions notwithstanding, metacognitive processes are of considerable adaptive and functional value (e.g., Benjamin, 2007).

## **Overview of metacognitive illusions**

This section presents an overview of illusions that afflict judgments about one's own learning, knowledge, and thinking (see Figure 19.1).

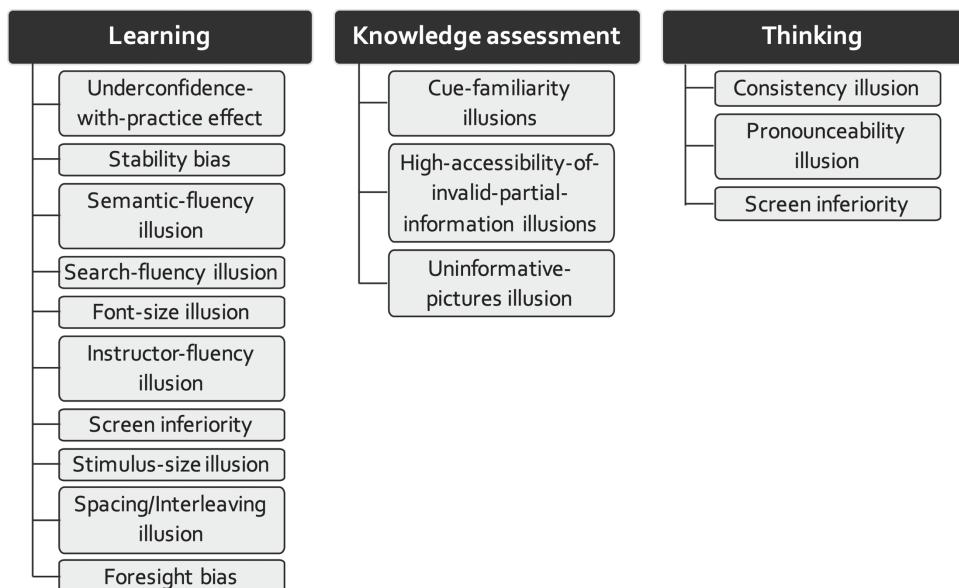


Figure 19.1 Overview of metacognitive illusions.

### **Metacognitive illusions of learning**

A well-replicated metacognitive illusion that impairs the calibration of predictions about learning is the *underconfidence-with-practice effect* (Koriat, 1997; see also Koriat et al., 2002). Koriat (1997) discovered that the calibration of JOLs deteriorates when people study and recall the same information across multiple study-test trials. Each study-test trial comprised a study phase in which people studied items such as word pairs (e.g., *glass – island*) and made JOLs, that is, predicted their chances of recalling each second word when seeing the first word at test. JOLs showed good calibration on the first study-test trial but shifted towards marked underconfidence from the second trial onwards. The reason for this was that people underestimated just how much memory performance benefits from study-test practice. Importantly, however, detrimental effects of study-test practice on JOL accuracy are limited to calibration. Repeated study-test trials typically improve JOL resolution, that is, increase the extent to which the JOLs differentiate between items that are remembered and items that are not remembered.

The *stability bias* (Koriat et al., 2004; Kornell & Bjork, 2009) is a metacognitive illusion in which people fail to anticipate future learning or future forgetting. Kornell and Bjork (2009) examined the impact of future study opportunities on JOLs. They asked participants to predict their chances of remembering recently studied items in tests that were announced to follow immediately upon the study phase or to take place after one, two, or three additional study-test trials. As expected, studying resulted in learning: Memory performance increased by 36% from Test 1 to Test 4. JOLs, in contrast, showed an unreliable increase of only 8%. The stability bias thus produced increasingly underconfident JOLs that failed to track performance improvements resulting from future study-test

trials. Koriat et al. (2004) showed that JOLs were insensitive to future forgetting. In their study, three groups of participants made JOLs for a memory test that was announced to take place immediately after studying, one day later, or one week later. Unsurprisingly, actual memory performance declined with longer retention intervals: Percentage correct dropped from 53% (immediate test) to 18% (test after one week). JOLs, however, were completely indifferent to the expected retention interval (immediate test: 52%; test after one week: 54%) and, consequently, severely overconfident with longer retention intervals. Here, the stability bias produced overconfident JOLs that failed to track performance impairments due to future forgetting.

A metacognitive illusion with dramatic effects on resolution is the *semantic-fluency illusion* (Benjamin et al., 1998). In Benjamin et al.'s (1998) experiment, participants answered general-knowledge questions and predicted the likelihood of recalling their answers on a free recall test in a second phase of the experiment. While actual free recall performance *increased* with the time it took people to answer the general-knowledge questions in Phase 1, JOLs *decreased* with response time in Phase 1. This dissociation resulted in negative gamma correlations between JOLs and free recall performance, indicating that people made lower JOLs for items they remembered than for items they did not remember. A related metacognitive illusion is the *search-fluency illusion* (Stone & Storm, 2021): People who used the internet to search for answers to general-knowledge questions made higher JOLs for the free recall of answers they found quickly rather than slowly, even though recall performance increased with search time.

The *font-size illusion* (Rhodes & Castel, 2008; for a review, see Luna et al., 2018) impairs the tracking of performance differences in JOLs. This illusion occurs when people make JOLs for study words that are printed in different font sizes (e.g., in a larger 48-point font and in a smaller 18-point font). People provide higher JOLs for large-font words than for small-font words, even though large-font words are remembered only slightly better than small-font words, if at all. Calibration and resolution typically remain unaffected by the font-size illusion.

Metacognitive illusions also afflict people's JOLs for complex study materials. An example is the *instructor-fluency illusion* (Carpenter et al., 2013). In Carpenter et al.'s (2013) study, students viewed one of two videotaped lectures. In the fluent video, the lecturer spoke freely and fluently and maintained eye contact with the camera. In the disfluent video, the same lecturer read the same content haltingly from notes, repeatedly flipped through her notes, and switched her gaze between camera and notes. Participants who viewed the fluent video predicted that they would remember a greater amount of information than those who viewed the disfluent video (48% v. 25%), even though actual memory performance was similar across the two groups (26% v. 22%). Viewing the fluent video thus led to severely overconfident JOLs.

The calibration of metacomprehension judgments predicting people's memory for texts is afflicted by *screen inferiority* (Ackerman & Goldsmith, 2011). Studying expository texts on screen rather than on paper results in overconfident predictions of one's performance on later tests. As a consequence, people do not devote enough study time to texts that are presented on screen and achieve low test performance.

### ***Metacognitive illusions of knowledge assessment***

Researchers have extensively examined the monitoring of whether information is available in memory – the feeling of knowing – with FOK judgments. FOK judgments are

subject to *cue-familiarity illusions*. In a study by Schwartz and Metcalfe (1992), participants studied word pairs for a cued recall test. For each target that participants could not recall at test, they indicated their confidence in identifying it in a later recognition memory test. Results showed higher FOK judgments when the pair's cue word had been presented in an initial priming phase, even though priming did not affect actual memory. Reder and Ritter (1992) presented participants with arithmetic problems (e.g.,  $18 \times 27$ ) and asked them to quickly judge whether they could retrieve the answer from memory or had to compute it. People's FOK judgments were the higher the more often they had seen parts of the problems in a previous training phase. This was even true when people did not have the answer in memory because the problems were novel (e.g., repeated exposure to  $18 + 27$  increased FOK judgments for  $18 \times 27$ ). These findings demonstrate that high familiarity of the information used to prompt FOK judgments can produce unduly high feelings of knowing that also fail to track differences in memory performance across conditions (or the lack thereof).

Another prominent illusion that afflicts FOK judgments involves *high accessibility of invalid partial information*. In a study by Koriat (1995), people answered general-knowledge questions and predicted their chances to recognize the correct answer in multiple-choice versions of the questions. Results showed that FOK judgments were based on the amount, intensity, and vividness of partial information that came to mind when searching for the answer in memory (e.g., someone expects that he knows the capital of Sweden because he can recall the capital of Denmark, watched a documentary about Norway, and feels knowledgeable about European geography). Reliance on accessibility of partial information produces illusory FOK judgments for deceptive items that bring to mind mostly incorrect information. For instance, the partial information people access when asked *To what country does Corsica belong?* typically points towards the wrong answer *Italy* rather than the correct answer *France*, resulting in too high FOK judgments. Thus, highly accessible invalid partial information produces unduly high feelings of knowing.

The *uninformative-pictures illusion* (Cardwell et al., 2017) demonstrates that the presence of irrelevant information can inflate people's assessments of their knowledge. This illusion was found when people judged their knowledge of natural and mechanical processes. Cardwell et al. (2017) reported that judgments of understanding were higher when descriptive phrases such as *how rainbows form* or *how radios work* were preceded by uninformative photos (e.g., a picture of a rainbow in the sky). Presumably, the photos made it easier to generate relevant thoughts and images, which people erroneously interpreted as indicative of profound knowledge and understanding of the to-be-judged processes.

### ***Metacognitive illusions of thinking***

Metacognitive illusions also occur in thinking tasks. An example is the *consistency illusion* (Williams et al., 2020). Williams et al. (2020) presented six letters one of which was highlighted and instructed participants to find all five-letter words that began with the highlighted letter and could be formed by rearranging the presented letters. People's retrospective confidence judgments about whether they had found all available words were higher when the highlighted letter was always in the same position (consistent) rather than when its position varied across trials (inconsistent). In contrast, position

consistency did not affect the number of words participants found. This finding shows that superficial procedural consistency can inflate confidence in one's performance in thinking tasks.

Topolinski et al. (2016) reported the *pronounceability illusion* in quick judgments of solvability for anagrams (e.g., *edisepo*, solution: *episode*; *ekisepto*, no solution). People rated easy-to-pronounce anagrams (e.g., *giloc*, solution: *logic*) more often as solvable than hard-to-pronounce anagrams (e.g., *ogil*), even though easy-to-pronounce anagrams are harder to solve than hard-to-pronounce anagrams.

### **Text box 19.2 Creating a font-size illusion in class**

This is a simplified and abbreviated version of Undorf and Zimdahl's (2019) Experiment 2. Preprogrammed versions of this classroom demonstration (in English and German) including stimuli, instructions, and worksheets, as well as a template for analysis are available at <https://osf.io/p5gau>.

#### **Method**

##### ***Participants***

The font-size illusion is quite large and should be found with 20 to 30 participants.

##### ***Materials***

The experiment requires four font sizes. Select the smallest font size so that one can just read the words and select the largest font size so that the words fill a large part of the display. Then, select the two intermediate font sizes so that differences between all successive font sizes are equal perceptually. Undorf and Zimdahl (2019) used 9-point, 29-point, 93-point, and 294-point Arial fonts. The experiment requires 40 nouns of the same length (e.g., five letters). Assign ten randomly selected words to each font size.

##### ***Procedure***

Ideally, all participants should view the stimuli from a similar distance. The experiment consists of three phases: a study phase, a filler task, and a test phase. Prior to the study phase, inform participants that they will study 40 words for a later test, in which they will be asked to write down as many study words as they can remember. Also inform participants that, during the study phase, they are required to estimate the chance of remembering each item at test on a percentage scale by writing down any number between 0 and 100. In the study phase, present each word for 2 seconds via projector. Immediately afterwards, present the JOL prompt “The chance to recall (0%–100%)?” Participants can write down their JOLs on a sheet of paper that lists word numbers in a column. After the study phase, participants work on an unrelated filler task for 90 seconds (e.g., solving arithmetic problems). In the test phase, participants are given 4 minutes to write down as many studied words as they can remember on a sheet of paper.

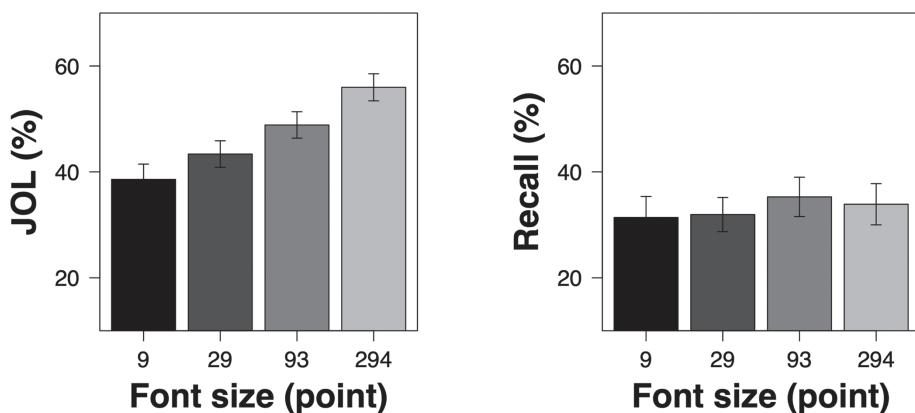


Figure 19.2 Expected results from the font size demonstration.

Note: Data taken from Undorf and Zimdahl (2019, Experiment 2, no-chin rest condition).

## Analysis

A font size illusion is present when JOLs increase with increasing font size, whereas recall performance is similar across all font sizes (see Figure 19.2 for typical results). A one-way ANOVA on JOLs using font size as a repeated-measures factor should reveal a significant effect. In contrast, a similar ANOVA on recall performance should be insignificant.

## Theoretical accounts

Metacognitive illusions provide important insights into the basis of metacognition. First and foremost, systematic dissociations between metacognitive judgments and actual performance favor inferential accounts over direct-access accounts of metacognition (Dunlosky & Metcalfe, 2009; Koriat, 2007).

### *Direct-access versus inferential accounts of metacognition*

*Direct-access accounts* of metacognition propose that metacognitive monitoring involves direct access to one's own ongoing cognitive processes. According to direct-access accounts, JOLs rely on people's observations of the memory traces that are formed during learning and of how trace strength increases as learning proceeds (Koriat, 1997). Similarly, FOK judgments are presumed to arise from searching one's memory for traces of the currently unrecallable information and reading out whether a relevant trace exists and how strong it is (Hart, 1965). Because direct-access accounts assume that metacognitive judgments rely on the to-be-judged cognitions themselves, these accounts predict that metacognitive judgments are highly accurate throughout. Consequently, direct-access accounts of metacognition are incompatible with the metacognitive illusions we presented in the previous section and, more generally, with any systematic dissociation between metacognitive judgments and actual performance (e.g., Koriat, 2007).

Metacognitive illusions do, however, provide strong evidence for *inferential accounts* of metacognition. Inferential accounts assume that people infer the state of their cognitive system from cues and heuristics. JOLs are assumed to rely on cues that are available at study and pertain to the stimuli or learning conditions, such as semantic associations between the two words of a pair (e.g., no association: *light – box*; strong association: *zebra – stripe*; Koriat, 1997), the font size of study words (Rhodes & Castel, 2008), the length and complexity of sentences from to-be-studied texts (Maki et al., 2005), or whether texts are presented on screen or on paper (Ackerman & Goldsmith, 2011). FOK judgments are assumed to rely on the familiarity of the cues used to probe memory (Reder & Ritter, 1992; Schwartz & Metcalfe, 1992) and on the accessibility of partial information about the elusive information (Koriat, 1993, 1995).

According to inferential accounts of metacognition, metacognitive judgments are accurate when based on cues that are predictive of actual performance. In contrast, metacognitive illusions emerge when people base metacognitive judgments on invalid cues or fail to take cues into account that are predictive of actual performance. Examples of valid cues include “semantic association” or “text complexity” (Koriat, 1997; Maki et al., 2005). Basing one’s JOLs on the invalid cue “font size” produces the font-size illusion (see above and Text box 19.2), and people’s failure to incorporate the valid cue “future study opportunities” in their JOLs underlies the stability bias (see above).

### **Theory- and experience-based processes**

Inferential accounts of metacognition distinguish between two types of processes underlying metacognitive judgments: theory-based and experience-based processes (e.g., Dunlosky & Tauber, 2014; Koriat & Levy-Sadot, 1999). *Theory-based processes* comprise the deliberate, analytic use of explicit beliefs and knowledge about cognition in general and about one’s own cognitive processes in particular. Prominent beliefs about memory include the beliefs that long-term memory has a limited capacity (held by 69% of a representative sample of 1,000 adult Norwegians; Magnussen et al., 2006), that dramatic events are remembered better than everyday events (70%; Magnussen et al., 2006), and that memory starts to decline by one’s 70th birthday at the latest (85%; Magnussen et al., 2006). When basing metacognitive judgments on theory-based processes, people are aware of the basis of their judgments (e.g., *I expect to remember this situation, because it frightens me*).

*Experience-based processes* are by-products of cognitive processes that influence people’s metacognitive judgments through non-analytic inferential processes operating below full awareness. Experience-based processes have the phenomenal quality of direct and unexplained intuitions (e.g., Koriat & Levy-Sadot, 1999). They often involve feelings of fluency, that is, the ease of processing information during reading, learning, retrieval, or problem solving. When basing metacognitive judgments on experience-based processes, people are unaware of the basis of their judgments (e.g., *I will certainly remember this information, because I have a strong feeling that I will*).

Theory-based and experience-based processes both contribute to accurate metacognitive judgments and metacognitive illusions. Correct explicit knowledge and beliefs may foster accurate metacognitive judgments. For instance, the belief that dramatic events are remembered better than everyday events may help learners to accurately predict that they will recall emotional stimuli better than neutral stimuli (Undorf et al., 2018; Zimmerman & Kelley, 2010). At the same time, basing metacognitive judgments on incomplete or faulty explicit beliefs about cognition harms metacognitive accuracy. For instance, people’s

knowledge about how emotion impacts memory is often incomplete insofar as they are unaware of the fact that memorial benefits of emotional stimuli do not extend to the cued recall of word pairs. Thus, basing JOLs on one's belief about memory for emotional information may contribute to unduly high JOLs for emotional word pairs (Undorf & Bröder, 2020; Zimmerman & Kelley, 2010). Also, the incorrect belief that the capacity of long-term memory is limited may contribute to underestimating the exceptionally good memory performance for pictures and, consequently, underconfident JOLs (e.g., Undorf & Bröder, 2021).

It is important to note that holding metacognitive beliefs does not guarantee that these affect metacognitive judgments. For instance, the widely held belief that memory declines with old age did not affect the JOLs students made in seven experiments reported by Tauber et al. (2019). Often, metacognitive knowledge and beliefs must be activated to be incorporated into metacognitive judgments (Undorf & Erdfelder, 2015).

Basing metacognitive judgments on experience-based cues and heuristics that are predictive of actual performance results in accurate metacognitive judgments, whereas reliance on invalid experience-based cues and heuristics produces metacognitive illusions. Because the validity of non-analytic cues can vary between situations, reliance on one and the same cue may produce accurate metacognitive judgments under some conditions and metacognitive illusions under other conditions. One example is retrieval fluency. Koriat and Ma'ayan (2005) found that basing JOLs for word pairs on the ease of retrieving information from memory contributed to JOL accuracy. In their study, JOLs were elicited with a delay after studying. Prior to making each JOL, participants saw the first word of the respective pair and were asked to recall the second word. The quicker participants retrieved the second word prior to making their JOLs, the higher were their JOLs and their performance on a final test. Thus, basing JOLs on retrieval fluency contributed to accurate JOLs. At the same time, however, retrieval fluency underlies the semantic-fluency illusion, which is due to basing JOLs for the free recall of words on the fluency of retrieving these words in a completely unrelated task (see above).

In recent years, many studies addressed the relative contributions of experience-based and theory-based processes to metacognitive illusions (for a review, see Undorf, 2020). This literature suggests, for instance, that the font-size illusion is mainly due to explicit beliefs about memory (but see Yang et al., 2018). People were found to believe that words printed in larger fonts are easier to remember than words printed in smaller fonts (e.g., Mueller et al., 2014; Undorf & ZimdaHL, 2019). Also, independent measures of fluency such as self-paced study time or lexical decision time were similar for small-font and large-font words (e.g., Mueller et al., 2014). Most importantly, using a wide range of font sizes (6 point to 500 point) dissociated perceptual fluency and JOLs. While JOLs increased with font size across the whole range of font size, perceptual fluency was lower for very small and very large font sizes than for intermediate font sizes (e.g., Undorf & ZimdaHL, 2019).

In contrast, other illusions were found to be due to experience-based processes. One example is the *stimulus-size illusion* (Undorf et al., 2017). This illusion afflicts JOLs for visual stimuli that gradually increase in size: Stimuli that increased quickly rather than slowly received higher JOLs, even though clarification speed did not affect memory performance. The finding that quick clarification did not increase JOL directly, but only indirectly through higher fluency (independently measured as the time it took people to identify the stimuli) indicates that the stimulus-size illusion is due to experience-based

processes. Moreover, participants were unaware of differences in clarification speed and did not believe quick clarification to help memory (Undorf et al., 2017).

In summary, the very existence of metacognitive illusions demonstrates that metacognitive judgments are inferences from cues people regard as informative about cognition in general and their own cognition in particular. Metacognitive illusions emerge when people base their metacognitive judgments on invalid cues or fail to rely on valid cues. The cues that produce metacognitive illusions can be broadly categorized into (1) faulty beliefs about cognition that affect metacognitive judgments through theory-based processes and (2) invalid non-analytic cues and heuristics that affect metacognitive judgments through experience-based processes.

## Applied perspectives

Metacognitive monitoring plays an integral role for effective self-regulation of cognition and behavior (e.g., Bjork et al., 2013; Metcalfe, 2009). Illusions and biases in metacognitive judgments can therefore lead to ineffective self-regulation and poor performance in cognitive tasks.

### *Real-life consequences of metacognitive illusions*

There is ample evidence for detrimental effects of metacognitive illusions and biases on cognitive performance (cf. Text box 19.1 above). Many studies addressed the negative consequences of overconfidence (see Chapter 18). For instance, Dunlosky and Rawson (2012) provided clear evidence that overconfidence can produce underachievement. In their experiments, students learned key-term definitions across alternating study and test phases (e.g., *The just-world hypothesis is the strong desire or need people have to believe that the world is an orderly, predictable, and just place, where people get what they deserve*). In each test phase, students judged the accuracy of their recalls. Individual differences in overconfidence predicted performance on the final test: The less overconfident people were when judging the accuracy of their recalls, the better was their final test performance. Also, experimentally reducing overconfidence in the students' judgments improved their final test performance.

Other research focused on adverse effects of specific metacognitive illusions. An educationally relevant illusion is the *spacing* (or *interleaving*) *illusion*, that is, people's failure to appreciate that long-term retention and inductive reasoning benefit from the spacing, rather than massing, of practice trials (e.g., Kornell & Bjork, 2008a; Logan et al., 2012). This illusion contributes to learners' robust preference for blocking repeated study trials (Tauber et al., 2013) and for restudying high priority information sooner rather than later (Cohen et al., 2013), even though restudying after (longer) intervals is more effective.

The surprising finding that learners who have the option to self-regulate their learning by dropping flashcards perform less well than learners who may not drop flashcards and, consequently, spend similar study time on all information, has been connected to the stability bias mentioned above (Kornell & Bjork, 2008b). Presumably, learners who underestimate the learning that might result from additional studying and disregard future forgetting drop flashcards too early and, consequently, perform poorly.

Although often studied from an educational perspective, negative effects of metacognitive illusions on cognitive performance are by no means restricted to educational settings. For instance, Hargis and Castel (2018) discuss adverse consequences of metacognitive

illusions on health. The authors focus on how metacognitive illusions may contribute to the difficulty of taking medications as prescribed, which is experienced by many people and older adults in particular. Overconfidence in one's own learning and memory of which medication to take in what dosage or at what time of day may prevent patients from taking notes in the doctor's office or from recognizing that they have forgotten important medication-related information. The stability bias may contribute to patients overestimating how far into the future they will remember medication-related information provided by the doctor. This example demonstrates that metacognitive illusions play a role in the healthcare sector, with potentially dramatic implications for patients and considerable costs for society (e.g., falls, heart failure, and preventable hospitalizations).

### ***Mending metacognitive illusions***

In view of the adverse consequences of inaccurate metacognitive judgments, an obvious question is whether metacognitive illusions can be mended. Research suggests that this is indeed possible, even though it is not necessarily easy. Sidi et al. (2017) demonstrated that alleviating screen inferiority is relatively easy: Emphasizing that tasks were of high importance eliminated people's overconfidence in their solutions to challenging problems when the problems were presented on the screen rather than on paper. Similarly, Koriat et al. (2004) found that JOLs were sensitive to future forgetting when the announced retention interval was manipulated within participants rather than between participants: JOLs declined monotonically with retention interval when participants knew during study that some items would be tested after 10 minutes, whereas other items would be tested after one day or one week.

In other cases, however, metacognitive illusions are difficult to overcome. One example is the interleaving illusion. In a study by Yan et al. (2016), people learned the styles of artists from examples of the artists' paintings. During study, paintings were presented blocked by artist for some artists and interleaved for other artists. On the test, participants assigned new paintings to the studied artists. Almost all participants achieved better classification performance for artists whose paintings were studied interleaved rather than blocked. After the test, however, they incorrectly judged that blocking was more effective than interleaving. Even when receiving personalized feedback about their test performance and detailed explanations why interleaving is superior to blocking, at least 50% of participants continued to show the interleaving illusion. After conducting six experiments with a total of more than 700 participants, Yan et al. (2016) concluded that "mending the metacognitive illusion that blocking is more effective for learning proved a daunting task" (p. 931). Why is the interleaving illusion so difficult to overcome? Yan et al. (2016) identified three reasons. First, learners experience blocked practice as easier and more fluent than interleaved practice. Second, most people have a strong belief that blocking is more effective than interleaving. Finally, learners often assume that general principles of learning may not apply to themselves. When informed that interleaving leads to better learning in 90% of learners, between 48% and 69% of the learners placed themselves in the remaining 10% of learners.

Mending metacognitive illusion is further complicated by the fact that successful debiasing may not generalize beyond the context in which it originally took place (e.g., Koriat & Bjork, 2006). Koriat and Bjork (2006) examined procedures for alleviating the *foresight bias*, that is, inflated JOLs for to-be-remembered information that appears obvious at study but is relatively difficult to remember at test (e.g., *water – well* appears semantically

related and highly memorable at study, but when *water* is presented alone at test, many other likely responses such as *drink*, *cool*, and *swim* come to mind and reduce memory performance). Two different debiasing procedures effectively alleviated the foresight bias. First, studying the same items across multiple study-test cycles considerably reduced the foresight bias beginning with the second study-test cycle. Second, educating participants about the foresight bias after the first study-test cycle and instructing them to avoid this bias resulted in a similar amount of debiasing from the second study-test cycle on. Importantly, however, only the debiasing achieved by explicit instruction yielded transfer to a new set of items (Koriat & Bjork, 2006).

Despite these difficulties in mending specific metacognitive illusions, it is certainly possible to create conditions that promote high monitoring accuracy. For instance, metacognitive monitoring accuracy in the classroom can be improved by using the *wait-generate-validate strategy* (Hausman et al., 2021). Wait refers to the recommendation that one should judge one's understanding not immediately after learning but should wait for a couple of hours or one day. After this delay, one should try to actively generate the material from memory without the help of study materials. Finally, one needs to validate the accuracy and completeness of the information one has generated. When following these steps, it is relatively easy to accurately assess one's learning and understanding. The resulting improvement in monitoring can significantly increase learning.

In summary, debiasing and creating favorable conditions for accurate metacognitive monitoring are promising approaches to ameliorating the negative effects of metacognitive illusions on cognitive performance.

## Conclusions

Metacognitive illusions have attracted a great deal of interest from researchers and, because of their detrimental effects on cognitive performance, from practitioners. There is no doubt that the study of metacognitive illusions has substantially advanced our understanding of metacognition. However, it is important to keep in mind that metacognitive illusions are by-products of functional metacognitive processes and that metacognition is quite accurate by and large.

## Summary

- Metacognition refers to people's assessments of their cognition (metacognitive monitoring) and to the use of this information for self-regulation (metacognitive control).
- Metacognitive illusions are systematic dissociations between cognition and metacognition that have been found in domains such as learning, knowledge assessment, and thinking.
- Metacognitive illusions undermine self-regulation and performance in cognitive tasks.
- The existence of metacognitive illusions indicates that people cannot directly access their cognition, but infer the state of their cognitive system from cues and heuristics.
- Faulty explicit beliefs about cognition as well as invalid non-analytic cues and heuristics contribute to metacognitive illusions.
- The negative effects of metacognitive illusions on cognitive performance can be ameliorated by debiasing techniques and by creating favorable conditions for assessing one's cognitive processes.

## Further reading

Dunlosky and Metcalfe's (2009) book about metacognition provides an overview of the accuracies and inaccuracies of metacognition. Contributions of fluency to metacognitive illusions are reviewed in Undorf (2020). Bjork et al. (2013) explore how metacognitive illusions can impair self-regulated learning. Evidence-based recommendations for improving metacognition are provided by Hausman et al. (2021).

## Acknowledgment

Manuscript preparation was supported by grants UN 345/1-3 and UN 345/2-1 from the Deutsche Forschungsgemeinschaft and by a Margarete von Wrangell fellowship from the state of Baden-Württemberg to Monika Undorf.

## References

- Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied*, 17(1), 18–32.
- Benjamin, A. S. (2007). Memory is more than just remembering: Strategic control of encoding, accessing memory, and making decisions. *Psychology of Learning and Motivation*, 48(7), 175–223.
- 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(1), 55–68.
- Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, 64, 417–444.
- Bröder, A., & Undorf, M. (2019). Metamemory viewed through the judgment lens. *Acta Psychologica*, 197, 153–165.
- Cardwell, B. A., Lindsay, D. S., Förster, K., & Garry, M. (2017). Uninformative photos can increase people's perceived knowledge of complicated processes. *Journal of Applied Research in Memory and Cognition*, 6(3), 244–252.
- Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: Instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review*, 20(6), 1350–1356.
- Cohen, M. S., Yan, V. X., Halamish, V., & Bjork, R. A. (2013). Do students think that difficult or valuable materials should be restudied sooner rather than later? *Journal of Experimental Psychology: Learning, Memory and Cognition*, 39(6), 1682–1696.
- Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Los Angeles, CA: Sage.
- Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, 22(4), 271–280.
- Dunlosky, J., & Tauber, S. K. (2014). Understanding people's metacognitive judgments: An isomechanism framework and its implications for applied and theoretical research. In T. J. Perfect & D. S. Lindsay (Eds.), *The Sage handbook of applied memory* (pp. 444–464). Los Angeles, CA: Sage.
- Flavell, J. H. (1971). First discussant's comments: What is memory development the development of? *Human Development*, 14(4), 272–278.
- Gaviria, C., & Corredor, J. (2021). Illusion of explanatory depth and social desirability of historical knowledge. *Metacognition and Learning*. <https://doi.org/10.1007/s11409-021-09267-7>
- Hargis, M. B., & Castel, A. D. (2018). Improving medication understanding and adherence using principles of memory and metacognition. *Policy Insights from the Behavioral and Brain Sciences*, 5(2), 147–154. <https://doi.org/10.1177/2372732218781643>

- Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, 56(4), 208–216.
- Hausman, H., Myers, S. J., & Rhodes, M. G. (2021). Improving metacognition in the classroom. *Zeitschrift für Psychologie*, 229(2), 89–103.
- Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100(4), 609–639.
- Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General*, 124(3), 311–333.
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349–370.
- Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *Cambridge handbook of consciousness* (pp. 289–325). New York: Cambridge University Press.
- Koriat, A., & Bjork, R. A. (2006). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(5), 1133–1345.
- Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, 133(4), 643–656.
- Koriat, A., & Levy-Sadot, R. (1999). Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 483–502). New York: Guilford Press.
- Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, 52(4), 478–492.
- Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, 131(2), 147–162.
- Kornell, N., & Bjork, R. A. (2008a). Learning concepts and categories: Is spacing the “enemy of induction”? *Psychological Science*, 19(6), 585–592.
- Kornell, N., & Bjork, R. A. (2008b). Optimising self-regulated study: The benefits – and costs – of dropping flashcards. *Memory*, 16(2), 125–136.
- Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, 138(4), 449–468.
- Logan, J. M., Castel, A. D., Haber, S., & Viehman, E. J. (2012). Metacognition and the spacing effect: The role of repetition, feedback, and instruction on judgments of learning for massed and spaced rehearsal. *Metacognition and Learning*, 7(3), 175–195.
- Luna, K., Nogueira, M., & Albuquerque, P. B. (2018). Words in larger font are perceived as more important: Explaining the belief that font size affects memory. *Memory*, 27(4), 555–560.
- Magnussen, S., Andersson, J., Cornoldi, C., De Beni, R., Endestad, T., Goodman, G. S., ... & Zimmer, H. D. (2006). What people believe about memory. *Memory*, 14(5), 595–613.
- Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology*, 97(4), 723–731.
- Mercier, H., Trouche, E., Yama, H., Heintz, C., & Girotto, V. (2015). Experts and laymen grossly underestimate the benefits of argumentation for reasoning. *Thinking & Reasoning*, 21(3), 341–355.
- Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, 18(3), 159–163.
- Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*, 70, 1–12.

- Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*, 20(2), 378–384.
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95(1), 109–133.
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). New York: Academic Press.
- Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning*, 13(2), 179–212.
- Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(3), 435–451.
- Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, 137(4), 615–625.
- Schwartz, B. L., & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(5), 1074–1083.
- Sidi, Y., Shpigelman, M., Zalmanov, H., & Ackerman, R. (2017). Understanding metacognitive inferiority on screen by exposing cues for depth of processing. *Learning and Instruction*, 51, 61–73.
- Stone, S. M., & Storm, B. C. (2021). Search fluency as a misleading measure of memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(1), 53–64.
- Tauber, S. K., Dunlosky, J., Rawson, K. A., Wahlheim, C. N., & Jacoby, L. L. (2013). Self-regulated learning of a natural category: Do people interleave or block exemplars during study? *Psychonomic Bulletin & Review*, 20(2), 356–363.
- Tauber, S. K., Witherby, A. E., & Dunlosky, J. (2019). Beliefs about memory decline in aging do not impact judgments of learning (JOLs): A challenge for belief-based explanations of JOLs. *Memory & Cognition*, 47(6), 1102–1119.
- Topolinski, S., Bakhtiari, G., & Erle, T. M. (2016). Can I cut the Gordian knot? The impact of pronounceability, actual solvability, and length on intuitive problem assessments of anagrams. *Cognition*, 146, 439–452.
- Undorf, M. (2020). Fluency illusions in metamemory. In A. M. Cleary & B. L. Schwartz (Eds.), *Memory quirks: The study of odd phenomena in memory* (pp. 150–174). New York: Routledge.
- Undorf, M., & Bröder, A. (2020). Cue integration in metamemory judgements is strategic. *Quarterly Journal of Experimental Psychology*, 73(4), 629–642.
- Undorf, M., & Bröder, A. (2021). Metamemory for pictures of naturalistic scenes: Assessment of accuracy and cue utilization. *Memory & Cognition*, 49(7), 1405–1422.
- Undorf, M., & Erdfelder, E. (2015). The relatedness effect on judgments of learning: A closer look at the contribution of processing fluency. *Memory & Cognition*, 43(4), 647–658.
- Undorf, M., Söllner, A., & Bröder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & Cognition*, 46(4), 507–519.
- Undorf, M., & Zimdahl, M. F. (2019). Metamemory and memory for a wide range of font sizes: What is the contribution of perceptual fluency? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(1), 97–109.
- Undorf, M., Zimdahl, M. F., & Bernstein, D. M. (2017). Perceptual fluency contributes to effects of stimulus size on judgments of learning. *Journal of Memory and Language*, 92, 293–304.
- Williams, E. F., Duke, K. E., & Dunning, D. (2020). Consistency just feels right: Procedural fluency increases confidence in performance. *Journal of Experimental Psychology: General*, 149(12), 2395–2405.

- Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, 145(7), 918–933.
- Yang, C., Huang, T. S.-T., & Shanks, D. R. (2018). Perceptual fluency affects judgments of learning: The font size effect. *Journal of Memory and Language*, 99, 99–110.
- Zimmerman, C. A., & Kelley, C. M. (2010). “I’ll remember this!” Effects of emotionality on memory predictions versus memory performance. *Journal of Memory and Language*, 62(3), 240–253.

## 20 Fake news and participatory propaganda

*Stephan Lewandowsky*

Everybody worries about “fake news”. In every member state of the European Union, at least half of respondents in a recent large survey ( $N \approx 27,000$ ) say they come across fake news once a week or more (Directorate-General for Communication, 2018). Similarly, in the US, 89% of adults indicated that they come across made-up news intended to mislead the public at least sometimes (Mitchell et al., 2019). There is abundant evidence that misinformation can have adverse consequences for individuals and societies as a whole (Lewandowsky et al., 2017, 2012). Public awareness of those adverse consequences is high, with more than eight out of ten EU citizens identifying fake news as a problem for democracy in general (Directorate-General for Communication, 2018). Accordingly, concern about fake news has become a focal point of policy discussions, with the European Union about to announce various legislative initiatives such as the Digital Services Act that seek to provide greater accountability of social media platforms, which have been identified as a major vector of misinformation (Kozyreva et al., 2020). Public opinion seems supportive of legislative measures, with eight out of ten Americans believing that steps should be taken to restrict fake news (Mitchell et al., 2019) and with four out of ten EU citizens assigning responsibility to national authorities (Directorate-General for Communication, 2018).

The widespread public concern about misinformation gives rise to a conundrum because misinformation does not multiply and spread on its own – it is spread by people who *choose* to spread it. Without active support from members of the public, much misinformation would likely remain a fringe phenomenon. An illustrative example of the path from fringe to mainstream involves the “pizzagate” conspiracy theory of 2016. This theory originated with a tweet that linked presidential candidate Hillary Clinton to a pedophilia ring operating out of the basement of a pizza parlor in Washington, DC. The first occurrence of the #pizzagate hashtag has been traced to a trolling account that often tweeted pro-Nazi content and was ultimately suspended by Twitter (Metaxas & Finn, 2019). The hashtag was then actively retweeted by accounts based mainly in the Czech Republic, Cyprus, and Vietnam (Fisher et al., 2016). The theory was actively promoted by right-wing influencers, including the son of President Trump’s former National Security Adviser. The conspiracy theory quickly migrated from Twitter onto Reddit and far-right sites such as 4Chan, where it became linked with a specific pizza parlor. Eventually, discussion on a dedicated subreddit (*r/pizzagate*) began to reveal private information about employees at the pizza parlor and stores nearby. Approximately a month after the theory first appeared, an armed individual entered the pizza parlor and fired shots inside in search of a (non-existent) basement (Fisher et al., 2016). By late December 2016, some six weeks

after the first tweet, 46% of Trump voters considered the pizzagate conspiracy theory to be true (Frankovic, 2016).

Pizzagate is a particularly prominent example of how conspiracy theories and other misinformation can move from the fringe into mainstream public discourse (boyd, 2017). Several important issues arise from pizzagate that deserve to be explored. The first issue relates to the architecture of social media and how it facilitates the spread of misinformation and conspiracy theories. These consequences of the online ecosystem have been examined in detail elsewhere (e.g., Kozyreva et al., 2020; Lorenz-Spreen et al., 2020) and, without wishing to downplay their importance, I set them aside for the remaining discussion.

The second issue, which I focus on here, relates to the psychological and cognitive variables that drive people's engagement with information that is *prima facie* unlikely (would a political party *really* run a child sex ring out of a downtown pizza parlor?), unsubstantiated by any evidence (other than Democratic staffers ordering or eating pizza), and easily disproven (there was no basement). Why do people believe, and more importantly in the present context, why do they *share* such information? And who are the people who share low-quality information? Are there any individual characteristics that predict who might share low-quality information? Finally, how can we describe and understand people's engagement and sharing of low-quality information at a macro level? How do strategic actors and political-manipulation entrepreneurs interface with the public?

## **Human attention and misinformation**

Journalists have long known that "if it bleeds, it leads". People seek out news that is predominantly negative (Soroka et al., 2019) or awe-inspiring (Berger & Milkman, 2012), and they preferentially share messages couched in moral-emotional language (Brady et al., 2017). Digital media amplify the role of emotion because the degree of moral outrage elicited by reports online is considerably greater than for encounters in person or in conventional media (Crockett, 2017). By design or otherwise, false content exploits this attentional bias: Misinformation on Facebook during the 2016 US presidential campaign was particularly likely to provoke voter outrage (Bakir & McStay, 2018) and fake news titles have been found to be substantially more negative in tone, and display more negative emotions such as disgust and anger, than real news titles (Paschen, 2019).

One factor determining sharing therefore lies in the nature of the material being created in the first place. This affords an opportunity for malicious content creators to design misleading or false material that is likely to be shared because it exploits these attentional biases. The intentions of content creators may therefore differ from those of users who subsequently share the content. Users may sincerely believe content that they find alarming, or they may be unsure what to think and hope to prompt discussion or commentary from others by sharing (e.g., because if it *were* true it would be interesting; Altay et al., 2021).<sup>1</sup>

## **Online reputation and sharing**

At first glance, recent research has repeatedly shown that only relatively few people actually share fake news (Grinberg et al., 2019; Guess et al., 2019). For example, during the US Presidential campaign in 2016, only 0.1% of Twitter users, known as

“superspreaders”, were responsible for 80% of retweets of fake news (Grinberg et al., 2019). This figure likely represents an underestimate because much of the relevant research on fake news has been confined to tracking the sharing of weblinks to a limited set of fake news websites (Kozyreva et al., 2020). It is likely that the number of people who share misleading or false content is considerably greater if all possible sources were considered. Unfortunately this is not readily traceable because automatic content classification is imperfect.

Nonetheless, it is clear that the majority of people do not share misleading content. Indeed, the majority of people indicate that they do not share any content at all (Roozenbeek & van der Linden, 2020). One reason people generally seek to avoid sharing of low-quality information is because they do not wish to jeopardize their reputation (Altay et al., 2020; Waruwu et al., 2020). When people are asked how much they would need to be paid to share fake news from their personal account, somewhere between 40% and 50% of participants insisted on payment of \$1,000 or more, with only around 20% indicating willingness to share for free (Altay et al., 2020). Users also report being embarrassed when they shared news that others in their group discovered to be false (Waruwu et al., 2020). This response appears well-calibrated to the fact that sharing of a single fake news story (out of four in total) causes a dramatic decline in trust in the source (Altay et al., 2021).

The desire to preserve one’s reputation thus provides a major safeguard against the sharing of low-quality information. There are, however, several factors that can undermine that safeguard. Those factors range from the purely cognitive to the political.

### **Information load, cognitive capacity, and decision quality**

The increasing abundance of information online has measurable consequences: Whereas in 2013 the most popular “hashtags” on Twitter remained popular for 17.5 hours, by 2016 a hashtag’s life in the limelight had dropped to 11.9 hours (Lorenz-Spreen et al., 2019). The same declining half-life was observed for Google queries and movie ticket sales (Lorenz-Spreen et al., 2019). If one takes these measures to be a proxy for the “attentional load” of the global public, such that people have to divide and shift their attention increasingly rapidly between different sources, then it is unsurprising that at least some people become susceptible to fake news and begin to share it. There is much evidence that information overload makes it harder for people to make good decisions about what to look at, spend time on, what to believe, and what to share (Hills, 2019; Hills et al., 2013). To illustrate, whereas traditional news consumption entailed relatively few decisions (e.g., which newspaper to buy or subscribe to), we now face a multitude of online micro-decisions for every article that we choose to read from a scattered array of sources. Although this potentially increases the diversity of our news diet, it also multiplies the opportunities for error and renders careful examination of the trustworthiness of a source increasingly difficult. Information overload can also contribute to polarization and dysfunctional disagreement between well-meaning and rational actors (Pothos et al., 2021). That is, despite their good-faith efforts, overload may prevent actors from forming compatible representations of complex problems: The complexity mandates a simplification of representations, which necessarily introduces the potential for conflict and incompatibilities between actors that may result in persistent disagreement (Pothos et al., 2021).

An obvious corollary to the adverse consequences of information overload are the observed effects of individual differences in various measures of cognitive capacity.

Belief in fake news – and intention to share – has been associated with insufficient analytic reasoning and deliberation (Pennycook & Rand, 2019), and the ability to resist false information after it has been corrected is correlated with working memory capacity (Brydges et al., 2018). Moreover, a consistent finding in the literature is that older individuals are more likely to consume and share fake news (Grinberg et al., 2019, Guess et al., 2020c, 2019, 2021). This finding is particularly troubling in light of the fact that, at least in the United States, older citizens are more likely to vote than any other age group (Brashier & Schacter, 2020). All these findings point to the important role of cognitive skill or analytic capacity in enabling people to resist sharing fake news.<sup>2</sup>

A further metacognitive skill that has been linked to the spreading of fake news is the calibration between self-perceived ability and actual ability to discern between true and false information. A recent large-scale study of American news consumers ( $N \approx 15,000$ ) found that the vast majority of people overestimate their ability to differentiate between legitimate and false headlines. The greater the extent of overestimation, the more likely people were to visit fake news sites and share false content (Lyons et al., 2021). Over-claiming of knowledge has also been associated with preferences for right-wing populist parties (van Kessel et al., 2020) and anti-establishment voting (van Prooijen & Krouwel, 2020). It turns out that in addition to over-claiming of knowledge, the consumption and sharing of fake news is also preferentially associated with right-wing or conservative political views.

## **Conservatism**

The cognitive and psychological differences between liberals and conservatives have been subject to a large body of research (for a recent review, see Jost, 2017). Research attention has recently also turned to differences in susceptibility to fake news or other low-quality information between partisans of different stripes. One line of research has involved carefully curated stimuli that presented “bullshit”; that is, utterances designed to impress but generated without any concern for the truth (Pennycook et al., 2015). For example, Sterling et al. (2016) randomly generated sentences from a set of buzzwords, yielding syntactically correct but meaningless stimuli such as “we are in the midst of a self-aware blossoming of being that will align us with the nexus itself” or “consciousness is the growth of coherence, and of us”. People who were more inclined to endorse neoliberal economics were found to be more likely to rate these statements as profound. A similar association was reported by Pfattheicher and Schindler (2016) involving general conservatism, and Fessler et al. (2017) found that more conservative participants exhibited greater credulity for (false) information about potential hazards. For example, conservatives were more likely to believe that kale contains thallium than liberals (there is no good evidence that it does).

Another line of research has focused on big-data analyses in the “real world”, by examining consumption and sharing behavior on various social media platforms, such as Twitter (Grinberg et al., 2019; Nikolov et al., 2021) and Facebook (Guess et al., 2019, 2021), or by tracking browsing behavior over extended periods (Guess et al., 2020b, 2020c). Virtually all of those studies have focused on American audiences. The consistent finding that has emerged from this research program is that conservatives are more susceptible to consumption and sharing of misinformation. To illustrate, a recent large-scale study presented participants with 20 of the most popular news stories harvested from

social media every two weeks across a period of nearly six months. The statements could be unambiguously classified as true or false. It was found that conservatives are more likely to hold misconceptions than liberals, and are also somewhat less likely to believe in factual statements (Garrett & Bond, 2021). Although these data present a fairly unambiguous picture, with conservatives sharing more false information than liberals, there is evidence that intensity of partisanship contributes to the sharing of false information at the other end of the political spectrum as well (Nikolov et al., 2021).

I now turn to a closer examination of the motives and mechanisms underlying the sharing of false information. Although the data just reviewed showed that belief in false information (e.g., Garrett & Bond, 2021) and sharing of fake news (e.g., Grinberg et al., 2019) go hand in hand and are associated with a common driver – in this instance conservatism – other results, mainly from experimental studies, suggest that belief can be decoupled from sharing intentions (Pennycook & Rand, 2021). For example, in a large online experiment, Pennycook et al. (2020) presented participants with COVID-19 related headlines that were true or false. When asked about accuracy, participants were able to differentiate between true and false headlines, but when asked about their sharing intention, veracity of the headline had little impact. In another study (Pennycook et al., 2021), 16% of shares of false news headlines occurred despite the headline being identified as inaccurate. Although this figure may appear relatively modest, when scaled up to the information ecosystem as a whole, the purposeful sharing of false information significantly contributes to deterioration of information quality online, in particular because other users may amplify the cascade by sharing the false information unwittingly. Why, then, do some people knowingly share false information? I address this question through the lens of “participatory propaganda” (e.g., Asmolov, 2018; Wanless & Berk, 2017).

## **Participatory propaganda**

Digital media have obliterated the distinction between “propagandist” (e.g., organs of a totalitarian government) and the target audience. Instead of public opinion being manipulated by direct one-way communication, perceptions are now also shaped by co-opting community members, willingly or unwittingly, in shaping perceptions, cognitions, and behaviors of the target audience (Wanless & Berk, 2017). It is the opportunistic involvement of community members that characterizes participatory propaganda and identifies it as a distinct object of study. Existing research has identified several interesting threads, mainly through case studies. I review some of this research here, but the paucity of existing data is also opening up numerous avenues for future empirical research.

### *The digital tactics of strategic actors*

A core aspect of participatory propaganda is that it can only unfold after the disinformation ecosystem has been seeded by strategic actors in pursuit of a campaign. These strategic actors may range from individuals (e.g., Donald Trump’s use of Twitter to divert public attention; Lewandowsky et al., 2020a) to state-sponsored actors relying on an assortment of computational-propaganda tools. Wanless and Berk (2017) identified six digital tactics that strategic actors can use to seed and shape a participatory propaganda cascade.

- (1) Micro-targeting of an audience through selective advertising on social media to maximize the match between a persuasive message and its intended target audience (Matz et al., 2017).
- (2) Design of provocative content that evokes outrage or creates emotive memes that are more readily shared (Brady & Crockett, 2019; Spring et al., 2018).
- (3) Exploiting the existence, or encouraging the formation, of echo chambers. Echo chambers refer to informational silos of partisans, with little cross-party exchange of information (Guess et al., 2020c).<sup>3</sup>
- (4) Gaming of search results (Metaxas, 2009; Metaxas & DeStefano, 2005), for example by cross-linking false information on different sites, using automated “botnets”, or exploiting of data voids (Golebiewski & boyd, 2019).
- (5) Encouraging followers to share information or engage in other participatory forms of manipulation. The Chinese government frequently exhorts users to repost information, sometimes even offering rewards for shares (Repnikova & Fang, 2018).
- (6) Use of mainstream media, for example by creating trends on social media (via hashtags and so on) that are then picked up and reported by mainstream media. The inauthenticity of the original information is typically lost once mainstream media report about trends on social media.

Use of one or several of these digital tactics permits strategic actors to seed a cascade of participatory propaganda. Once started, cascades can be monitored by online tools that afford further opportunity to strategic actors for shaping and fine-tuning (Wanless & Berk, 2017). The small literature on participatory propaganda has focused mainly on how the audience, which includes the media and the public at large, respond and get involved in strategically seeded information cascades.

### ***Involvement of the media in participatory propaganda***

I begin by considering the unwitting involvement of major media, which have been shown to be susceptible to manipulation by strategic actors, notwithstanding their express commitment to resist such manipulation. Specifically, it has been shown that leading American media (*New York Times* and *ABC Headline News*) adjusted their coverage in response to (frequently) misleading utterances by Donald Trump during his presidency. Much has been written about Trump’s masterful use of Twitter (e.g., Enli, 2017; Lee & Xu, 2018). Of particular relevance here is a recent text corpus analysis that explored potential associations between leading mainstream media coverage in the US and Donald Trump’s tweets. The study showed that Trump’s tweets demonstrably diverted media attention away from issues that were damaging to the president (Lewandowsky et al., 2020a).

Beyond affecting media coverage, Trump’s misleading or false statements, often but not exclusively communicated on Twitter, tended to trigger supportive information cascades on social media propagated by his millions of followers, culminating in the violent insurrection on 6 January 2021 that was motivated by Trump’s fabricated claim that his reelection had been “stolen” from him. Although this claim was shown to be false by virtually all mainstream media in the US and dismissed by the courts, it was able to gather pace on social media (Kirk & Schill, 2021; Munn, 2021). In the five months following the 6 January insurrection, across 23 surveys an average of 72% of Republicans and 78% of Trump voters denied that Biden was the legitimate winner of the election, with no sign of a downward trend (Jacobson, 2021).

### ***When expressive responding trumps perception***

The participatory adherence to misinformation among followers of Donald Trump can be traced back to the very early days of his presidency, when White House officials falsely claimed that more people attended Trump's inauguration than any other previously. This claim was readily falsifiable by a range of evidence, including public transport data (ridership of the Washington, DC, Metro system) and photographs of the crowds present on the National Mall during the inauguration. Nonetheless, the Trump administration stuck to the claim and it soon became a polarizing issue. Schaffner and Luks (2018) conducted a study within two days of the controversy erupting that explored the impact of the administration's claim. Participants were presented with two side-by-side photographs of the inaugurations of Barack Obama in 2009 and Donald Trump in 2017 and were asked to choose the photo that had more people in it. The photographs were unlabeled, and the differences in crowd size so unambiguous, that it was virtually impossible for honest responses to be incorrect. Accordingly, among non-voters and Clinton voters, only 3% and 2% of respondents, respectively, chose the incorrect picture. Among Trump voters, by contrast, this proportion was 15% overall. When responses were broken down further by level of respondents' education, the error rate rose to 26% among highly educated Trump voters, compared to 1% for highly educated Clinton voters. For participants with low education, the gap between Trump (11%) and Clinton (2%) voters was considerably smaller. Given that the evidence was unequivocal and the task trivial, Schaffner and Luks (2018) interpreted these results as revealing "expressive responding" of partisans. More highly educated respondents were more likely to recognize the unlabeled pictures, thus providing an incentive to express their support for Trump in this controversy. Instead of genuinely believing a misconception, partisans effectively chose to participate in propaganda on behalf of their leader, even if in this instance the audience was merely an unknown experimenter.

The study by Schaffner and Luks (2018) provides a striking visualization of partisans' willingness to set aside unambiguous perceptual evidence in favor of participating in the promulgation of a politically concordant falsehood. The proportion of people who were willing to do this meshes well with the proportion of knowing shares of false headlines observed by Pennycook et al. (2021). The moment people share information online that they know to be false, they become a willing agent in the participatory propaganda ecosystem.

### ***The hallmarks of participatory propaganda***

Several state actors have been identified as strategic actors that often rely on triggering participatory propaganda. Foremost among them are the Russian (Paul & Matthews, 2016) and Chinese (King et al., 2014; Roberts, 2018) governments. It has been estimated that the Chinese government posts around 450 million social media comments per year (King et al., 2017), much of it posted by a "50-cent army" of operatives who are paid to disseminate messages on the regime's behalf. A notable aspect of those messages is that they frequently involve distraction rather than attempts to persuade (King et al., 2017): Speech that is considered inconvenient or dangerous by a regime or other strategic actors is drowned out rather than being banned or confronted outright (Applebaum, 2019). Asmolov (2019, p. 13) expressed this mechanism particularly well:

propaganda has become less interested in changing people's opinion about a specific object or in convincing people that it is either truth or fiction. The main purpose of 21st century propaganda is to increase the scope of participation in relation to the object of propaganda. In a digital environment relying on user participation, propaganda is a technology of power that drives the socialization of conflicts and a tool for increasing the scope of contagion. While participation in political debates is often considered to be an important feature of democracy, propaganda allows us to define the structure and form of participation in a way that serves only those who generate propaganda, and minimizing the constructive outcomes of participation.

Crucial to the success of this new form of propaganda is the opportunistic participation by users who willingly spread information that supports their political views, regardless of whether or not they know it to be false (Starbird, 2019). It is this "entanglement of sincere activists, journalists, and [misleading] information operations" (Wilson & Starbird, 2020, p. 6) that contributes to the success of participatory propaganda. Several hallmarks of participatory propaganda and its success have been identified.

#### *Obscuring origin and existence*

Public participation in the sharing, and also shaping (e.g., via quoted Tweets), of disinformation obscures the origin and intent of the original source. "Fake news" remains a work of fiction until the audience mistakes it for real news (Tandoc et al., 2018), and cascades of misinformation can contain any number of combinations of beliefs about the veracity of the information (Giglietto et al., 2019). As the cascade unfolds, chances are that an increasing proportion of users involved are unaware of the falsehood of the information they share. The mixture of news, entertainment, and personal information that characterizes most social-media feeds makes it particularly difficult to discern the intention behind any given piece of information (Starbird et al., 2019).

As a result, participatory propaganda not only obscures the origins of strategic disinformation campaigns but it may also obscure their very presence. Starbird et al. (2019) reports a case-study of strategic computational propaganda in which existing online communities surrounding the #BlackLivesMatter movement became unwitting hosts of a Russian state-sponsored influence operation.

#### *The coveted authentic users*

Over time, information that is entrenched in a participatory propaganda cascade becomes "internalized". That is, "external cultural artifacts are integrated into the cognitive process and help to define our human relationship with reality" (Asmolov, 2019, p. 14). For example, "likes" on social media become ingrained into our views of the information we encounter, and eventually the strategic origins of the information are completely lost and the cascades are sustained by the unwitting involvement of "authentic" users. Authentic and unwitting users are the most coveted audience for strategic actors (Wanless & Berk, 2019), in part because platforms have begun to remove inauthentic accounts more aggressively.

The unwitting involvement of authentic users also serves to provide "social proof" of the information via consensus effects. People's attitudes towards controversial material

are in part shaped by their perceptions of the social consensus among other readers (e.g., Lewandowsky et al., 2019a). The more people think that other readers agree on a position, the more likely they are to be swayed in that direction, irrespective of the particular *content* being circulated (Lewandowsky et al., 2019a). Successful participatory propaganda cascades can therefore send powerful social signals that help shape public perception through (false and strategically orchestrated) perceptions of consensus.

### *Conflict in perpetuity*

Participatory propaganda is characterized by two further attributes: First, it is almost perpetual. Most campaigns are launched by strategic actors long before an election or other target events (Wanless & Berk, 2021). Second, all participatory propaganda cascades are inevitably conflictual (Asmolov, 2019). The objective nearly always is to polarize and enhance conflict, rather than to persuade.

### **Conclusions**

Participatory propaganda has only recently been identified as a distinct form of propaganda that deserves to be studied in its own right. The literature on this concept is still in its infancy and is dominated by case-studies and conceptual analyses. Although these analyses have provided a rough first sketch of the terrain, much empirical work remains to be done to turn this sketch into an accurate map. Embarking on this empirical journey is an urgent task because of the applied implications of the issues just reviewed. Text box 20.1 provides an example of how participatory propaganda can be studied in the classroom.

#### **Text box 20.1 A demonstration experiment**

Some of the phenomena described in this chapter are difficult to demonstrate in the “classroom”, whether it is a virtual meeting place or a physical space, because they are observable mainly at scale and involve a relatively small proportion of people. I suggest instead to use a version of a message-framing study reported by Altay and Mercier (2020), which shows that people’s willingness to share identical information online can be dramatically affected by the way it is framed. This text box provides all necessary details to set up a simplified classroom experiment.

Although the original study investigated six different messages in as many conditions, I suggest using only two messages that were found to differ maximally in eliciting sharing intentions.

### **Materials**

The materials involve information about the safety of vaccines, an issue that is highly topical at the time of this writing. The messages for this experiment present identical statistics about the medical consensus on vaccinations but using two different frames:

- “90% of medical scientists think that vaccines are safe.”
- “10% of medical scientists don’t think that vaccines are safe.”

Setting aside the possibility of ambiguity (“don’t know” responses), both messages convey an identical statistic, namely the scientific consensus on vaccinations which is known to be a powerful communication tool to assuage public concern about vaccine safety (van der Linden et al., 2015). Nonetheless, Altay and Mercier (2020) found that the positive framing (“90% ... ”) elicited far greater willingness to share the message on social media ( $M = 3.87$  on a five-point scale) than the negative framing (“10% ... ”;  $M = 2.03$ ).

### **Power and participants**

The effect size for the between-groups comparison between these two messages computed from Altay and Mercier’s Table 1 was  $d = 1.56$ , computed using GPower (Mayr et al., 2007). To detect an effect of that size with a power of .90 and  $\alpha = .05$  (using a two-tailed t-test) would require ten participants in each group. To detect an effect that is half the size as that reported by Altay and Mercier would require  $n = 36$  per group.

### **Procedure**

Participants are randomly assigned to read one or the other statement, before answering the question “How likely would you be to pass along this statement to other people.” Responses are collected on a five-point Likert scale ranging from 1 (very unlikely to pass along) to 5 (very likely to pass along). Self-reports of sharing intention (using a ternary scale no/maybe/yes) have been shown to correlate well ( $r = .44$ ) with actual sharing on Twitter (Mosleh et al., 2020).

### **Analysis plan**

The analysis consists of a simple between-groups *t*-test on the numeric responses to the single item. Each participant contributes one observation to the analysis.

## **Applied implications**

At least two questions arise from the issues just reviewed. First, how relevant is the phenomenon of participatory propaganda to society as a whole? Second, assuming it is relevant, how can it be counteracted?

Concerning the first question, most recent big-data analyses of online information diets consistently, and at first glance encouragingly, concluded that misinformation constitutes only a small share of overall news consumption in the US (Allen et al., 2020; Grinberg et al., 2019; Guess et al., 2019, 2020b, 2020c, 2021). However, those conclusions are all based on analysis of a limited number of websites that are known purveyors of fake news or other low-quality information. Any consumption or sharing of news on those sites is therefore a fairly unambiguous marker of misinformation. While this limited focus provides methodological rigor and avoids fuzzy boundaries between “true” and “false” information, it also necessarily ignores other sources of misinformation, thus providing only a lower-bound estimate of people’s misinformation exposure (Kozyreva et al., 2020). Evidence that the scale of the problem may be bigger

than is apparent from analyses that focus on fake-news websites was provided in a study by Vargo et al. (2018), which showed that, although fake news did not dominate the media landscape (during 2014–2016), it was closely intertwined with American partisan media (e.g., Fox News); each influenced the other's agendas across a wide range of topics, including the economy, education, the environment, international relations, religion, taxes, and unemployment. Thus, fake news sources may not feature prominently in people's news diet, but it can nonetheless set political agendas and determine media coverage. Participatory propaganda plays an integral part in this process, as illustrated by the "Pizzagate" affair discussed at the outset.

A further cautionary element arises from theoretical research showing how motivated minorities can significantly destabilize a majority of actors in informational networks (Galam, 2002; Juul & Porter, 2019; Lewandowsky et al., 2019b). For example, in an agent-based simulation, Lewandowsky et al. (2019b) showed that an evidence-resistant minority (i.e., science "deniers") can delay the formation of a scientific consensus and, when given disproportionate prominence in the media, can prevent the public from recognizing the existence of a scientific consensus. Taken together, the available evidence warrants concern about the effects of misinformation, spread via participatory propaganda, in democratic societies.

In light of this concern, what countermeasures can be implemented? A considered treatment of this question is beyond the scope of this chapter; in-depth explorations of those issues can be found in Kozyreva et al. (2020) and Lewandowsky et al. (2020b). Nonetheless, it is worth providing a thumbnail sketch of available options. Epstein et al. (2021) investigated a toolkit of options to enhance the accuracy of people's sharing decisions. They found that four interventions decreased the sharing of false information about COVID-19:

- A "long evaluation" treatment, consisting of judging the accuracy of eight non-COVID-related headlines (half of which were true, half false), with feedback provided after each judgment.
- An "importance" treatment, consisting of asking participants "How important is it to you that you share only news articles on social media (such as Facebook and Twitter) if they are accurate?"
- A "tips" treatment consisting of four simple digital literacy tips, taken from an intervention developed by Facebook (Guess et al., 2020a).
- A combination of the "tips" treatment and providing normative information, namely that eight out of ten past survey respondents said it was "very important" or "extremely important" to share only accurate news online, and that this was true of both Democrats and Republicans.

Another approach is known as inoculation (Lewandowsky & van der Linden, 2021). It involves people being informed about specific rhetorical techniques by which they might be misled before being exposed to them. Similar to a vaccine, this treatment is thought to develop "cognitive antibodies" which help people detect persuasive messages that seek to mislead them.

Although these interventions are readily available and can potentially be deployed at scale, they cannot by themselves be sufficient to reform an infrastructure that is built on the commodification of human attention (Lewandowsky et al., 2017). Instead, what is needed is a change in the infrastructure itself, either through platform initiatives or, where

necessary, regulation (Lewandowsky et al., 2020b). Encouragingly, it has been shown that even seemingly trivial changes to platform features can have far-reaching consequences. For example, in India in 2018, false rumours about child kidnappers shared via WhatsApp's unlimited forward facility were implicated in at least 16 mob lynchings, leading to the killings of 29 innocent people (Dixit & Mac, 2018). In response, WhatsApp introduced several small changes to their app, including the identification of forwarded messages as being forwarded (whereas previously they appeared to originate with the person who last forwarded them), and curtailing the number of recipients a message could be forwarded to at the same time (thereby slowing large cascades of messages). These relatively small changes may have contributed to the cessation of lynch killings in India since 2018 (de Freitas Melo et al., 2019).

## Summary

- There has been much public concern worldwide over misinformation and “fake news” spreading online, in particular on social media.
- But misinformation does not spread on its own, it is spread by people who choose to spread it.
- Most misinformation is spread by relatively few people – so-called “superspreaders”.
- Why do some people share information that they *know* to be false, thereby engaging in “participatory propaganda”?
- One cognitive factor that may facilitate sharing of low-quality information is cognitive overload arising from the oversupply of information sources.
- Another factor is political conservatism, which has repeatedly been found to be associated with sharing of misinformation.
- Participatory propaganda cascades are pernicious because the public’s involvement obscures the origin and intent of the original source, and because users may eventually unwittingly but authentically believe the false information.

## Notes

- 1 Other attributes of the message, such as its authoritativeness or markers of popularity, play no apparent role in self-reported sharing intentions (Buchanan, 2020).
- 2 Although cognitive capacity is known to decline with age, the observed association between advanced age and increased fake-news consumption arguably cannot be explained by cognitive decline alone. Other factors such as social goals and insufficient digital literacy may also play a role (Brashier & Schacter, 2020).
- 3 The nature of echo chambers, in particular whether they are the result of algorithmic streaming of newsfeeds or emerge from users’ self-selected aggregation, is a matter of intense academic debate (Stewart et al., 2019).

## Further reading

A broader and more complete theoretical account of what drives sharing online has been provided by Van Bavel et al. (2021). The broader background of how human cognition interacts with online technology was explored by Kozyreva et al. (2020) and Lorenz-Spreen et al. (2020). Anyone interested in the policy implications of the tension between human cognition and online technologies can consult the extensive report for the European Commission provided by Lewandowsky et al. (2020b).

## Acknowledgment

Preparation of this paper was facilitated by a Humboldt Award from the Humboldt Foundation to the author and by a grant from the Volkswagen Foundation for the project “Reclaiming individual autonomy and democratic discourse online.” The author was also supported by an ERC Advanced Grant (PRODEMINFO) during part of this work. Address correspondence to the author at the School of Psychological Science, University of Bristol, 12a Priory Road, Bristol BS8 1TU, United Kingdom. E-mail: stephan.lewandowsky@bristol.ac.uk. Personal web page: [www.cogsciwa.com](http://www.cogsciwa.com).

## References

- Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6, eaay3539.
- Altay, S., de Araujo, E., & Mercier, H. (2021). “If this account is true, it is most enormously wonderful”: Interestingness-if-true and the sharing of true and false news. *Digital Journalism*. doi: 10.1080/21670811.2021.1941163.
- Altay, S., Hacquin, A.-S., & Mercier, H. (2020). Why do so few people share fake news? It hurts their reputation. *New Media & Society*. doi: 10.1177/1461444820969893
- Altay, S., & Mercier, H. (2020). Framing messages for vaccination supporters. *Journal of Experimental Psychology: Applied*, 26, 567–578.
- Applebaum, A. (2019). The new censors won’t delete your words—they’ll drown them out. Retrieved from [www.washingtonpost.com/amphhtml/opinions/global-opinions/the-new-censors-wont-delete-your-words--theyll-drown-them-out/2019/02/08/c8a926a2-2b27-11e9-984d-9b8fba003e81\\_story.html](http://www.washingtonpost.com/amphhtml/opinions/global-opinions/the-new-censors-wont-delete-your-words--theyll-drown-them-out/2019/02/08/c8a926a2-2b27-11e9-984d-9b8fba003e81_story.html)
- Asmolov, G. (2018). The disconnective power of disinformation campaigns. *Journal of International Affairs*, 71, 69–76.
- Asmolov, G. (2019). The effects of participatory propaganda: From socialization to internalization of conflicts. *Journal of Design and Science*, 6, 1–25.
- Bakir, V., & McStay, A. (2018). Fake news and the economy of emotions. *Digital Journalism*, 6, 154–175.
- Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49, 192–205.
- boyd, d. (2017). Hacking the attention economy. *Data & Society*. Retrieved from <https://points.datasociety.net/hacking-the-attention-economy-9fa1daca7a37>
- Brady, W. J., & Crockett, M. (2019). How effective is online outrage? *Trends in Cognitive Sciences*, 23, 79–80.
- Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114, 7313–7318.
- Brashier, N. M., & Schacter, D. L. (2020). Aging in a fake news era. *Current Directions in Psychological Science*, 29, 316–323.
- Brydges, C., Gignac, G., & Ecker, U. (2018). Working memory capacity predicts ongoing reliance on misinformation: A latent-variable analysis. *Intelligence*, 69, 117–122.
- Buchanan, T. (2020). Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation. *PLOS One*, 15, e0239666.
- Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, 1, 769–771.
- de Freitas Melo, P., Vieira, C. C., Garimella, K., de Melo, P. O. V., & Benevenuto, F. (2019). Can WhatsApp counter misinformation by limiting message forwarding? In H. Cherifi, S. Gaito, J. Mendes, E. Moro, & L. Rocha (Eds.), *International conference on complex networks and their applications VIII. COMPLEX NETWORKS 2019* (pp. 372–384). Studies in Computational Intelligence, 881. Cham: Springer. [https://doi.org/10.1007/978-3-030-36687-2\\_31](https://doi.org/10.1007/978-3-030-36687-2_31)

- Directorate-General for Communication. (2018). *Flash Eurobarometer 464: Fake news and disinformation online*. Retrieved from <https://op.europa.eu/en/publication-detail/-/publication/2d79b85a-4cea-11e8-be1d-01aa75ed71a1/language-en>
- Dixit, P., & Mac, R. (2018). How WhatsApp destroyed a village. Retrieved from [www.buzzfeednews.com/article/pranavdixit/whatsapp-destroyed-village-lynchings-rainpada-india](http://www.buzzfeednews.com/article/pranavdixit/whatsapp-destroyed-village-lynchings-rainpada-india)
- Enli, G. (2017). Twitter as arena for the authentic outsider: Exploring the social media campaigns of Trump and Clinton in the 2016 US presidential election. *European Journal of Communication*, 32, 50–61.
- Epstein, Z., Berinsky, A. J., Cole, R., Gully, A., Pennycook, G., & Rand, D. G. (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review*, 2(3). doi: 10.37016/mr-2020-71
- Fessler, D. M. T., Pisor, A. C., & Holbrook, C. (2017). Political orientation predicts credulity regarding putative hazards. *Psychological Science*, 28, 651–660.
- Fisher, M., Cox, J. W., & Hermann, P. (2016). Pizzagate: From rumor, to hashtag, to gunfire in DC. Retrieved from [www.washingtonpost.com/local/pizzagate-from-rumor-to-hashtag-to-gunned-in-dc/2016/12/06/4c7def50-bbd4-11e6-94ac-3d324840106c\\_story.html](http://www.washingtonpost.com/local/pizzagate-from-rumor-to-hashtag-to-gunned-in-dc/2016/12/06/4c7def50-bbd4-11e6-94ac-3d324840106c_story.html)
- Frankovic, K. (2016). Belief in conspiracy theories depends largely on which side of the spectrum you fall on. Retrieved from <https://today.yougov.com/topics/politics/articles-reports/2016/12/27/belief-conspiracies-largely-depends-political-iden>
- Galam, S. (2002). Minority opinion spreading in random geometry. *European Physical Journal B*, 25, 403–406.
- Garrett, R. K., & Bond, R. M. (2021). Conservatives' susceptibility to political misperceptions. *Science Advances*, 7, eabf1234.
- Giglietto, F., Iannelli, L., Valeriani, A., & Rossi, L. (2019). "Fake news" is the invention of a liar: How false information circulates within the hybrid news system. *Current Sociology*, 67, 625–642.
- Golebiewski, M., & boyd, d. (2019). *Data voids: Where missing data can easily be exploited* (Tech. Rep.). N.pl.: Data & Society.
- Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363, 374–378.
- Guess, A., Aslett, K., Tucker, J., Bonneau, R., & Nagler, J. (2021). Cracking open the news feed. *Journal of Quantitative Description: Digital Media*, 1.
- Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020a). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, 117, 15536–15545.
- Guess, A. M., Lockett, D., Lyons, B., Montgomery, J. M., Nyhan, B., & Reifler, J. (2020b). "Fake news" may have limited effects on political participation beyond increasing beliefs in false claims. *Harvard Kennedy School Misinformation Review*, 1(1). doi: 10.37016/mr-2020-004
- Guess, A. M., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5, eaau4586.
- Guess, A. M., Nyhan, B., & Reifler, J. (2020c). Exposure to untrustworthy websites in the 2016 U.S. election. *Nature Human Behavior*, 4, 472–480.
- Hills, T. T. (2019). The dark side of information proliferation. *Perspectives on Psychological Science*, 14, 323–330.
- Hills, T. T., Noguchi, T., & Gibbert, M. (2013). Information overload or search-amplified risk? Set size and order effects on decisions from experience. *Psychonomic Bulletin & Review*, 20, 1023–1031.
- Jacobson, G. C. (2021). Driven to extremes: Donald Trump's extraordinary impact on the 2020 elections. *Presidential Studies Quarterly*, 51, 492–521. doi: 10.1111/psq.12724
- Jost, J. T. (2017). Ideological asymmetries and the essence of political psychology. *Political Psychology*, 38, 167–208.
- Juul, J. S., & Porter, M. A. (2019). Hipsters on networks: How a minority group of individuals can lead to an antiestablishment majority. *Physical Review E*, 99, 022313.

- King, G., Pan, J., & Roberts, M. E. (2014). Reverse-engineering censorship in China: Randomized experimentation and participant observation. *Science*, 345, 1251722–1251722.
- King, G., Pan, J., & Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, 111, 484–501.
- Kirk, R., & Schill, D. (2021). Sophisticated hate stratagems: Unpacking the era of distrust. *American Behavioral Scientist*. doi: 10.1177/00027642211005002
- Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the Internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, 21, 103–156.
- Lee, J., & Xu, W. (2018). The more attacks, the more retweets: Trump's and Clinton's agenda setting on twitter. *Public Relations Review*, 44, 201–213.
- Lewandowsky, S., Cook, J., Fay, N., & Gignac, G. E. (2019a). Science by social media: Attitudes towards climate change are mediated by perceived social consensus. *Memory & Cognition*, 47, 1445–1456.
- Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the post-truth era. *Journal of Applied Research in Memory and Cognition*, 6, 353–369.
- Lewandowsky, S., Ecker, U. K. H., Seifert, C., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13, 106–131.
- Lewandowsky, S., Jetter, M., & Ecker, U. K. H. (2020a). Using the president's tweets to understand political diversion in the age of social media. *Nature Communications*, 11, 5764.
- Lewandowsky, S., Pilditch, T. D., Madsen, J. K., Oreskes, N., & Risbey, J. S. (2019b). Influence and seepage: An evidence-resistant minority can affect public opinion and scientific belief formation. *Cognition*, 188, 124–139.
- Lewandowsky, S., Smillie, L., Garcia, D., Hertwig, R., Weatherall, J., Egidy, S., ... Leiser, M. (2020b). *Technology and democracy: Understanding the influence of online technologies on political behaviour and decision making*. Tech. Rep. EUR 30422 EN. Luxembourg: Publications Office of the European Union.
- Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, 32, 348–384. doi: 10.1080/10463283.2021.1876983
- Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R., & Hertwig, R. (2020). How behavioural sciences can promote truth and, autonomy and democratic discourse online. *Nature Human Behaviour*, 4, 1102–1109.
- Lorenz-Spreen, P., Mønsted, B. M., Hövel, P., & Lehmann, S. (2019). Accelerating dynamics of collective attention. *Nature Communications*, 10, 1759.
- Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B., & Reifler, J. (2021). Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences*, 118, e2019527118.
- Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 48, 12714–12719.
- Mayr, S., Erdfelder, E., Buchner, A., & Faul, F. (2007). A short tutorial of GPower. *Tutorials in Quantitative Methods for Psychology*, 3, 51–59.
- Metaxas, P., & Finn, S. (2019). Investigating the infamous #Pizzagate conspiracy theory. *Technology Science*. Retrieved from <https://techscience.org/a/2019121802/>
- Metaxas, P. T. (2009). Web spam, social propaganda and the evolution of search engine rankings. In *International conference on web information systems and technologies* (pp. 170–182). Heidelberg: Springer Verlag.
- Metaxas, P. T., & DeStefano, J. (2005). Web spam, propaganda and trust. Adversarial information retrieval (airweb), WWW 2005 Conference, Chiba, Japan.

- Mitchell, A., Gottfried, J., Stocking, G., Walker, M., & Fedeli, S. (2019). Many Americans say made-up news is a critical problem that needs to be fixed. Retrieved from [www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/](http://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/)
- Mosleh, M., Pennycook, G., & Rand, D. G. (2020). Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *PLOS One*, 15, e0228882.
- Munn, L. (2021). More than a mob: Parler as preparatory media for the U.S. Capitol storming. *First Monday*. doi: 10.5210/fm.v26i3.11574
- Nikolov, D., Flammini, A., & Menczer, F. (2021). Right and left, partisanship predicts (asymmetric) vulnerability to misinformation. *Harvard Kennedy School Misinformation Review*. doi: 10.37016/mr-2020-55
- Paschen, J. (2019). Investigating the emotional appeal of fake news using artificial intelligence and human contributions. *Journal of Product and Brand Management*, 29, 223–233.
- Paul, C., & Matthews, M. (2016). *The Russian “firehose of falsehood” propaganda model*. Santa Monica, CA: RAND Corporation (Tech. Report).
- Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making*, 10, 549–563.
- Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592, 590–595.
- Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31, 770–780.
- Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39–50.
- Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, 25, 388–402.
- Pfattheicher, S., & Schindler, S. (2016). Misperceiving bullshit as profound is associated with favorable views of Cruz, Rubio, Trump and conservatism. *PLOS One*, 11, e0153419.
- Pothos, E. M., Lewandowsky, S., Basieva, I., Barque-Duran, A., Tapper, K., & Khrennikov, A. (2021). Information overload for (bounded) rational agents. *Proceedings of the Royal Society B: Biological Sciences*, 288, 20202957.
- Reznikova, M., & Fang, K. (2018). Authoritarian Participatory Persuasion 2.0: Netizens as thought work collaborators in China. *Journal of Contemporary China*, 27, 763–779.
- Roberts, M. E. (2018). *Censored: Distraction and diversion inside China's great firewall*. Princeton, NJ: Princeton University Press.
- Roozenbeek, J., & van der Linden, S. (2020). Breaking harmony square: A game that “inoculates” against political misinformation. *Harvard Kennedy School Misinformation Review*, 1(8). doi:10.37016/mr-2020-47
- Schaffner, B. F., & Luks, S. (2018). Misinformation or expressive responding? What an inauguration crowd can tell us about the source of political misinformation in surveys. *Public Opinion Quarterly*, 82, 135–147.
- Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, 116, 18888–18892.
- Spring, V. L., Cameron, C. D., & Cikara, M. (2018). The upside of outrage. *Trends in Cognitive Sciences*, 22, 1067–1069.
- Starbird, K. (2019). Disinformation’s spread: Bots, trolls and all of us. *Nature*, 571, 449.
- Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as collaborative work. *Proceedings of the ACM on Human-Computer Interaction*, 3, 1–26.
- Sterling, J., Jost, J. T., & Pennycook, G. (2016). Are neoliberals more susceptible to bullshit? *Judgment and Decision Making*, 11, 352–360.
- Stewart, A. J., Mosleh, M., Diakonova, M., Arechar, A. A., Rand, D. G., & Plotkin, J. B. (2019). Information gerrymandering and undemocratic decisions. *Nature*, 573, 117–121.

- Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining “fake news”. *Digital Journalism*, 6, 137–153.
- Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K., & Tucker, J. A. (2021). Political psychology in the digital (mis)information age: A model of news belief and sharing. *Social Issues and Policy Review*. Retrieved from <https://doi.org/10.31234/osf.io/u5yts>
- van der Linden, S. L., Clarke, C. E., & Maibach, E. W. (2015). Highlighting consensus among medical scientists increases public support for vaccines: Evidence from a randomized experiment. *BMC Public Health*, 15, 1–5.
- van Kessel, S., Sajuria, J., & Van Hauwaert, S. M. (2020). Informed, uninformed or misinformed? A cross-national analysis of populist party supporters across European democracies. *West European Politics*, 44, 585–610.
- van Prooijen, J.-W., & Krouwel, A. P. M. (2020). Overclaiming knowledge predicts anti-establishment voting. *Social Psychological and Personality Science*, 11, 356–363.
- Vargo, C. J., Guo, L., & Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, 20, 2028–2049.
- Wanless, A., & Berk, M. (2017). Participatory propaganda: The engagement of audiences in the spread of persuasive communications. Proceedings of the Social media and social order, culture conflict 2.0 conference.
- Wanless, A., & Berk, M. (2019). The audience is the amplifier: Participatory propaganda. In P. Baines, N. O’Shaughnessy, & N. Snow (Eds.), *The Sage handbook of propaganda* (pp. 85–104). London: Sage.
- Wanless, A., & Berk, M. (2021). The changing nature of propaganda: Coming to terms with influence in conflict. In T. Clack & R. Johnson (Eds.), *The world information war* (pp. 63–80). London: Routledge.
- Waruwu, B. K., Tandoc, E. C., Duffy, A., Kim, N., & Ling, R. (2020). Telling lies together? Sharing news as a form of social authentication. *New Media & Society*, 23, 2516–2533. doi: 10.1177/1461444820931017
- Wilson, T., & Starbird, K. (2020). Cross-platform disinformation campaigns: Lessons learned and next steps. *Harvard Kennedy School Misinformation Review*, 1(1). doi: 10.37016/mr-2020-002

# 21 Positivity biases

*Carla A. Zimmerman and W. Richard Walker*

In October 2020, a few weeks before the United States Presidential election, the second author of this chapter (RW) received an e-mail from a news outlet. The e-mail asked for an interview about the obvious and seemingly disastrous impact that positive emotions had wreaked upon the world in 2020. The rose-colored glasses that people had worn throughout the year had obviously blinded them to harsh realities that the year had forced us all to confront. The second author considered the argument the reporter was trying to make. The world in 2020 was awash in stupidity, there was wanton disregard for facts and common sense, and venom was a regular part of discourse. *Somehow positive emotions were to blame.*

To be clear, positive emotions can lead the mind astray. Positive illusions can promote unrealistic thinking that can create blind-spots preventing people from accepting painful realities. These illusions are present in our judgments of self, others, and the future, as well as in how we direct our attention and remember the past. This truth must be held in balance with another truth that is often overlooked – *a mind biased towards positivity is often healthier, more purposeful, and more able to cope with adversity than a mind that favors negativity.* This chapter will review evidence that supports this thesis in four sections. The first section examines the deleterious effects of negative emotions, specifically depression, anxiety, and anger. The second section reviews positivity biases that help people achieve their goals. The third section examines positivity biases that are related to self-protection, self-esteem, and overall psychological health. The fourth section reviews research on the Fading Affect Bias (FAB), a phenomenon that shows that most people can minimize the negativity of life events while retaining their positive aspects. Table 21.1 presents a summary of the positivity biases discussed in this chapter.

## **Depression, anxiety, and anger mislead the mind**

The negativity bias of depression is one of the most widely documented phenomena in psychology (e.g., Beck, 1976; Gotlib & Joorman, 2010; Joorman & Stanton, 2016). Indeed, the negativity bias is pervasive in the minds of the depressed; it affects their perceptions of themselves (Hards et al., 2020) and the future (Bjareheded et al., 2010). Depressed people see their lives as lacking joy, meaningful human interaction, and purpose.

Studies that examine the autobiographical memories of clinically depressed people typically find three things. First, depressed people tend to recall memories with far less detail than their non-depressed counterparts. Second, depressed people ruminate about their negative experiences more than non-depressed people. Third, depressed people

*Table 21.1* Summary of positive illusions

<i>Name</i>	<i>Definition</i>
<i>Age-related Positivity Effect</i>	Attention: A focus on positive stimuli over negative stimuli which increases with age. Memory: Greater recall of positive over negative information which increases with age.
<i>Better than Average Effect</i>	Rating oneself or close others as better than the average person in ability, traits, behaviors, and wellbeing.
<i>Fading Affect Bias</i>	Negative affect associated with autobiographical memories tends to fade more than positive affect.
<i>Optimistic Update Bias</i>	Updating judgments to a greater extent for positive outcomes compared to negative outcomes.
<i>Self-enhancing Exaggeration</i>	Exaggerating one's abilities in self-reports in comparison to objective information.
<i>Self-serving Attributions</i>	Making internal attributions for positive outcomes and external attributions for negative outcomes.
<i>Unrealistic Idealization</i>	Rating romantic partners as more similar to an ideal partner than the romantic partner would self-report.
<i>Unrealistic Optimism Effect</i>	Viewing positive outcomes for the self as more likely to occur than base-rates would suggest or as more likely for the self as compared to others.

recall negative experiences more quickly than non-depressed people (Dalgleish et al., 2007; Williams & Scott, 1988).

The question of whether the negativity bias reflects the lived experience of depressed people or a retrospective bias has also been addressed. Urban et al. (2018) examined 158 participants with a history of depression and 1,499 with no such history. Participants were asked to keep a daily record of their positive and negative emotions for eight days. At the end of the study, participants were asked to recall how often they had experienced the same positive and negative emotions during the past eight days. Participants with a history of depression experienced fewer daily positive emotions and more negative emotions compared to those without a history of depression. When examining the retrospective recall of emotions, both groups overestimated the frequency of negative emotions, but this bias was much greater for participants with a history of depression.

Anxiety is perhaps the next most damaging emotional experience. In the grips of anxiety, a person may perceive risks that are not real or exaggerate ones that pose little to no actual risk. People who are prone to anxiety often express an attentional bias towards threats (Hunt et al., 2006), a phenomenon known as threat magnification. Goodwin et al. (2017) reviewed 21 articles examining the tendency for people suffering from general anxiety disorder to exhibit biases towards threatening stimuli. They found that this bias extended across a wide variety of stimuli including threatening or negatively valenced words, images, and faces.

One of the most common but least understood anxieties that people suffer from is a fear of heights. Krupic et al. (2021) examined acrophobia using a virtual reality paradigm. Participants entered a simulated elevator that led them to the top of a skyscraper. The door would open to a wooden plank at the exit of the elevator, looking over an urban environment. These simulations worked very effectively, increasing feelings of tension,

reducing electrodermal activity, and reducing the ability of participants to focus on stimuli or thoughts other than retreat or escape.

The experience of anger is often intense, but short-lived. Because of its fleeting nature, anger is a difficult emotion to study. Researchers have successfully identified one important cognitive bias associated with it. The Hostile Attribution Bias is the tendency to interpret other people's actions as having hostile intent (Nasby et al., 1980). First identified in children, it was later found that this bias could be directly related to physical aggression (Holtzworth-Munroe & Hutchinson, 1993), relational aggression (Crick et al., 2002), and even early death (Barefoot et al., 1989).

To understand how hostile attribution relates to anger, Banks et al. (2018) examined 80 adolescents who were shown a series of animated vignettes depicting ambiguous social situations. The researchers assessed levels of aggression, hostile attribution bias, and heart rates in response to the vignettes. Most of the participants' heart rates showed a pattern of slowing followed by acceleration, which mimicked a threat response. This suggests that ambiguous information was perceived as threatening by most of the participants. However, for those participants with higher levels of hostile attribution, their heart rates remained elevated for a longer period. In short, these individuals stayed mad longer.

Taken together, the evidence is clear. The negative emotions of depression, anxiety, and anger are associated with significant biases that prevent people from accurately interpreting themselves and the circumstances that surround them.

## **Positive illusions help people attain life goals**

In this section, we review research suggesting that positive illusions help us achieve important goals, whether it be broad goals of maintaining affect or seeking knowledge, achieving academic goals, or fulfilling our need to belong through relationships with others.

As we age, the amount of time we have left decreases. This perception of decreasing time influences which goals we prioritize and which we set aside (Carstensen et al., 1999). Socioemotional selectivity theory proposes that we can classify our social goals into two basic categories – knowledge seeking and emotional regulation. When we perceive that there is ample time for us to achieve our goals, we focus on the future and seeking knowledge. When time feels limited, we focus on the present and regulating the types of emotions we experience. This phenomenon explains the age-related positivity effect – the finding that biases in memory recall and attention towards positive information increase with age (Carstensen & Mikels, 2005).

The age-related positivity effect in attention demonstrates a tendency to focus on positive stimuli that becomes stronger with age (Reed et al., 2014). For example, when eye movements are tracked during a video of two speakers discussing the positive and negative effects of aging, the amount of time spent focusing on the negative speaker decreases with age (Li et al., 2011). Younger adults should be more motivated to seek out knowledge and information; therefore, a focus on both speakers is consistent with this motivation. Older adults, however, should be motivated to maintain desired emotional states; by directing attention away from the negative speaker, they avoid an unwelcome impact on their emotions, successfully meeting their goal.

Interestingly, the age-related positivity effect appears to vary based on culture. Fung et al. (2008) studied visual attention to positive and negative faces in Chinese participants.

In contrast to the findings of Western studies, as discussed above, the older Chinese adults showed an attentional preference towards fearful faces as compared to happy faces. A later cross-cultural study noted that, while American participants showed an age-related positivity effect in attention, the opposite was seen with Chinese participants (Fung et al., 2019).

An age-related positivity effect also occurs for memory, particularly autobiographical memory. Kennedy et al. (2004) asked 300 nuns to complete a survey about their health activities, symptoms of stress, emotions, and feelings of loneliness in 1987. In 2001, they were divided into one of three conditions – to rate how they answered this survey in 1987 with a focus on accuracy, their current emotional state, or no particular focus (control). Their current mood was measured after rating their past responses. In the control group, the oldest participants showed a stronger positivity bias than the younger group; that is, the difference between 1987 and 2001 ratings was larger in the older nuns compared to the younger nuns. The older nuns in this group also experienced a larger change in positive mood after completing the survey, suggesting that this positively biased remembering of 1987 led to a more positive mood in the present. Importantly, the other two conditions demonstrated that goal activation influences the age-related positivity effect – when the goal was to focus on one's current emotions, a positivity effect was present. When the nuns were told to focus on accuracy, there was instead a negativity effect! Therefore, we see that the age-related positivity effect in attention and memory helps people to achieve relevant goals – when regulating one's current emotional state is less relevant, the age-related positivity effect disappears.

An important question in this area is whether older adults remember more positive events than younger adults, leading to a more positive perception of past events, or if older adults reinterpret their past events more positively. Research indicates that positive and negative memories are equally likely to be recalled by older and younger adults (Schryer & Ross, 2014) and these memories are similar in detail across age groups (Gallo et al., 2013). This suggests that the age-related positivity effect is not due to a selective memory for positive events in older adults, but a more positive interpretation of memories.

The benefits of positive illusions are not limited to older adults. A well-established positive illusion is the “better than average” effect, sometimes referred to as the “Lake Wobegon effect” (Kruger, 1999). Indeed, people rate themselves as better than the average person on a variety of measures, including desirable personality traits (Ziano et al., 2021), environmentally friendly behaviors (Bergquist, 2020), and driving ability (Nees, 2019). People even report that they are happier and more satisfied with life than others (Wojcik & Ditto, 2014). This tendency is exaggerated when well-being and happiness are portrayed as being especially desirable. Similarly, Brown (2012) found a stronger better than average effect for traits considered important by participants. The better than average effect extends to judgments of bias in self and others. Pronin et al. (2002) found that American college students rated themselves as less susceptible to bias than the average American and their classmates, while travelers in an airport rated themselves as less susceptible to bias than the average traveler. Notably, the better than average effect was strongest for undesirable biases and weak for socially desirable biases, suggesting that the better than average effect is beneficial for self-enhancement.

Research conducted on students suggests that positive illusions about the self improve performance and motivation. Chung et al. (2016) followed college students across their four-year academic careers. Students reported ratings of their academic ability compared to the average student at their university and their adjustment to university at multiple time points. Their grade point average (GPA), enrollment, and graduation were examined

as outcomes. Results indicated that the better than average effect increased over time – the longer students were enrolled in school, the better they felt they compared in ability to their peers. Additionally, those who felt that they were better than average at the start of their academic careers had a better adjustment to university life, higher GPAs, more consistent enrollment, and were more likely to graduate than those who did not initially feel better than average. Thus, the better than average effect resulted in higher performance and a greater likelihood of completing a degree.

Not only do people often believe themselves to be better than average, but they may also describe themselves in terms that conflict with the available evidence. For example, people might exaggerate their abilities in a domain that is particularly important to their self-concepts. One line of research examining this phenomenon focuses on the tendency for students to exaggerate their academic abilities by reporting a higher GPA than the GPA reported in university records. Willard and Gramzow (2009) propose that this form of exaggeration is due to motivation, and in fact, students with a high need for achievement show greater levels of academic exaggeration. Exaggeration was associated with an approach orientation towards academics and a greater perception of challenges versus threats. A crucial question in this area of study is whether such effects are driven by positive emotions or by avoidance of failure. In a study to delineate these motivations, Gramzow et al. (2008) asked students to report their GPA and prior grades before having their physiological responses recorded. During this recording, students were interviewed about their academic performance and their judgments of such performance compared to their peers. Students who exaggerated their GPAs showed a significantly greater increase in GPA at the end of the semester than those who did not exaggerate. Also, they showed a calmer physiological response during the academic performance interview than their non-exaggerating peers. Thus, a self-enhancement bias in reported GPA seems to not only improve academic performance and motivation but does so without increasing anxiety.

Positive illusions in memory, attention, and self-judgments promote goal achievement, motivation, and persistence in achieving these goals. If we consider the need to belong as a fundamental human need (Baumeister & Leary, 1995), we might consider the development and maintenance of interpersonal relationships as another goal that is common to humanity.

Similar to positive illusions about the self, we also hold positive illusions about those close to us and the nature of our relationships, including friends, family members, children, and romantic partners (Cohen & Fowers, 2004; Endo et al., 2000; Murray et al., 1996; Wenger & Fowers, 2008). We not only rate these people as better than average but feel our relationships are stronger than the average relationship. These illusions appear to promote behaviors that initiate and maintain relationships, helping us fulfill this fundamental need.

Our judgments about others have important effects on our interactions and the behaviors of others. Rosenthal and Jacobson (1968) created expectations in teachers that certain students would experience a “spurt” of academic achievement. This expectation created a self-fulfilling prophecy, in that the designated “spurters” showed a greater increase in IQ than the “non-spurters”.

Research on self-fulfilling prophecies indicates that our expectations of others influence our behaviors towards them and, in turn, their behavior. For example, positive illusions about a friend’s romantic interest in us lead to more frequent romantic behaviors, which in turn predicts that friend’s reciprocal interest (Lemay & Wolf, 2016). In this way, overly optimistic judgments of a potential romantic partner’s interest can lead them to become more interested. Positive illusions extend beyond attempts to initiate a relationship into

short- and long-term relationship maintenance. Murray et al. (1996) followed dating couples for one year. Each member of a couple completed a survey describing themselves, their partners, and their level of relationship satisfaction. Overall, the more partners idealized the other, the more satisfied they were with the relationship. Additionally, the more idealized a partner was, the happier they were with the relationship. Idealizing one's partner and being idealized by one's partner were associated with important factors for relationship maintenance – fewer conflicts and less ambivalence about the relationship. Indeed, couples who idealized one another were less likely to break up throughout the yearlong study. This may again illustrate a form of self-fulfilling prophecy created by positive illusions – those who saw the best in their partners, and whose partners saw the best in them, had fewer relationship conflicts and were in turn, more likely to remain together.

Miller et al. (2006) suggested that positive illusions and idealization of one's partner could prevent a decrease in love over time. They followed married couples for 13 years, during which the couples were surveyed on the amount of positive and negative behaviors their partner engaged in, their perceptions of their partner's agreeableness, and their level of marital love. Positive illusions – where a perceived level of agreeableness contradicted reports of behavior – were associated with a greater level of love in newlywed couples at the beginning of the study. These illusions were also associated with a lesser decline in love over time. Taken together, we see that idealized views of our partners – and our partners' idealized view of ourselves – lead to behaviors that create a self-fulfilling prophecy of positive behaviors, increased relationship satisfaction in the short term, and a slower decline of marital love in the long term.

### **Positive illusions: self-protection and self-esteem**

Taylor and Brown (1994) argued that positive illusions promote psychological well-being. In this section, we will review research on positive illusions that help to buffer the self, promote self-esteem and self-worth, and provide protection in times of stress and crisis.

Self-enhancement biases, such as the better than average effect discussed above, serve a protective function (Taylor & Brown, 1988). Brown (2012) found that the better than average effect was stronger after negative feedback. This suggests that the tendency to rate ourselves as better than others helps to restore self-worth, as it appears more strongly after receiving threatening information about the self. These biases, however, appear to be influenced by situational and cultural contexts. After completing either a difficult or easy task with performance feedback, participants whose failure during the difficult task was viewed by an experimenter showed a decreased better than average effect in comparison to participants whose failure was not witnessed by another person (Brown & Gallagher, 1992). In contrast, those who succeeded by completing the easy task rated themselves more highly when the success was seen by others. This suggests that people must balance competing motives when making social comparisons. When one's deficits are apparent to others, it can cause social backlash to portray one's self as better than others, and this reduces the strength of self-enhancement. A cross-cultural study comparing the better than average effect in participants from the United States and Norway, where social norms discourage self-promotion, found that Norwegian participants showed less of a self-enhancement bias than American participants. In addition, self-esteem was less strongly related to self-enhancement in Norwegian participants compared to American participants. This indicates a cultural influence in how people judge themselves in relation to others and in the motivations underlying such judgments (Silvera & Seger, 2004).

In addition to the better than average effect, people show self-serving biases in how they make attributions for outcomes. Broadly speaking, people tend to make internal, self-focused attributions for positive outcomes and external attributions for negative outcomes (Mezulis et al., 2004). Intriguingly, the tendency to make self-serving attributions does not appear to be an automatic process – Sakaki and Murayama (2013) provided evidence that people tend to attribute failures initially to their abilities. When cognitive load is increased through a divided attention task, attributions to abilities are greater than when cognitive load is low. External attributions for poor performance were associated with higher levels of intrinsic motivation, supporting the protective nature of self-serving attributions. This tendency appears to be related to self-esteem: In Brown et al. (2009), participants from the US and China received false feedback indicating that they received high or low scores on a social sensitivity task. Across cultures, those with high self-esteem were more likely to make self-serving attributions about their performance than those with low self-esteem, that is, attributing high performance to their ability and low performance to an issue with test accuracy. In addition, self-serving attributions for failure are stronger in situations of threat, including situations where a task was viewed as important and success was expected (Campbell & Sedikides, 1999).

Positivity biases not only affect our attributions for success and failure, but they can also be found in our expectations of the future. The unrealistic optimism effect refers to a tendency to view positive outcomes for the self as likelier to occur than base rates would suggest and/or to view positive outcomes for the self as more likely to occur in comparison to others (Shepperd et al., 2015). Weinstein (1980) asked college students to rate the likelihood of positive and negative events occurring in their future compared to other students at their university. Overall, students judged positive events as more likely to happen to them compared to others, and negative events as less likely to happen to them compared to others. An unrealistic optimism effect was particularly likely for events that were perceived as controllable. Studies have shown that this effect also occurs in terms of judging negative workplace events, such as experiencing a health and safety issue or being bullied by co-workers (Caponecchia, 2010) or in judging the likelihood of developing strep throat or being a victim of homicide (Weinstein, 1987). We are unrealistically optimistic about a broad range of positive and negative outcomes.

In addition to feeling as though we are luckier than others, we do not appropriately update our predictions in the face of contradictory information about base rates. Research in this area is similar to that of the unrealistic optimism effect – participants are asked to rate the likelihood of different events happening to themselves in comparison to another person. Following these estimates, they are given base-rate information about the events and provide another rating of the likelihood of these events occurring. These updated ratings consistently show that people make greater adjustments to their ratings when the base rates indicate a positive outcome is more likely or a negative event is less likely rather than the opposite – the optimistic update bias. This effect is particularly noticeable for self-related judgments rather than overall judgments of base rates in general (Garrett & Sharot, 2014) or judgments about the likelihood of events occurring to others (Kuzmanovic et al., 2015). The optimistic update bias is influenced by several individual differences. For example, Kuzmanovic et al. (2015) found that this bias was strongest for those high in trait optimism, and manipulations to induce optimism can increase the bias in mildly depressed people (Yoshimura & Hashimoto, 2020). This is particularly interesting given that depressed individuals do not show the optimistic update bias (Korn et al., 2014).

Both the unrealistic optimism effect and the optimistic update bias have drawn concerns that an unrealistic view of risks can lead to poor decision-making or increased risk-taking (Strecher et al., 1995). However, there are benefits to unrealistic optimism – it has been linked with fewer physical symptoms of stress (Scheier & Carver, 1985) as well as better psychological adjustment, lower depression, and greater self-esteem (Lapsley & Hill, 2010). In patients with chronic illness, unrealistic optimism is associated with better physical function via increased positive coping strategies (Fournier et al., 2002).

### **The fading affect bias**

This chapter has argued that positive illusions are often beneficial to people as they try to achieve their life goals, motivate themselves to perform challenging tasks, boost their self-esteem, and nurture a sense of optimism for the future. At the heart of this argument, however, is that when people are inevitably confronted with negative life events, they can cope with those events so that they can move forward in life. But what does this coping actually look like? For the past 25 years, studies of autobiographical memory have revealed that most people are able to resolve the negativity of life events while retaining the positivity of life events.

The negative affect associated with autobiographical memories tends to fade more than the positive affect, a phenomenon known as the Fading Affect Bias or FAB (Walker et al., 1997, 2003a). Several laboratories around the world have replicated the FAB using a variety of populations and methodologies (e.g., Ritchie et al., 2015a; Walker & Skowronski, 2009). While there have been some disruptions of this effect caused by individual differences in depression (Walker et al., 2003b), anxiety (Walker et al., 2014), and narcissism (Ritchie et al., 2015b), the consensus is that the FAB reflects evidence of a healthy coping mechanism operating in autobiographical memory.

The FAB is typically assessed by diary or retrospective procedures. The diary procedure requires that participants keep track of their life events for a period ranging from two weeks to three and a half months. Participants are then tested on the contents of their diaries. Participants rate how positive or negative the events were at the time of the event (initial) and again at the time of the test (current). The change in these affect ratings reflects the perceived change in the emotionality of the events. The retrospective method works similarly, except that participants are asked to recall the events from memory and then they make the ratings of event emotion (both initial and current) in retrospect. Despite the concerns about retrospective biases, the methods produce strikingly similar results (Walker & Skowronski, 2009). Indeed, the results have been replicated using a variety of methods and populations around the world (Ritchie et al., 2015a).

#### **Text box 21.1 Assessing the Fading Affect Bias in a classroom exercise**

##### **Retrospective recall**

Give students ten minutes to recall two specific autobiographical events from their personal past (one positive, one negative). The event descriptions must include information such as time, place, and sensory details and should be between 100–200 words long.

## Affect ratings

Provide students with a seven-point scale ranging from -3 (Extremely Unpleasant) to +3 (Extremely Pleasant) with a midpoint of 0 (Neutral). Ask students to rate the initial affect of the events when they occurred. Next, ask students to rate the current affect of the events. Participants can make these ratings in reverse order to minimize concerns of regression to the mean or use a qualitative approach described below. Remind students that some emotions may not change in emotion, some might become stronger or weaker, and even some emotions make switch from positive to negative or vice versa.

## Calculate difference scores

Subtract the current affect rating from the initial affect rating for each event. For instance, an initial rating of -3 and a current rating of -1 would result in a difference score of -2. Each participant will provide two difference scores (one positive, one negative). These can be analyzed using a repeated measures *t*-test. Evidence of the FAB is observed in larger difference scores for negative events than for positive events. Changes in emotion can also be categorized into four groups: Fixed Affect (No Change), Fading Affect (Reduction Intensity), Flourishing Affect (Increase Intensity), and Flexible Affect (Valence Reversal) and the frequencies can be analyzed using a chi-square test.

## Qualitative method

Instead of relying upon rating scales, students can provide emotion words to describe the initial and current affect of their events. Changes in emotion words would reflect potential changes in emotion (e.g., fear to anxiety, anger to relief).

Research on the FAB has found that this phenomenon is meaningfully related to processes known to be integral to healthy coping, such as social rehearsal. In a series of studies, Skowronski et al. (2004) examined positive and negative events that were shared frequently (ten times or more) or infrequently (five times or less) with other people. The results showed that frequently shared events showed a stronger FAB than events that were infrequently shared. In a final study in which the social rehearsals were directly manipulated in a laboratory setting, the relationship between social rehearsal and affective fading was found to be causal. The more often participants shared events during the study, the stronger the FAB those participants evinced.

Another piece of evidence that shows that the FAB is meaningfully related to the functions of a healthy mind comes from a study that examined the impact of the personality characteristic of grit on how emotions fade. Grit has been described as the ability to persevere in the pursuit of long-term goals. Walker et al. (2020) had 197 participants complete a grit assessment and categorized participants as Low Grit ( $n = 44$ ), Moderate Grit ( $n = 55$ ), Moderate-High Grit ( $n = 53$ ), and High Grit ( $n = 45$ ). Participants then recalled four event memories (two positive, two negative) and made ratings of initial and current affect. The results showed that participants with higher levels of grit had a stronger FAB compared to participants with lower levels of grit.

While much of the research on the FAB has been conducted on commonplace life events, some studies have examined how people cope with traumatic events. Henderson et al. (2015) examined essays written by African-Americans who were coping with violent and nonviolent deaths of close family members or friends. While some events were still emotionally open, a majority of the events had been successfully resolved. The perceived change in the emotional intensity of the events was examined and revealed that the negative emotion showed evidence of substantial fading, a finding consistent with the FAB. Another study involving trauma examined Filipinos' memories of super-typhoon Haiyan, the most destructive typhoon ever to hit the islands (Bond et al., 2015). Three years later, the memory vividness rated by participants closely resembled ratings observed for flashbulb memories. The results also showed that negative emotional intensity for the event faded after the event, a finding consistent with the FAB.

The 2020 pandemic offered a unique opportunity to examine the FAB in the context of ongoing stress and trauma. Would the autobiographical memories of participants show evidence of healthy emotional coping or would they reveal the effects of depression or anxiety? A study began in the summer of 2020 and is continuing beyond the publication of this chapter. Participants were asked to keep one-week diaries of positive and negative events and were later tested on the contents of those diaries after retention intervals. While the specifics need to be properly analyzed and peer-reviewed, a few tentative observations can be shared. First, these data appear to replicate the FAB for pandemic-related events. Second, reports of stress appeared to increase over three semesters. This likely reflects pandemic fatigue. Finally, the events recorded by participants appear more generic than the events captured in prior studies. Indeed, some participants observed that many events seemed very similar. These observations are consistent with prior research on the FAB – people are often resilient in the face of adversity, but resilience is not synonymous with invulnerability.

### **Text box 21.2 How positive emotions become associated with irrationality – a Candide explanation**

This chapter argues that positive emotions can help people think more clearly, make good decisions, and overcome obstacles they may not otherwise be able to surmount. This assertion is at the core of positive psychology. But many scholars consider positive emotions inherently misleading. Indeed, feelings of happiness and positivity are oftentimes associated with foolishness and stupidity. Why?

The answer has to do more with philosophy than with psychology and the philosopher who most passionately made this linkage was François-Marie Arouet, better known by his pen name Voltaire (Brooks, 1964; Cronk, 2009). Voltaire was the most important French philosopher of the Enlightenment. He developed an early cynicism that was directed at authority figures, traditional morality, and the church. At 16, he had an affair with a 27-year-old married mother of three. Four years later, his father sent him to The Hague as punishment after it was revealed that he was publishing embarrassing poems that lampooned the church and figures of nobility. Ten years later, he was banished to England after getting into a heated argument with the Chevalier de Rohan. While in England, he read the works of Isaac Newton, which convinced him to embrace empiricism. Upon his return to

France, Voltaire published *Lettres Philosophiques* that distilled his newfound ideals in the form of scathing critiques of the church. His work was openly burned in the streets of Paris.

In 1759, after a lifetime in which he had faced imprisonment, banishment, harassment from the church, and the death of a woman he loved, he wrote *Candide*. This work was a scathing satire that ridiculed Gottfried Leibniz, a German philosopher who posited that humans exist in the best of all possible universes. The satire follows the travails of Candide, an optimistic young man who grows up in an idyllic setting, being taught philosophy and seeking the affections of the beautiful Cunégonde. His advances are met with punishment, including banishment, conscription into an army, physical abuse, and near execution. He learns that his home was destroyed and his family murdered. Cunégonde had been captured, raped, and sold into slavery. Through the next 20 chapters, Candide is repeatedly forced to suffer and defend his optimism. He visits the fabled El Dorado, a city that is covered in gold. He grows weary of the leisure and leaves, concluding that man was not meant for such idleness. In the waning chapters of the text, he talks with a character named Martin, a pessimist who professes that disease, despair, and death are the only fortunes of life. Candide finally reunites with Cunégonde, who is no longer young or beautiful. He buys her from slavery and, with a collection of characters he has met, buys a farm and resolves to live his life simply. Instead of pursuing optimistic dreams that have led him to misery, he suggests instead that humans should seek a life of unassuming work, concluding that “we must cultivate our garden”.

That a figure such as Voltaire embraced science and empiricism and so strongly linked the positive emotions of hope, happiness, and serenity with irrationality, stupidity, and folly meant that the die was cast in minds of many empiricists of the Enlightenment. Science was about cold, hard, and ultimately *negative* facts. Voltaire viewed positive emotions as Chimera in the mind. A Chimera is a fanciful creature of Greek mythology with the head of a lion, the body of a goat, and the tail of a serpent (Figure 21.1). In some tales, it could fly or breathe fire. In the parlance of Voltaire, a Chimera was something that was wished for but could never be achieved. Hence, positive emotions could only mislead the mind – never help or inform it.



Figure 21.1 Chimera by Pearson Scott Foresman, 2016, donated to the Wikimedia Foundation. Retrieved from [https://commons.wikimedia.org/wiki/File:Chimera\\_\(PSF\).jpg](https://commons.wikimedia.org/wiki/File:Chimera_(PSF).jpg)

## Conclusion

As this chapter is being completed, vaccines are being administered to the world's population. Economies are starting to reopen and a sense of normalcy is starting to return. In some cases, the optimism is premature and many regions are seeing their caseloads spike. In other cases, lockdown protocols have stoked anxieties that will likely be felt for years to come. But there is an undeniable feeling of guarded optimism that is starting to wash over the world. People are starting to look forward to the future. What does this tell us about the positive biases described in this chapter? It tells us that these biases more often defend the mind than cripple it. The psychological pain of depression, anxiety, and anger distort reality by turning vipers into dragons. Positivity creates a platform that allows people to formulate and achieve long-term goals, striving to succeed rather than retreating from failure. Positive biases boost esteem and engender a set of defenses that serve to protect the self from harm. And finally, when tragedy strikes, the pain is often short-lived and the healthy mind can recover quickly and effectively.

Voltaire would consider the positive biases discussed in this chapter to be Chimera, fanciful and fearsome monsters leading the mind away from reality. Instead, we would suggest that a more suitable mythological allusion befitting positivity would be the Pegasus – a winged creature allowing the individual to rise above the troubles of the world, to gain perspective, and to overcome the insurmountable.

## Summary

- Positive illusions are present in attention, memory, and judgment.
- These illusions motivate and enhance goal achievement.
- Positive illusions boost self-esteem and assist in recovering from negative events.

## Further reading

Taylor and Brown (1988) pioneered the idea that positive illusions can be beneficial for well-being. The following readings are suggested for delving deeper into specific illusions: For a meta-analytic review of the age-related positivity effect, see Reed et al. (2014) which notes that experimental procedures influence the strength of the age-related positivity effect. Brown (2012) presents a series of studies demonstrating the motivational, self-enhancing nature of the “better than average” effect. Shepperd et al.’s (2015) comprehensive review of unrealistic optimism describes the different forms unrealistic optimism can take, why it occurs, and the positive and negative outcomes of holding unrealistically optimistic beliefs. Walker and Skowronski (2009) provide a review of the literature on the fading affect bias, describing its functions in autobiographical memory.

## References

- Banks, D. M., Scott, B. G., & Weems, C. F. (2018). Anxiety, hostile attributions, and differences in heart rate response to ambiguous situational vignettes in adolescents. *Emotion*, 18(2), 248–259.
- Barefoot, J. C., Dodge, K. A., Peterson, B. L., Dahlstrom, W. G., & Williams Jr, R. B. (1989). The Cook-Medley hostility scale: Item content and ability to predict survival. *Psychosomatic Medicine*, 51(1), 46–57.
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3), 497–529.

- Beck, A. T. (1976). *Cognitive therapy and emotional disorders*. New York: International Universities Press.
- Bergquist, M. (2020). Most people think they are more pro-environmental than others: A demonstration of the better-than-average effect in perceived pro-environmental behavioral engagement. *Basic and Applied Social Psychology*, 42(1), 50–61.
- Bjärehed, J., Sarkohi, A., & Anderson, G. (2010). Less positive or more negative? Future directed thinking in mild to moderate depression. *Cognitive Behaviour Therapy*, 39(1), 37–45.
- Bond, G. D., Walker, W. R., Bargo, A. J. B., Bansag, M. J., Self, E. A., Henderson, D. X., Anu, R. M., Sum, L. S., & Alderson, C. J. (2015). Fading affect bias in the Philippines: Confirmation of the FAB in positive and negative memories but not for death memories. *Applied Cognitive Psychology*, 30(1), 51–60.
- Brooks, R. A. (1964). *Voltaire and Leibniz*. Geneva: Droz.
- Brown, J. D. (2012). Understanding the better than average effect. *Personality and Social Psychology Bulletin*, 38(2), 209–219.
- Brown, J. D., Cai, H., Oakes, M. A., & Deng, C. (2009). Cultural similarities in self-esteem functioning: East is East and West is West, but sometimes the twain do meet. *Journal of Cross-Cultural Psychology*, 40(1), 140–157.
- Brown, J. D., & Gallagher, F. M. (1992). Coming to terms with failure: Private self-enhancement and public self-effacement. *Journal of Experimental Social Psychology*, 28(1), 3–22.
- Campbell, W. K., & Sedikides, C. (1999). Self-threat magnifies the self-serving bias: A meta-analytic integration. *Review of General Psychology*, 3(1), 23–43.
- Caponecchia, C. (2010). It won't happen to me: An investigation of optimism bias in occupational health and safety. *Journal of Applied Social Psychology*, 40(3), 601–617.
- Carstensen, L. L., Isaacowitz, D. M., & Charles, S. T. (1999). Taking time seriously: A theory of socioemotional selectivity. *American Psychologist*, 54(3), 165–181.
- Carstensen, L. L., & Mikels, J. A. (2005). At the intersection of emotion and cognition: Aging and the positivity effect. *Current Directions in Psychological Science*, 14(3), 117–121.
- Chung, J., Schriber, R. A., & Robins, R. W. (2016). Positive illusions in the academic context: A longitudinal study of academic self-enhancement in college. *Personality and Social Psychology Bulletin*, 42(10), 1384–1401.
- Cohen, J. D., & Fowers, B. J. (2004). Blood, sweat, and tears: Biological ties and self-investment as sources of positive illusions about children and stepchildren. *Journal of Divorce and Remarriage*, 42(1–2), 39–59.
- Crick, N. R., Grotjeter, J. K., & Bigbee, M. A. (2002). Relationally and physically aggressive children's intent attributions and feelings of distress for relational and instrumental peer provocations. *Child Development*, 73(4), 1134–1142.
- Cronk, N. (2009). *The Cambridge companion to Voltaire*. Cambridge: Cambridge University Press.
- Dalgleish, T., Williams, J. M. G., Golden, A.-M. J., Perkins, N., Barrett, L. F., Barnard, P. J., Yeung, C. A., Murphy, V., Elward, K. T., & Watkins, E. (2007). Reduced specificity of autobiographical memory and depression: The role of executive control. *Journal of Experimental Psychology: General*, 136(1), 23–42.
- Endo, Y., Heine, S. J., & Lehman, D. R. (2000). Culture and positive illusions in close relationships: How my relationships are better than yours. *Personality and Social Psychology Bulletin*, 26(12), 1571–1586.
- Fournier, M., Ridder, D., & Bensing, J. (2002). Optimism and adaptation to chronic disease: The role of optimism in relation to self-care options of type 1 diabetes mellitus, rheumatoid arthritis and multiple sclerosis. *British Journal of Health Psychology*, 7(4), 409–432.
- Fung, H. H., Gong, X., Ngo, N., & Isaacowitz, D. M. (2019). Cultural differences in the age-related positivity effect: Distinguishing between preference and effectiveness. *Emotion*, 19(8), 1414–1424.
- Fung, H. H., Isaacowitz, D. M., Lu, A. Y., Wadlinger, H. A., Goren, D., & Wilson, H. R. (2008). Age-related positivity enhancement is not universal: Older Chinese look away from positive stimuli. *Psychology and Aging*, 23(2), 440–446.

- Gallo, D. A., Korthauer, L. E., McDonough, I. M., Teshale, S., & Elizabeth, L. (2013). Age-related positivity effects and autobiographical memory detail: Evidence from a past/future source memory task. *Memory*, 19(6), 641–652.
- Garrett, N., & Sharot, T. (2014). How robust is the optimistic update bias for estimating self-risk and population base rates? *PLoS ONE*, 9(6), e98848.
- Goodwin, H., Yiend, J., & Hirsch, C.R. (2017). Generalized anxiety disorder, worry and attention to threat: A systematic review. *Clinical Psychology Review*, 54, 107–122.
- Gothlib, I. H., & Joormann, J. (2010). Cognition and depression: Current status and future directions. *Annual Review of Clinical Psychology*, 6, 285–312.
- Gramzow, R. H., Willard, G., & Mendes, W. B. (2008). Big tales and cool heads: Academic exaggeration is related to cardiac vagal reactivity. *Emotion*, 8(1), 138–144.
- Hards, E., Ellis, J., Fisk, J., & Reynolds, S. (2020). Negative view of the self and symptoms of depression in adolescents. *Journal of Affective Disorders*, 262, 143–148.
- Henderson, D. X., Bond, G. D., Alderson, C. J., & Walker, W. R. (2015). This too shall pass: Evidence of coping and fading emotion in African Americans' memories of violent and nonviolent death. *Omega: Journal of Death and Dying*, 71(4), 291–311.
- Holtzworth-Munroe, A., & Hutchinson, G. (1993). Attributing negative intent to wife behavior: The attributions of maritally violent versus nonviolent men. *Journal of Abnormal Psychology*, 102(2), 206–211.
- Hunt, C., Keogh, E., & French, C. C. (2006). Anxiety sensitivity: The role of conscious awareness and selective attentional bias to physical threat. *Emotion*, 6, 418–428.
- Joormann, J., & Stanton, C. H. (2016). Examining emotion regulation in depression: A review and future directions. *Behaviour Research and Therapy*, 86, 35–49.
- Kennedy, Q., Mather, M., & Carstensen, L. L. (2004). The role of motivation in the age-related positivity effect in autobiographical memory. *Psychological Science*, 15(3), 208–214.
- Korn, C. W., Sharot, T., Walter, H., Heekeren, H. R., & Dolan, R. J. (2014). Depression is related to an absence of optimistically biased updating about future life events. *Psychological Medicine*, 44(3), 579–592.
- 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(2), 221–232.
- Krupic, D., Zuro, B., & Corr, P.J. (2021). Anxiety and threat magnification in subjective and physiological responses of fear of heights induced by virtual reality. *Personality and Individual Differences*, 169(1), 109720.
- Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2015). Self-specific optimism bias in belief updating is associated with high trait optimism. *Journal of Behavioral Decision Making*, 28(3), 281–293.
- Lapsley, D. K., & Hill, P. L. (2010). Subjective invulnerability, optimism bias and adjustment in emerging adulthood. *Journal of Youth and Adolescence*, 39(8), 847–857.
- Lemay, E. P., & Wolf, N. R. (2016). Projection of romantic and sexual desire in opposite-sex friendships: How wishful thinking creates a self-fulfilling prophecy. *Personality and Social Psychology Bulletin*, 42(7), 864–878.
- Li, T., Fung, H. H., & Isaacowitz, D. M. (2011). The role of dispositional reappraisal in the age-related positivity effect. *Journals of Gerontology – Series B Psychological Sciences and Social Sciences*, 66 B(1), 56–60.
- Mezulis, A. H., Abramson, L.Y., Hyde, J. S., & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin*, 130(5), 711–747.
- Miller, P. J. E., Niehuis, S., & Huston, T. L. (2006). Positive illusions in marital relationships: A 13-year longitudinal study. *Personality and Social Psychology Bulletin*, 32(12), 1579–1594.
- Murray, S. L., Holmes, J. G., & Griffin, D. W. (1996). The self-fulfilling nature of positive illusions in romantic relationships: Love is not blind, but prescient. *Journal of Personality and Social Psychology*, 71(6), 1155–1180.

- Nasby, W., Hayden, B., & DePaulo, B. M. (1980). Attributional bias among aggressive boys to interpret unambiguous social stimuli as displays of hostility. *Journal of Abnormal Psychology, 89*(3), 459–468.
- Nees, M. A. (2019). Safer than the average human driver (who is less safe than me)? Examining a popular safety benchmark for self-driving cars. *Journal of Safety Research, 69*, 61–68.
- Pronin, E., Lin, D.Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*(3), 369–381.
- 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, 29*(1), 1–15.
- Ritchie, T. D., Batteson, T. J., Bohn, A., Crawford, M. T., Ferguson, G. V., Schrauf, R. W., Vogl, R. J., & W. Walker, W. R. (2015). A pancultural perspective on the fading affect bias in autobiographical memory. *Memory, 23*(2), 278–290.
- Ritchie, T. D., Walker, W. R., Marsh, S., Hart, C., & Skowronski, J. J. (2015). Narcissism distorts the fading affect bias in autobiographical memory. *Applied Cognitive Psychology, 29*(1), 104–114.
- Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review, 3*(1), 16–20.
- Sakaki, M., & Murayama, K. (2013). Automatic ability attribution after failure: A dual process view of achievement attribution. *PLoS ONE, 8*(5), e63066.
- Scheier, M. F., & Carver, C. S. (1985). Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. *Health Psychology, 4*(3), 219–247.
- Schryer, E., & Ross, M. (2014). Does the age-related positivity effect in autobiographical recall reflect differences in appraisal or memory? *Journals of Gerontology – Series B Psychological Sciences and Social Sciences, 69*(4), 548–556.
- Shepperd, J. A., Waters, E. A., Weinstein, N. D., & Klein, W. M. P. (2015). A primer on unrealistic optimism. *Current Directions in Psychological Science, 24*(3), 232–237.
- Silvera, D. H., & Seger, C. R. (2004). Feeling good about ourselves. *Journal of Cross-Cultural Psychology, 35*(5), 571–585.
- Skowronski, J. J., Gibbons, J. A., Vogl, R. J., & Walker, W. R. (2004). The effects of social disclosure on the affective intensity provoked by autobiographical memories. *Self and Identity, 3*, 285–309.
- Strecher,V.J., Kreuter, M.W., & Kobrin, S.C. (1995). Do cigarette smokers have unrealistic perceptions of their heart attack, cancer, and stroke risks? *Journal of Behavioral Medicine, 18*(1), 45–54.
- Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*(2), 193–210.
- Taylor, S. E., & Brown, J. D. (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin, 116*(1), 21–27.
- Urban, E. J., Charles, S. T., Levine, L. J., & Almeida, D. M. (2018). Depression history and memory bias for specific daily emotions. *PLoS ONE, 13*(9), 1–15.
- Walker, W. R., Alexander, H., & Aune, K. (2020). Higher levels of grit are associated with a stronger fading affect bias. *Psychological Reports, 123*(1), 124–140.
- Walker, W. R., & Skowronski, J. J. (2009). The fading affect bias: But what the hell is it for? *Applied Cognitive Psychology, 23*(8), 1122–1136.
- Walker, W. R., Skowronski, J. J., Gibbons, J. A., Vogl, R. J., & Thompson, C. P. (2003a). On the emotions that accompany autobiographical memories: Dysphoria disrupts the fading affect bias. *Cognition and Emotion, 17*(5), 703–723.
- Walker, W. R., Skowronski, J. J., & Thompson, C. P. (2003b). 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.
- Walker, W. R., Yancu, C. N., & Skowronski, J. J. (2014). Trait anxiety reduces affective fading for both positive and negative autobiographical memories. *Advances in Cognitive Psychology, 10*(3), 81–89.
- Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39*(5), 806–820.

- Weinstein, N. D. (1987). Unrealistic optimism about susceptibility to health problems: Conclusions from a community-wide sample. *Journal of Behavioral Medicine*, 10(5), 481–500.
- Wenger, A., & Fowers, B. J. (2008). Positive illusions in parenting: Every child is above average. *Journal of Applied Social Psychology*, 38(3), 611–634.
- Willard, G., & Gramzow, R. H. (2009). Beyond oversights, lies, and pies in the sky: Exaggeration as goal projection. *Personality and Social Psychology Bulletin*, 35(4), 477–492.
- Williams, J. M. G., & Scott, J. (1988). Autobiographical memory in depression. *Psychological Medicine*, 18(3), 689–695.
- Wojcik, S. P., & Ditto, P. H. (2014). Motivated happiness: Self-enhancement inflates self-reported subjective well-being. *Social Psychological and Personality Science*, 5(7), 825–834.
- Yoshimura, S., & Hashimoto, Y. (2020). The effect of induced optimism on the optimistic update bias. *BMC Psychology*, 8(1), 1–7.
- Ziano, I., Mok, P. Y., & Feldman, G. (2021). Replication and extension of Alicke (1985) better-than-average effect for desirable and controllable traits. *Social Psychological and Personality Science*, 12(6), 1005–1017.

# **Part III**

# **Memory**



Taylor & Francis  
Taylor & Francis Group  
<http://taylorandfrancis.com>

## 22 Moses illusion

*Felix Speckmann and Christian Unkelbach*

Next time you are at a cocktail party or similar social event, try asking your conversation partners the following question: “How many animals of each kind did Moses take on the Ark?”. You will most likely receive a skeptical glance and the hesitant answer: “Two”. Then, you can proceed to tell them that it was in fact not Moses, but Noah who took the animals on the Ark in the biblical story. Erickson and Mattson (1981) were the first to show that people fall for this kind of trick question (i.e., answer the question as if it were formulated correctly) and named this effect the Moses illusion. Since their seminal work, research has addressed when the illusion occurs, what the underlying processes are, and what moderates the strength of the illusion. Because apart from being an amusing demonstration of a cognitive fallacy and a potential icebreaker at a party, understanding the underpinnings of this illusion and why people fall for it illuminates the way people understand and comprehend language. Similar to how optical illusions help understanding how the visual system works, such cognitive illusions provide insights on how the cognitive system works. In this chapter, we explain the basic Moses illusion-paradigm by reviewing the original study by Erickson and Mattson (1981). We then examine several explanations of the illusion and how compatible they are with the available evidence. We continue by addressing several moderators of the illusion: What increases illusion strength, what decreases it, and what has no effect on it. Finally, we will close with implications for language comprehension and daily life.

### The Moses illusion-paradigm

In their study, Erickson and Mattson (1981) were interested in language comprehension; specifically, why sometimes the addressee of a question misunderstands a question even though they possessed all relevant knowledge to understand the question. To investigate this phenomenon, they asked participants to read aloud questions presented on a screen and then answer them aloud as well, which was recorded using a microphone. They also informed participants that they will see several questions that “have something wrong with [them]” (p. 542), giving them an example of such a question and then telling them to reply “wrong” or similarly when faced with such a question. Examples for such “distorted questions” are “How many animals of each kind did Moses take on the Ark?” or “In the biblical story, what was Joshua swallowed by?”. The “undistorted” versions of these questions would be “How many animals of each kind did Noah take on the Ark?” and “In the biblical story, what was Jonah swallowed by?”. Participants then proceeded to look at, read aloud, and respond aloud to 20 questions, which consisted of four target questions (“distorted questions”) and 16 distractor questions. Afterwards, in the second

part of the experiment, participants saw only the beginning of all questions shown before and had to complete each question by adding the remaining words. Finally, participants answered four questions designed to check whether they had the necessary knowledge to answer the distorted questions correctly. For example, if a participant could not answer the question “Who was it that took the animals on the Ark?” correctly, they did not possess the required knowledge to correctly respond to the distorted question and their response did not count towards the percentage of times the illusion occurred.

For the original Moses question, 26 out of 27 participants had the required knowledge (i.e., Noah took the animals on the Ark), but nonetheless 21 of those 26 fell for the illusion and answered “two”; a correct response if the question were undistorted, but incorrect because the question is faulty. We call this type of response a “Moses response” in the following. For the other three distorted questions, the required knowledge was present less frequently (range 9–16 out of 27) and they only led to Moses responses about 43% of the time. These results provided the first evidence of the Moses illusion. Since then, this robust illusion has been replicated many times, but what are the processes underlying this illusion?

## **Explaining the Moses illusion**

Presently, four explanations are discussed in the literature to account for the Moses illusion. We first explain each one and then present research that supports or contradicts the respective explanations.

### ***Cooperative communication setting***

The first explanation for the Moses illusion is the assumption of a cooperative communication setting (Grice, 1975). According to Grice, cooperative communication should follow four maxims (i.e., quantity, quality, relation, and manner). In a nutshell, these maxims imply that cooperative communication should not include trick questions. Thus, if you are a participant in the experiment described above and you read one of the distorted questions and notice the distortion, given you did not read the instructions with great attention and missed the part about the “wrong” questions, you might assume that this distortion is actually an error on the researcher’s part. You would then say “two”, assuming the experimenter mixed up “Moses” and “Noah”; and accordingly, the sensible action is to answer as if the question had been presented undistorted – providing a correct answer to a faulty question.

Within a cooperative communication setting (Grice, 1975), “two”, the Moses response, would be the correct response. Participants respond as if the questions were undistorted because they ignore the mistakes that the researchers ostensibly made. While the explanation is plausible, several studies found evidence against the Moses illusion as a consequence of cooperative communication. First, Reder and Kusbit (1991) explicitly tested this explanation by assigning participants to two different conditions. In a “gist” condition, participants were explicitly told to ignore possible distortions and answer questions as if they were correct; that is, the condition made the maxims of Grice explicit. In the “literal” condition, however, participants were told to say “can’t say” when the question was distorted, replicating the original experiment by Erickson and Mattson (1981). The cooperative communication setting explanation would predict that participants in both conditions noticed the distortion but react to it differently as per the instructions; that

is, participants in the gist condition should provide more Moses responses. However, participants who were told to ignore the distortion committed *less* errors; that is, when responding to distorted questions, participants in the gist condition committed significantly less errors (24.63%) than participants in the literal condition (38.31%).

Further evidence against the cooperative communication explanation came from Bredart and Modolo (1988) and Van Oostendorp and De Mul (1990): In both experiments participants responded to statements rather than questions (e.g., “Moses took two animals of each kind on the ark”) by indicating whether they were true or false. In this case, a cooperative communication setting would not predict that participants ignore distortions because the veracity of the statements is now most relevant, rather than the response to a (possibly distorted) question. In other words, if the question is about the number of animals, participants might be inclined to ignore distortions because that is not the main focus of the question. If they are asked to judge the correctness of a statement however, ignoring distortions would be counterproductive. Nevertheless, in both studies, participants showed the illusion.

Finally, Speckmann and Unkelbach (2021) used a multiple-choice response format to test if the illusion would persist. This response format included four responses options: the correct response or Moses response (i.e., “Noah” for undistorted questions and “Moses” for distorted questions), an obviously wrong foil (“three”), “can’t say” (the correct response to a distorted question), and “I don’t know” as an option to skip the question. Because “can’t say” was a response option for every question, it repeatedly reminded participants that some questions could not be answered. The cooperative communication setting explanation would predict that the percentage of Moses responses would be reduced in such a response format, but the results were comparable to those of earlier research. Taken together, these findings make a cooperative communication explanation of the Moses illusion unlikely.

### **Text box 22.1 Classroom demonstration of the Moses illusion**

This text box describes an adaptation of the multiple-choice format of the Moses illusion paradigm used by Speckmann and Unkelbach (2021).

#### **Method**

#### **Materials**

A selection of questions in both distorted and undistorted forms is given in Table 22.1. One may use either this selection of statements, or a larger subset or the whole set of questions from the appendix in Speckmann and Unkelbach (2021). The statements presented to the participants should be half distorted and half undistorted questions (within manipulation of distortion) and no question should be presented in both versions. This can be achieved through online tools (e.g., Qualtrics) or by passing out paper-pencil questionnaires with an *a priori* randomized set of questions. Each question has four different response options. The first response option is the correct response if the question is undistorted and the Moses response if the question is distorted (e.g., “two” to the Moses question). The second option is a foil that also depends on the question and is always (and rather obviously) incorrect. The third

Table 22.1 Examples of distorted and undistorted questions, with respective correct (Moses) response and foils (only relevant in a multiple-choice design)

<i>Distorted</i>	<i>Undistorted</i>	<i>Answer</i>	<i>Foil</i>
By flying a kite, what did <i>Edison</i> discover?	By flying a kite, what did <i>Franklin</i> discover?	Electricity	Gravity
Who found the glass slipper left at the ball by <i>Snow White</i> ?	Who found the glass slipper left at the ball by <i>Cinderella</i> ?	Prince	Stepmother
What is the name of the Mexican dip made with mashed-up <i>artichokes</i> ?	What is the name of the Mexican dip made with mashed-up <i>avocados</i> ?	Guacamole	Salsa
What country was Margaret Thatcher <i>president</i> of?	What country was Margaret Thatcher <i>prime minister</i> of?	United Kingdom	France
What is the name of the kimono-clad courtesans who entertain <i>Chinese</i> men?	What is the name of the kimono-clad courtesans who entertain <i>Japanese</i> men?	Geisha	Samurai
Who is the video game character and Italian plumber who is <i>Sony</i> 's mascot?	Who is the video game character and Italian plumber who is <i>Nintendo</i> 's mascot?	Mario	Sonic
Who is the dictator of <i>South Korea</i> ?	Who is the dictator of <i>North Korea</i> ?	Kim Jong-Un	Fidel Castro
What is the name of Leonardo da Vinci's famous painting of a woman that is displayed in the <i>Pompidou</i> in Paris?	What is the name of Leonardo da Vinci's famous painting of a woman that is displayed in the <i>Louvre</i> in Paris?	Mona Lisa	The Scream
How many doors does an Advent <i>wreath</i> have?	How many doors does an Advent <i>calendar</i> have?	24	365
What is the name of the device that tells the <i>temperature</i> by measuring the incidence of sunlight on a dial?	What is the name of the device that tells the <i>time</i> by measuring the incidence of sunlight on a dial?	Sundial	Oscillator

Note: The first five statements are from Reder and Kusbit (1991), the second five (and all foils) are from Speckmann and Unkelbach (2021).

option is always “This question can't be answered in this form” (rather than the ambiguous “can't say”) and the fourth option is always “Don't know”. The last option allows participants to skip a question if they feel they do not possess the relevant knowledge to answer. The order of response options may be randomized for each question (but it likely will not have a large effect on the results, cf. Speckmann & Unkelbach, 2021).

### **Participants and design**

The average observed effect size in Speckmann and Unkelbach (2021) was  $d = .57$ , suggesting that a sample of  $N = 35$  would be sufficient to show an effect with 90% power and  $\alpha = .05$ , although using less than the full set of 40 questions will likely reduce the effect size and increase the needed sample size. Question type (distorted v. undistorted) is manipulated within participants.

### **Procedure**

Participants receive a questionnaire containing half distorted and half undistorted questions. They are informed that some questions will appear that cannot be answered and that the correct response in this case is “This question can’t be answered in this form”. They are also informed that they can select the response option “Don’t know” to skip a question that they do not know the answer to.

### **Results**

Moses responses (i.e., first response option to only the distorted questions) are coded as 1 and all other responses are coded as 0. All values are then added up and divided by the total number of distorted questions to compute the percentage of Moses responses for each participant. To provide an inferential test, this number is compared to the percentage of Moses responses one would expect from chance level. If participants respond randomly, they should select each of the three response options (the fourth one being a skip option rather than a response) one-third of the time. Thus, a one-sample  $t$ -test comparing the percentage of Moses responses to the chance level of a third should show that participants give more Moses responses than could be expected by chance.

### **Imperfect encoding**

The imperfect encoding explanation locates the cause within participants’ perception of the question. For example, when presented with the question “How many animals of each kind did Moses take on the ark?”, participants may spontaneously encode the question in its undistorted form (i.e., with Noah). In other words, participants may “auto-correct” questions while reading or hearing them. This explanation suggests the illusion is due to participants’ correct responding to a question they incorrectly encoded. The relevant distinction to the cooperative communication explanation above resides in the fact that participants are assumed to correctly understand the Moses part of the question, while for the imperfect encoding explanation, Moses is never correctly encoded as part of the question.

Consequently, a direct test of this explanation would be ascertaining that participants encode the questions correctly and checking if this reduces the illusion. This is what Erickson and Mattson (1981) did in their original study. In their first experiment, participants saw each question presented on a monitor and had to first read aloud the whole question before responding to the question as quickly as possible. Reading the question aloud should ensure correct encoding, but the illusion occurred nonetheless. Further evidence comes from Reder and Kusbit (1991) and Van Oostendorp and De Mul (1990) who measured

how much time participants spend reading a question. Longer reading times should be related to longer encoding times, so imperfectly encoding (i.e., without noticing the distortion) a distorted question should take less time to read than correctly encoding (i.e., noticing the distortion) a distorted question (Park & Reder, 2004). However, the results showed that participants spent more time reading a question when they did not notice the distortion, providing further evidence against the imperfect encoding explanation.

### ***Imperfect retrieval***

Assuming that participants encode the distorted question correctly, they still need to retrieve relevant information from memory to match the encoded question against. For example, when answering the original Moses question, participants could retrieve only the number of animals, rather than all associated information. This explanation follows because participants are entitled to retrieve only the relevant information (i.e., how many animals of each kind are on the ark), but not the irrelevant information (i.e., who brought the animals unto the ark). Thus, participants would not notice the distortion because the encoded question does not clash with the information retrieved from memory.

If the imperfect retrieval explanation is true, then it should be possible to manipulate memory retrieval to affect illusion strength. For example, having participants study the statements before responding to them should improve accessibility in memory and lead to less Moses responses. However, Reder and Kusbit (1991) did exactly that and found no supportive evidence for the imperfect retrieval explanation. In their second experiment, following the experiment described above, participants first studied correct statements directly related to the later questions (e.g., “Noah took two animals of each kind on the ark”). One-fourth of the statements pertained to later distorted questions and one-fourth pertained to later undistorted questions, resulting in priming for half of the total questions. Participants then responded to all questions (i.e., half distorted and half undistorted), again under instructions of “gist” (i.e., participants should ignore errors) and “literal” (i.e., replicating the original Erikson & Mattson, 1981, conditions). The results showed that while participants responded much quicker and more accurately to questions which they had previously studied, the overall pattern from Experiment 1 remained stable: Responses in the literal condition were slower and provided more, not less, Moses responses compared to the gist condition.

The imperfect retrieval explanation would suggest that, for distorted questions, the accuracy difference between gist and literal conditions should diminish for studied statements because the distortion should become more obvious in comparison to the studied statement. This should lead to a decrease in accuracy for the gist condition as distortions become harder to ignore and to an increase in the literal condition as distortions become easier to notice. However, this change is not reflected in the data, thus providing evidence against the imperfect retrieval explanation.

### ***Imperfect matching (partial matching)***

This explanation was suggested by Reder and Kusbit (1991), who argued that, despite correctly encoding the question and correctly retrieving relevant memories, the matching process between encoding and retrieval could be flawed. Partial matching implies that participants do not try to match each word of a sentence and only accept a 100% match, but rather that they accept smaller mismatches and treat the match as valid if the

question is close enough to the memory structure. This leads to the testable hypothesis that increasing the mismatch between encoded question and retrieved information from memory should lead to fewer Moses responses. That is, if the encoded distorted term deviates too much from the retrieved undistorted term, participants should notice the distortion. Such a condition was, for example, established by Erickson and Mattson (1981), when they replaced Noah with Nixon (Exp. 3).

Kamas et al. (1996) argued that the semantic relatedness between the undistorted and distorted terms plays a key role in determining which distortions participants detect. They proposed a semantic network in which activation spreads between different semantic concepts, with higher activation spread between more closely related concepts (Kamas & Reder, 1995). Applied to the Moses illusion, this model predicts distortions are less likely to be detected the stronger the connection between distorted and undistorted terms are. For example, when examining the original Moses question, Noah and Moses are highly connected because they both appear in the Old Testament of the Bible and their names are both Hebrew, they are often depicted as old men, their stories both deal with water, and so forth. All of these shared features can be seen as semantic nodes that connect to both Moses and Noah, making their overall connectedness strong and the distortion highly unlikely to be noticed. Nixon, on the other hand, has little to nothing in common with Noah, resulting in less shared features, less connectedness, and thus, a higher likelihood of the distortion being detected.

This model explains why manipulations based on the word level did not lead to increased detection rates of distortions. And while others have suggested that not only semantic relatedness, but also phonetic similarity can decrease distortion detection (Shafto & MacKay, 2000), the partial matching explanation is the explanation for the Moses illusion that is compatible with most of the available data.

## Moderators

After considering the potential explanations, we will now take a closer look at different moderators of the illusion. Which factors increase illusion strength (i.e., reduce distortion detection and increases the number of Moses responses) and which factors decrease it?

### *Semantic relatedness*

A lot of early research examined which features contained in the questions would influence the illusion. For instance, two studies by van Oostendorp and De Mul (1990) and van Oostendorp and Kok (1990) found that the illusion becomes stronger the more semantically related the distorted term is to the undistorted term, which follows directly from the partial matching hypothesis. For example, a name that is somewhat semantically related to Moses and still results in the Moses illusion albeit much less frequently is Adam. From a partial matching perspective, this reduced semantic relatedness makes sense because Adam has many unique semantic nodes that are not shared by Moses (e.g., Garden Eden, Eve, the apple, etc.) and thus an increased detection of distortions is the logical outcome.

### *Statements instead of questions*

Bredart and Modolo (1988) found that, when using statements instead of questions, changing the focus of the sentence can lead to a decrease in Moses responses. In two conditions,

the sentences either focused on the (un)distorted term (placed in the left cleft phrase) or on a different part of the sentence (placed in the right cleft phrase). For example, participants in the first condition would read “It was Moses who took two animals of each kind on the ark”, whereas participants in the second condition would read “It was two animals of each kind that Moses took on the ark”. The focus on the distorted term in the left cleft phrase resulted in fewer Moses responses, but a conceptual replication by Kamas et al. (1996, Experiment 1) suggests that this is due to a shift in response bias rather than improvement of distortion detection. Rather than using differently structured sentences, they presented participants with statements prior to the questions. In these statements, they printed in bold either the (un)distorted term (i.e., “MOSES”), the part of the statement that would be the correct response in the later questions (i.e., “TWO”), or nothing at all. They found that capitalizing the (un)distorted term increased “can’t say” responses for both the undistorted and distorted questions. This increase in both correct detections and false alarms suggests that the manipulation did not increase participants’ sensitivity to distortions, but rather their response bias to respond “can’t say” to any question.

### ***Situational manipulations***

Other research has investigated ways to decrease the Moses illusion. One study by Song and Schwarz (2008) found that low processing fluency (see Unkelbach & Greifeneder, 2013) when reading the questions attenuated the Moses illusion. In both experiments, the authors manipulated the question font to be either easy to read or hard to read and found that participants in the hard-to-read font condition gave more “can’t say” responses to the distorted question whereas no participant responded “can’t say” to the undistorted question. However, the authors only used two questions in each experiment, one distorted (the original Moses question) and one undistorted (“What country is famous for cuckoo clocks, chocolate, banks, and pocketknives?” – Switzerland). Importantly, the authors did not use distorted and undistorted versions of the same questions, but rather different questions entirely. As such, the possibility that the results were due to a bias shift rather than increased sensitivity cannot be fully ruled out.

Further research by Lee et al. (2015, Experiment 1) found that fish odor (a “fishy smell”) increased participants’ sensitivity to distortions. The authors used fish oil and water (control condition) to make two different booths smell like fish or smell neutral before assigning participants to one of the booths based on condition. Participants then responded to one distorted question and one undistorted question. Participants without knowledge about the biblical story of Noah were excluded and the results showed that participants in the booth with fish smell answered “can’t say” to the distorted question (but not the undistorted question) more often than participants in the neutral smelling booth. However, as with the previously mentioned study, this experiment only used the original single Moses question as the distorted question and the question about Switzerland as the undistorted question, limiting the generalizability of these findings (see Judd et al., 2012, on the necessity of stimulus sampling).

### ***Interindividual differences***

A different line of research investigates the influence of individual differences on the susceptibility to the Moses illusion. Hannon and Daneman (2001) replicated earlier findings that increased semantic relatedness between the distorted and the undistorted word increased

the illusion, but they assessed two cognitive measures for each participant: knowledge access and working memory span. The knowledge access measure consisted of statements that required participants to “access and reason about prior knowledge in long-term memory” (p. 452), similar to how the detection of distorted questions requires matching with existing memory structures. The working memory span measure was designed to measure participants’ working memory during reading, specifically their processing and storage capacity. The authors found that both measures combined accounted for 36% of the variance in distortion sensitivity, suggesting that how susceptible people are to the Moses illusion differs between individuals and is moderated by how well they access information from memory and how well they process and store information while reading.

Another individual factor that influences the strength of the Moses illusion is age. Umanath et al. (2014) investigated whether older age (and thus, more lifetime memories) makes one more resilient against the Moses illusion. Furthermore, they examined if older age protects participants from the negative memorial consequences earlier research had found (Bottoms et al., 2010). They found that, when faced with distorted questions, older adults responded incorrectly more often (50%) than younger adults did (41%), but they were less suggestible than younger adults. After the typical Moses illusion-paradigm, participants had to respond to open-ended questions targeted at the (un)distorted term (e.g., “Who took two animals of each kind on the ark?”),<sup>1</sup> and suggestibility was measured as the percentage of responses that had been influenced by the distorted question in the main paradigm (e.g., the response “Moses” to the previous example question). Younger adults were more suggestible (6%) than older adults (4%), but suggestibility was low overall. This pattern could be due to older adults’ prior knowledge leading to stronger memory structures, which on one hand increase partial matching (leading to more Moses responses) but also protect them against memorial consequences (leading to less susceptibility to suggestion).

### ***Expertise***

One aspect that seems plausible to moderate the Moses illusion is expertise. It is reasonable to assume that experts are less susceptible to the Moses illusion when it comes to their field of expertise. Cantor and Marsh (2017) investigated this expertise hypothesis. They constructed an array of 60 questions that pertained to either history or biology and recruited their participants from graduate programs in biology and history, making each participant an expert in one of the fields. They then instructed participants to specifically search for errors (i.e., distortions) within the questions they would see and respond “wrong” if they noticed an error. Experts gave less Moses responses in their field of expertise and a signal detection analysis showed that this was due to improved detection rather than just a shift of response bias (i.e., a general tendency to state “can’t say”).

### ***Motivation***

So far, the presented research used questions and statements to which the answers had no consequences for participants. Going back to the cocktail party situation, one may ask if the response would change if a \$5 bet is made on the correctness of the answer.

Speckmann and Unkelbach (2021) argued that being highly motivated should reduce the number of Moses responses. They used a multiple-choice format and provided participants with monetary incentives: They monetarily rewarded correct responses and monetarily punished incorrect ones. Monetary incentives are commonly assumed

to increase effort (Bonner & Sprinkle, 2002) and effort could lead to a more rigorous matching procedure between the question and memory structures. In other words, if participants stand to gain a substantial amount of money through correct responses, they should be motivated to “think harder” before responding to the questions.

Participants were assigned to one of three conditions: high incentives, low incentives, or no incentives. However, despite their increased effort as evidenced by higher response times, higher incentives only lead to marginally less Moses responses. Furthermore, participants in the high incentives condition chose to skip a question more often, thereby neither gaining nor losing money, suggesting a shift of response bias to avoid risk. This means that even high monetary incentives do not make people use a complete matching strategy instead of a partial matching strategy.

## **Implications**

The Moses illusion is a highly robust phenomenon. However, unlike other cognitive illusions in this book, the Moses illusion in its experimental form has no real-life analogy. Nevertheless, by studying the Moses illusion, one learns much about the process of comprehending language. The illusion illustrates the interaction of bottom-up (i.e., encoding) and top-down (i.e., memory retrieval) processes in answering questions. In other words, the illusion illustrates the difference between what is said and what is understood and responded to. People perceive questions in an active and constructive manner that is adaptive in most situations. Most of the time, if someone makes a mistake while asking a question, they do not do so intentionally. In that case, the constructive process in language comprehension and the ignoring of mistakes (i.e., partial mismatches), fosters conversational flow between two conversation partners by blending together what is said, what is meant, and what is understood. Researchers can create conversation situations where the Moses illusion is maladaptive, but in most real-life conversational situations, the processes that underlie the illusion help communication partners understand each other, and are a potential showcase for the adaptive nature of cognition (see Reber & Unkelbach, 2010).

## **Conclusion**

The Moses illusion is a robust illusion, and partial matching is the best explanation for it. Because people do not fully match questions with the corresponding memory structures, but rather rely on heuristic decision-making based on the activation in the semantic network, they overlook the distorted term “Moses” and respond as if the question were undistorted. This seemingly faulty process may be factually adaptive in everyday life. People often lack the cognitive capacity to completely match each query they receive with their memory structures. Partial matching is thus a highly efficient strategy that is likely to produce the correct response more often than not. The research tradition addressing the illusion and its processes has nevertheless identified the conditions under which the processes lead to predictable errors. Going back to the example from the beginning of this chapter: If you do ask someone at a cocktail party “How many animals of each kind did Moses take on the ark?”, you will probably notice that the other person knows you are asking them a trick question and they will do their best to respond correctly, but given the circumstances, they will very likely be unable to do so.

## Summary

- People tend to overlook distortions in questions when the distortion consists of replacing a term (not the target of the question) with a semantically related but incorrect term. This is known as the Moses illusion.
- Partial (i.e., incomplete) matching with existing memory structures is the most supported explanation for this illusion.
- The Moses illusion is a robust phenomenon that has few real-world implications but provides insight into human cognition and language comprehension.

## Note

<sup>1</sup> This is not a direct example from the paper. The example the authors used is: “Whose famous soliloquy contained the phrase, ‘To be or not to be, That is the question?’” (p. 483), but we decided to apply this format to the original Moses question as an example for consistency.

## Further reading

Erickson and Mattson (1981) first demonstrated the Moses illusion, Reder and Kusbit (1991) empirically tested the different explanations of the illusion, and Kamas et al. (1996) discussed the importance of differentiating between sensitivity and response bias.

## References

- Bonner, S. E., & Sprinkle, G. B. (2002). The effects of monetary incentives on effort and task performance: Theories, evidence, and a framework for research. *Accounting, Organizations and Society*, 27(4–5), 303–345.
- Bottoms, H. C., Eslick, A. N., & Marsh, E. J. (2010). Memory and the Moses illusion: Failures to detect contradictions with stored knowledge yield negative memorial consequences. *Memory*, 18(6), 670–678.
- Bredart, S., & Modolo, K. (1988). Moses strikes again: Focalization effect on a semantic illusion. *Acta Psychologica*, 67(2), 135–144.
- Cantor, A. D., & Marsh, E. J. (2017). Expertise effects in the Moses illusion: Detecting contradictions with stored knowledge. *Memory*, 25(2), 220–230.
- Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, 20(5), 540–551.
- Grice, H. P. (1975). Logic of conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Speech acts* (Vol. 3, pp. 41–58). New York: Seminar Press.
- Hannon, B., & Daneman, M. (2001). Susceptibility to semantic illusions: An individual-differences perspective. *Memory & Cognition*, 29(3), 449–461.
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103(1), 54–69.
- Kamas, E. N., & Reder, L. M. (1995). The role of familiarity in cognitive processing. In R. F. Lorch & E. J. O’Brien (Eds.), *Sources of coherence in reading* (pp. 177–202). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Kamas, E. N., Reder, L. M., & Ayers, M. S. (1996). Partial matching in the Moses illusion: Response bias not sensitivity. *Memory & Cognition*, 24(6), 687–699.

- Lee, D. S., Kim, E., & Schwarz, N. (2015). Something smells fishy: Olfactory suspicion cues improve performance on the Moses illusion and Wason rule discovery task. *Journal of Experimental Social Psychology*, 59, 47–50.
- Park, H., & Reder, L. M. (2004). Moses illusion. In R. F. Pohl (Ed.), *Cognitive illusions* (pp. 275–291). Hove: Psychology Press.
- Reber, R., & Unkelbach, C. (2010). The epistemic status of processing fluency as source for judgments of truth. *Review of Philosophy and Psychology*, 1(4), 563–581.
- Reder, L. M., & Kusbit, G. W. (1991). Locus of the Moses illusion: Imperfect encoding, retrieval, or match? *Journal of Memory and Language*, 30(4), 385–406.
- Shafto, M., & MacKay, D. G. (2000). The Moses, mega-Moses, and Armstrong illusions: Integrating language comprehension and semantic memory. *Psychological Science*, 11(5), 372–378.
- Song, H., & Schwarz, N. (2008). Fluency and the detection of misleading questions: Low processing fluency attenuates the Moses illusion. *Social Cognition*, 26(6), 791–799.
- Speckmann, F., & Unkelbach, C. (2021). Moses, money, and multiple-choice: The Moses illusion in a multiple-choice format with high incentives. *Memory & Cognition*, 49(4), 843–862.
- Umanath, S., Dolan, P. O., & Marsh, E. J. (2014). Ageing and the Moses illusion: Older adults fall for Moses but if asked directly, stick with Noah. *Memory*, 22(5), 481–492.
- Unkelbach, C., & Greifeneder, R. (2013). The experience of thinking. In C. Unkelbach & R. Greifeneder (Eds.), *The experience of thinking: How the fluency of mental processes influences cognition and behavior*. Hove: Psychology Press.
- van Oostendorp, H., & De Mul, S. (1990). Moses beats Adam: A semantic relatedness effect on a semantic illusion. *Acta Psychologica*, 74(1), 35–46.
- van Oostendorp, H., & Kok, I. (1990). Failing to notice errors in sentences. *Language and Cognitive Processes*, 5(2), 105–113.

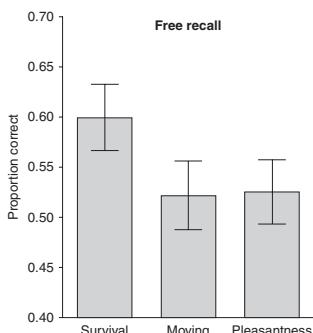
## 23 Survival processing effect

*Meike Kroneisen and Edgar Erdfelder*

Why do we remember? Research on memory is often concerned with the question 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 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 et al. (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 23.1 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 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 five seconds, and participants were asked to rate the words on five-point scales, with 1 indicating totally irrelevant (or unpleasant) and 5 signifying extremely relevant (or pleasant). The material consisted of typical exemplars 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



*Figure 23.1* Average proportion of correct recall, sorted by condition (survival scenario v. 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*, 33, 265. © 2007 American Psychological Association. Adapted with permission of the publisher).

retention than did the other encoding procedures. Figure 23.1 illustrates 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 et al., 2010), when survival processing is compared to other memory-enhancing encoding tasks (Kroneisen & Makerud, 2017; Nairne et al., 2008), or when using emotionally arousing control scenarios rather than the moving control scenario (e.g., Kang et al., 2008; Kroneisen et al., 2022). It could even be used to learn vocabulary in a second language (Kazanas et al., 2020). Furthermore, the survival processing advantage was successfully replicated as part of the Open Science Collaboration project (Open Science Collaboration, 2015). The survival processing effect also shows up in judgments of learning (Palmore et al., 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). Overall, it is a stable effect with medium to large effect sizes (Scofield et al., 2017).

Moreover, Weinstein et al. (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 et al., 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). Relatedly, Parker et al. (2021) more recently observed that survival processing not only boosts recall but also directed forgetting (i.e., more forgetting of material following the forget-cue in list-method directed forgetting). Also, survival processing does not enhance performance in a Stroop color-naming task (Kazanas et al., 2015). Apparently, considering one's survival when performing attention-based and memory tasks does not enhance cognitive performance in general. In addition, no survival processing benefit was found for memory of faces (Savine et al., 2011), for memory of serial order (Wöstenfeld et al., 2020) or for indirect memory tests (Tse & Altarriba, 2010) or when presenting the scenarios in a second language (Saraiva et al., 2020). 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 et al., 2012). Moreover, Klein (2013) demonstrated that no further context is necessary to produce the survival benefit on memory. He compared the grassland scenario with a context-free survival situation ("try to stay alive"). Both scenarios produced equivalent levels of recall. Klein also concluded that survival processing per se is too broad to be shaped by natural selection. 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.

### Text box 23.1 Classroom experiment

#### Material

Stimulus materials are taken from the updated Battig and Montague norms (Van Overschelde et al., 2004) and consist of 30 typical members drawn from 30 unique categories (see list below). All words are presented in random order.

---

*Word material 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

---

#### 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 two minutes 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 five seconds 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 one second 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 ten minutes.

### **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 et al. (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 et al., 2007, 2008) 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 et al. (2010, 2011) saw memory as future-orientated. 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 emotional 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 et al., 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; Kroneisen et al., 2022; Soderstrom & McCabe, 2011) or by stress (Smeets et al., 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 et al., 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 for "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 that explains the effect but specific aspects of arousal. 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 et al., 2009, 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 et al. (2014b) 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 to that required by the survival orienting task. Bell et al. (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 et al. (2014a) 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 et al. (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 scenario. Similar results were repeatedly observed by other researchers (e.g., Kroneisen & Erdfelder, 2011; Kroneisen et al., 2013, 2014, 2016; 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 et al. (2013) showed that the strength of the survival processing effect is a function of the number of unique relevance arguments generated per item. In line with this, Wilson (2016) showed that the survival scenario elicited more alternative uses in the Guilford's Alternate Uses Test compared to non-survival related conditions. Accordingly, Kroneisen et al. (2021) found that the survival processing effect is more pronounced for objects low in functional fixedness (i.e., objects that can be used in multiple ways) compared to objects high in functional fixedness that are linked to a

specific function. Moreover, Bell et al. (2015) 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. Four independent studies indeed showed that survival processing diminishes under working memory load (Kroneisen et al., 2014, 2016; Nouchi, 2013; Yang et al., 2021; see, however, Stillman et al., 2014, and Kroneisen et al., 2016, for some discrepant results and a potential explanation).

If participants generate unique relevance arguments during survival processing selectively (e.g., using a chair to defend oneself against predators; melting snow to drink it later ...), this should not only help them to recall or recognize these items later on, it should also allow them to specifically attribute the item to the survival scenario context. More precisely, participants should remember well that an item was encoded in a survival condition (i.e., they should have better source memory for the survival scenario compared to alternative scenarios). Indeed, there is evidence that survival processing improves memory for context (Clark & Bruno, 2016; Kroneisen & Bell, 2018; Misirlisoy et al., 2019; Nairne et al., 2012). However, there are also studies that did not find a source memory advantage for words processed in a survival scenario (Bröder et al., 2011; Savine et al., 2011; Nairne et al., 2015; Hou & Liu, 2019).

Current research goes beyond what can be learned from behavioral measures alone. Recently, Event Related Potential (ERP) research demonstrated that survival processing is associated with an increased frontal slow wave (Forester et al., 2019, 2020a). These findings suggest that survival processing is not associated with lower level encoding processes, which are sensitive to motivation and stimulus salience, but rather with more elaborate forms of encoding. This is also in line with the finding that reward motivation does not moderate the survival processing effect (Forester et al., 2020b).

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 et al., 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 increase or decrease 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 chose not 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 which are 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 is a robust and general phenomenon. In fact, the data support the conclusion that it is one of the most efficient encoding procedures identified in human episodic-memory research so far. Given this robustness and 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 of the survival processing advantage 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 thus “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 and stress within the survival situation. However, the empirical evidence for this account is inconsistent.
  - 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, Nairne and Pandeirada (2011) did not.
  - Richness of encoding: The survival scenario fosters 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 improved 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 overview of the literature can be found in Erdfelder and Kroneisen (2014) and in Kazanas and Altarriba (2015). 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* (Howe & Otgaar, 2014) that covers selected theoretical and empirical aspects of the survival processing effect.

## References

- Aslan, A., & Bäuml, K.-H. T. (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. (2014a). Dying scenarios improve recall as much as survival scenarios. *Memory*, 22, 51–64.
- Burns, D. J., Hart, J., Kramer, M. E., & Burns, A. D. (2014b). 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.
- Clark, D. P., & Bruno, D. (2016). Fit to last: Exploring the longevity of the survival processing effect. *Quarterly Journal of Experimental Psychology*, 69, 1164–1178.
- 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.
- Forester, G., Kroneisen, M., Erdfelder, E., & Kamp, S.-M. (2019). On the role of retrieval processes in the survival processing effect: Evidence from ROC and ERP analyses. *Neurobiology of Learning and Memory*, 166, Article 107083.
- Forester, G., Kroneisen, M., Erdfelder, E., & Kamp, S.-M. (2020a). Survival processing modulates the neurocognitive mechanisms of episodic encoding. *Cognitive, Affective, and Behavioral Neuroscience*, 20, 717–729.
- Forester, G., Kroneisen, M., Erdfelder, E., & Kamp, S.-M. (2020b). Adaptive memory: Independent effects of survival processing and reward motivation on memory. *Frontiers in Human Neuroscience*, 14(article 588100), 1–13.
- Hart, J., & Burns, D. J. (2012). Nothing concentrates the mind: Thoughts of death improve recall. *Psychonomic Bulletin & Review*, 19, 264–269.
- Hou, C., & Liu, Z. (2019). The survival processing advantage of face: The memorization of the (un)trustworthy face contributes more to survival adaptation. *Evolutionary Psychology*, 17, 1474704919839726.
- 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.
- Howe, M. L., & Otgaar, H. (2014). What kind of memory has evolution wrought? Introductory article for the Special Issue, Adaptive memory: The emergence and nature of proximate mechanisms. *Memory*, 22, 1–8.

- 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.
- Kazanas, S. A., & Altarriba, J. (2015). The survival advantage: Underlying mechanisms and extant limitations. *Evolutionary Psychology, 13*, 360–396.
- Kazanas, S. A., Altarriba, J., O'Brien, E. G. (2020). Paired-associate learning, animacy, and imageability effects in the survival advantage. *Memory & Cognition, 48*, 244–255.
- Kazanas, S. A., Van Valkenburg, K. M., & Altarriba, J. (2015). Survival processing and the Stroop task: Does the survival advantage depend on deeper processing during encoding? *Evolutionary Psychology, 13*, 1–8.
- 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. (2013). Does optimal recall performance in the adaptive memory paradigm require the encoding context to encourage thoughts about the environment of evolutionary adaptation? *Memory & Cognition, 41*, 49–59.
- 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., & Bell, R. (2018). Remembering the place with the tiger: Survival processing can enhance source memory. *Psychonomic Bulletin & Review, 25*, 667–673.
- 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., Kriechbauer, M., Kamp, S.-M., & Erdfelder, E. (2021). How can I use it? The role of functional fixedness in the survival-processing paradigm. *Psychonomic Bulletin & Review, 28*, 324–332.
- Kroneisen, M., Kriechbaumer, M., Kamp, S.-M., & Erdfelder, E. (2022). Realistic context doesn't amplify the survival processing effect: Lessons learned from Covid-19 scenarios. *Acta Psychologica, 222*, 103459.
- Kroneisen, M., & Makerud, S. E. (2017). The effects of item material on encoding strategies: Survival processing compared to the method of loci. *Quarterly Journal of Experimental Psychology, 70*, 1824–1836.
- Kroneisen, M., Rummel, J., & Erdfelder, E. (2014). Working memory load eliminates the survival processing effect. *Memory, 22*, 92–102.
- Kroneisen, M., Rummel, J., & Erdfelder, E. (2016). What kind of processing is survival processing? Effects of different types of dual-task load on the survival processing effect. *Memory & Cognition, 44*, 1228–1243.
- 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.
- Misirlisoy, M., Tanyas, H., & Atalay, N.B. (2019). Does survival context enhance memory for source? A within-subjects comparison. *Memory, 27*, 780–791.
- 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., Pandeirada, J. N., VanArdsall, J. E., & Blunt, J. R. (2015). Source-constrained retrieval and survival processing. *Memory & Cognition, 43*, 1–13.
- 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.
- Nairne, J. S., VanArdsall, J. E., Pandeirada, J. N. S., & Blunt, J. R. (2012). Adaptive memory: Enhanced location memory after survival processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 495–501.
- 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.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*, article 6251.
- 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.
- Parker, A., Parkin, A., & Dagnall, N. (2021). Effects of survival processing on list method directed forgetting. *Memory*, electronic preprint. doi: 10.1080/09658211.2021.1931338.
- 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.

- Saraiva, M., Garrido, M. V., & Pandeirada, J. N. S. (2020). Surviving in a second language: Survival processing effect in memory of bilinguals. *Cognition and Emotion*, 35(2), 417–424.
- 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.
- Scofield, J. E., Buchanan, E. M., & Kostic, B. (2017). A meta-analysis of the survival-processing advantage in memory. *Psychonomic Bulletin & Review*, 25, 997–1012.
- 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. (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.
- Yang, L., Truong, L., & Li, L. (2021). Survival processing effect in memory under semantic divided attention. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*. Advance online publication. <https://doi.org/10.1037/cep0000210>
- Weinstein, Y., Bugg, J. M., & Roediger, H. L. (2008). Can the survival recall advantage be explained by the basic memory processes? *Memory & Cognition*, 36, 913–919.
- Wilson, S. (2016). Divergent thinking in the grasslands: Thinking about object function in the context of a grassland survival scenario elicits more alternate uses than control scenarios. *Journal of Cognitive Psychology*, 28, 618–630.
- Wöstenfeld, F. O., Ahmad, S., Kroneisen, M., & Rummel, J. (2020). Does the survival memory advantage translate to serial recall? *Collabra: Psychology*, 6(1), 8.

## 24 Labeling and overshadowing effects

Rüdiger F. Pohl

The labeling 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 13, the misinformation effect in Chapter 26, and the hindsight bias in Chapter 27). Text box 24.1 lists a number of examples demonstrating labeling effects in different domains. This list already suggests that labeling effects may be rather widespread in our lives. More examples, from more applied domains, are given in Text box 24.2.

### Text box 24.1 Examples of labeling 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 et al., 2003). An odor that was introduced as “pleasant” received higher hedonic ratings than the same odor 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 et al., 1932). A blue-green color 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 24.2 Applied domains where labeling effects might occur

Apart from the cognitive paradigms reviewed in this chapter, labeling occurs in other, more applied contexts as well. For example, developmental psychologists study the effects of labeling on young infants’ ability to categorize objects by transitive inference (e.g., Sloutsky et al., 2001; Weller & Graham, 2001). Social

psychologists are interested on how labeling contributes to impression formation (Higgins et al., 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 et al., 1977; but see Katz et al., 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. 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, labeling 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 label 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).

## **Labeling effects and their theoretical accounts**

This section presents cognitive domains in which labeling effects have been found. I begin with the classical examples studying the effects of labeling on the memory for ambiguous line drawings and other visual objects. Using the misinformation design (see Chapter 26), other researchers investigated how labeling might affect the memory for color or speed. Another domain of labeling concerns judgments of taste and odor. And finally, I present an example from judging differently labeled political statements.

### ***Memory for visual objects***

In one of the most influential studies on the labeling effect, Carmichael et al. (1932) presented a set of relative 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 24.1 for the complete material). These words had 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% (with Label 1 or 2, respectively) could be classified as being closer to the named object 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”.

Text box 24.3 describes a classroom demonstration of labeling effects, which has been adapted from the described experiment by Carmichael et al. (1932). This classroom demonstration is a simplified version and should be fairly easy to set up and to analyze.

### **Text box 24.3 Classroom experiment demonstrating the labeling effect**

As a classroom demonstration of labeling effects, this text box describes an adapted version of the classical experiment by Carmichael et al. (1932) on the effects of labeling on the reproduction of ambiguous line drawings.

#### **Method**

##### ***Materials and participants***

The complete material is given in Table 24.1. 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 ( $\chi^2$ -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 versus 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 five seconds 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 labeling

Table 24.1 The 12 ambiguous visual stimuli with two lists of labels

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

Source: Carmichael et al. (1932).

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 participant 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 ( $\chi^2$ -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,  $\chi^2$ -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).

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 labeling 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 five minutes, 20 minutes, or two 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 S’s 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 catalog). 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 v. 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 color***

Another group of researchers looked at the effects of labeling on the memory for color. For example, Thomas and DeCapito (1966) found that an ambiguous blue-green color (with a wave length of 490 nm) that was called either "blue" or "green" by individual participants led to a differential generalization to bluer or greener colors, respectively.

Following this experimental idea, Bornstein (1976) first determined the point of subjective equality (PSE) for calling a color "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 eight distractors (with four items having shorter and four 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 seconds 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 labeling effect. The distribution of mean probabilities of saying “same color” moved towards the blue colors, when the label was “bluish”, and towards the green colors, 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 26), Loftus (1977) let her participants first view a series of color 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 color term. After a 20-minute filler task, a color recognition test was applied to both groups. It consisted of 30 simultaneously presented color strips covering the whole range of visible colors. Participants were asked to identify the color for ten 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 color 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 colors. Loftus interpreted these results as showing the blending of the original trace with the subsequently presented false information (see also Belli, 1988, and Pohl & Gawlik, 1995, for more evidence on blends in memory as well as a critical discussion of theoretical conclusions).

In another misinformation experiment on color 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 what was allegedly a taste study. After a 15 minute 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 minute distractor task, the same recognition test as in the Loftus (1977) study was administered. Participants were asked to pick the correct color from 30 color strips for ten 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) color, 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 labeling 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 1980s and 1990s. 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 labeling effects (see the discussion in Chapter 26 and the SARA model in Chapter 27).

More critique on the Loftus and Palmer (1974) results, however, accrued when two attempts to replicate them failed (McAllister et al., 1988; Read et al., 1978). Presumably, labeling 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, color, taste, or odor), 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 labeling effect (cf. the problems of replicating "verbal overshadowing" described below).

### ***Judgments of taste and odor***

A few studies investigated the effects of labels on taste judgments. Although these studies did not directly address the question of how labeling influences memory, it appears fairly safe to conclude that the observed judgment distortions also generalize to corresponding memory distortions. Besides, there is at least one explicit demonstration of labeling effects on taste memory (Melcher & Schooler, 1996).

In one of the first studies, Allison and Uhl (1964) asked experienced beer drinkers to taste different beer brands that were either correctly labeled or not labeled at all. One result was that participants were not able to reliably distinguish unlabeled beer brands. However, when the bottles were correctly labeled, participants rated the taste of their preferred brand as superior. In addition, beer from labeled bottles was generally judged to taste better than beer from unlabeled bottles. The authors concluded that perceptual characteristics (like the brand names) are much more important for taste judgments than actual taste characteristics. However, the results from later studies were not so clear. For example, Mauser (1979) replicated the study under more experimental control and found

that his participants (experienced beer drinkers, too) were able to distinguish unlabeled beer brands and exhibited similar preferences as with correctly labeled bottles.

In another study demonstrating labeling 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 labeling 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.

Plassmann et al. (2008) reported that wine was judged as tasting better when participants believed that it was more expensive. Interestingly, the authors scanned their participants’ brain activity during the task (using fMRI) and found increased activity in the medial orbitofrontal cortex, an area that is considered to encode experienced pleasantness, suggesting that participants did indeed feel more pleasant consuming an allegedly expensive wine.

Even labels that convey essentially no information about the tastiness of a product showed an influence. Lotz et al. (2013) let their participants taste coffee or chocolate that was either attached with a “Fair Trade” label or not. The results showed that products with a “Fair Trade” label were judged as tasting better than products without such a label. In addition, the authors found that subjectively experienced affect during tasting mediated the labeling effect.

Labeling effects were also reported for odor perception. Herz and von Clef (2001) presented a set of five different odors repeatedly in two sessions (one week apart). Each odor 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 the hedonic-value suggestion was manipulated as a *within-subjects* factor, thus allowing a strong test of the labeling effect. The five odors (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 odors. As a result, each odor was combined with a positive or a negative label in the two sessions (e.g., pine oil was either labeled as “Christmas tree” in Session 1, i.e., positive, or as “spray disinfectant” in Session 2, i.e., negative). These labels were given immediately before each odor was sniffed. The results revealed clear effects of the given labels. If an odor was suggested to be pleasant (positive), it received much higher hedonic ratings than when the same odor was suggested to be unpleasant (negative). Recall that these ratings were given by the same participants (and not different ones), thus showing indeed strong labeling effects on odor judgments.

### ***Other areas of judgment***

Generally, labeling effects might occur whenever labels are attached to something. This pertains, for example, to many areas of social cognition (e.g., stereotyping; see also Text box 24.2 and Chapter 16 on halo effects). One recent example comes from

studying agreement with political statements (see also Cohen, 2003): Neumann et al. (2020) presented a number of political statements that were either labeled as right-wing or not and asked participants to judge their agreement with these statements. The results showed an interaction between labeling condition and political orientation of the participants: Right-wing supporters agreed more with the statements when they were labeled as “right-wing” (compared to no label), whereas non-supporters of right-wing parties agreed less with them when they were labeled as “right-wing” (compared to no label).

## **Verbal overshadowing**

While, for the examples presented so far, the label was always provided by the experimenter, labeling 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 et al., 1997, for a review). Negative effects of verbalization on memory have meanwhile been shown for shapes, colors, 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 et al., 2008). Recently, the original study by Schooler and Engstler-Schooler (1990) has been replicated in an international replication project, as will be described below.

### **Verbal overshadowing on face recognition**

Introducing the concept of “verbal overshadowing”, Schooler and Engstler-Schooler (1990, Exp. 1 and 2) studied the negative effects of verbalization on visual memory for faces. Participants watched a video depicting a salient person (a bank robber). After a 20-minute 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 (including the target person). The results showed that recognition accuracy was substantially impaired in the verbalization group compared to the control group.

Several later studies, however, failed to replicate this effect, leading Schooler et al. (1997) to admit that the overshadowing 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.

In an attempt to test the robustness of this type of verbal overshadowing, the editors of *Perspectives on Psychological Science* initiated a large-scale replication project, involving 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 negative 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” (2014, p. 556).

The negative effect of verbalization nevertheless remained under debate and led to further studies some of which even reported positive effects of verbalization. The attention that this effect received is probably also due to its important practical implications, for example, in legal settings. Mickes and Wixted (2015; Mickes, 2016) stressed this applied perspective and warned against premature conclusions as long as the likelihood of identifying an innocent suspect from a target-absent lineup (the false identification rate) is not known. The studies so far only looked at the hit rate that was diminished in the verbalizing condition, but not at the false-alarm rate in target-absent lineups. The false-alarm rate is necessary to test whether the decreased hit performance in the verbalization condition is simply due to a stricter criterion in responding (in comparison to the control condition), and not to an impaired memory (Clare & Lewandowsky, 2004). In an accordingly improved study, Wilson et al. (2018) still found an overshadowing effect, even after controlling for criterion shifts.

Using a changed procedure, Sporer et al. (2016) reported facilitation effects of verbalization. They had their participants reread their verbal descriptions of a perpetrator (who was previously seen in a film) prior to identifying him in a lineup. This procedural change apparently helped in discriminating between the visual and the verbal memory trace of the seen face, presumably by reinstating the respective encoding context.

### ***Verbal overshadowing on other visual materials***

Schooler and Engstler-Schooler (1990, Exp. 3) presented one of three different color swatches (red, blue, or green) to their participants. Immediately after that, participants either wrote descriptions of the seen color (verbal), visualized it (visual), or did an unrelated task (control) for 30 seconds. In a subsequent recognition test, the target color was shown together with five distractors containing similar color hues. The whole procedure was then repeated for the other two color swatches. The results showed that, compared to the control condition, recognition for the original color was impaired when it had been verbally described, whereas visualizing the color had no such effect.

In a more recent study on the memory for color, Souza and Skóra (2017) did not replicate the verbal-overshadowing effect. They rather found *increased* performance after verbalization. During color presentation participants were instructed to either name the color aloud (labeling) or to repeat “ba-ba-ba” aloud (articulatory suppression). Later, using a color wheel, identification of the previously seen colors was better in the labeling than in the suppression condition. The authors argued that “labeling activates visual long-term categorical representations which help in reducing the noise in the internal representations of the visual stimuli” (p. 277). Using the same approach, Forsberg et al. (2020) replicated these findings and extended their investigation to age-related differences. Whereas both younger and older adults profited from labeling the colors, the effects of labels appeared to be somewhat different in both age groups boosting different processes.

### ***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. 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 et al. (2002), participants heard a voice speaking one sentence, then worked on a 10-minute 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 et al. (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 reanalyzed, a 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 who had given a verbal description of the original voice recognized it much less often than those who had not given such a description. But again, these studies also showed that the appearance of verbal overshadowing may hinge on subtle procedural conditions.

In a completely different domain, namely assessment of hunger, Creswell et al. (2018) measured subjectively felt hunger. They compared a non-verbal measure of hunger (i.e., squeezing a handheld dynamometer) with a verbal one that simply consisted of a rating scale (from 0 to 100) with verbally labeled end-points. The authors found that the non-verbal measure predicted later eating behavior better, but only when it was not contaminated by a preceding verbal measure, thus demonstrating the detrimental effect of verbal overshadowing.

Summarizing the studies reported in this section, it is clear that verbal overshadowing is not always observed, but rather depends on a number of experimental details (like features of the materials, the participants, and the exact procedure).

### ***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 labeling information (cf. 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 (described above) 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. They also assumed 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).

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 et al. (2008) identified several moderators of verbal overshadowing (like single v. 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).

More recently, Souza and Skóra (2017) discussed four different explanations of verbal overshadowing and facilitation effects for visual stimuli, especially color: (1) Encoding of the label simply distorts memory for the visual stimulus. (2) Both visual and verbal trace coexist (like in the dual-trace approach), but may interfere at retrieval. (3) The label

helps to make the visual stimulus more distinctive and acts as an additional retrieval cue (and thus augments memory). (4) The label activates categorical (visual) information that could be helpful or not (depending on the relation of the label to the visual stimulus). Souza and Skóra concluded from their studies (described above) that the last explanation received the most support from their finding of a facilitative effect of labels. But as outlined above, other studies found detrimental effects and would thus presumably support other explanations.

## Summary

- A label is something that is affixed to a stimulus. A labeling effect occurs if judging or recalling that stimulus is systematically influenced by and in the direction of the label.
- Labeling effects have been found for the recall of visual forms and color, and for the judgment of speed, taste, and odor.
- Labeling 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, colors, 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.
- Verbal overshadowing seems to be a fragile effect with small effect size.
- Verbalization can also improve memory.

## Further reading

An enlightening exercise in classical experimenting is to read the original labeling study by Carmichael et al. (1932). As examples of less-well researched labeling domains, the studies by Pohl et al. (2003) on taste and by Herz and von Clef (2001) on odor judgment are noteworthy. With respect to “verbal overshadowing”, Schoeller et al. (1997) presented an early review of the field, and Meissner and Brigham (2001) a meta-analysis. Schoeller (2002) discussed in detail several theoretical approaches. Meissner et al. (2008) presented a very thorough meta-analysis. More recently, the replication project yielded valuable results (Alogna et al., 2014).

## References

- Allison, R. I., & Uhl, K. P. (1964). Influence of beer brand identification on taste perception. *Journal of Marketing Research*, 1(3), 36–39.
- Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., Birt, A. R., ... & Zwaan, R. A. (2014). Registered replication report: Schoeller & Engstler-Schoeller (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.
- Clare, J., & Lewandowsky, S. (2004). Verbalizing facial memory: Criterion effects in verbal overshadowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(4), 739–755.

- Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology*, 85(5), 808–822.
- Creswell, K. G., Sayette, M. A., Schooler, J. W., Wright, A. G. C., & Pacilio, L. E. (2018). Visceral states call for visceral measures: Verbal overshadowing of hunger ratings across assessment modalities. *Assessment*, 25(2), 173–182.
- 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.
- Forsberg, A., Johnson, W., & Logie, R. H. (2020). Cognitive aging and verbal labeling in continuous visual memory. *Memory & Cognition*, 48(7), 1196–1213.
- 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. T. (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.
- Lotz, S., Christandl, F., & Fetchenhauer, D. (2013). What is fair is good: Evidence of consumers' taste for fairness. *Food Quality and Preference*, 30(2), 139–144.
- 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.
- Mauser, G. A. (1979). Allison & Uhl revisited: The effects of taste and brand name on perceptions and preferences. In W. L. Wilkie (Ed.), *NA – Advances in consumer research* (Vol. 6, pp. 161–165). Ann Arbor, MI: Association for Consumer Research.
- 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.
- Memon, A., & Meissner, C. A. (Eds.). (2002). Investigations of the effects of verbalization on memory [Special issue]. *Applied Cognitive Psychology*, 16(8).
- Mickes, L. (2016). The effects of verbal descriptions on eyewitness memory: Implications for the real-world. *Journal of Applied Research in Memory and Cognition*, 5(3), 270–276.
- Mickes, L., & Wixted, J. T. (2015). On the applied implications of the "verbal overshadowing effect". *Perspectives on Psychological Science*, 10(3), 400–403.
- 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.

- Neumann, H., Thielmann, I., & Pfattheicher, S. (2020). Labelling affects agreement with political statements of right-wing populist parties. *PLOS ONE*, 15(11), e0239772.
- Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, 76, 241–263.
- Paivio, A. (1971). *Imagery and verbal processes*. Ballein, MO: Holt, Rinehart, & Winston.
- Perfect, T. J., Hunt, L. J., & Harris, C. M. (2002). Verbal overshadowing in voice recognition. *Applied Cognitive Psychology*, 16, 973–980.
- Plassmann, H., O'Doherty, J., Shiv, B., & Rangel, A. (2008). Marketing actions can modulate neural representations of experienced pleasantness. *Proceedings of the National Academy of Sciences*, 105(3), 1050–1054.
- 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). New York: 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.
- Souza, A. S., & Skóra, Z. (2017). The interplay of language and visual perception in working memory. *Cognition*, 166, 277–297.
- Sporer, S. L., Kaminski, K. S., Davids, M. C., & McQuiston, D. (2016). The verbal facilitation effect: Re-reading person descriptions as a system variable to improve identification performance. *Memory*, 24(10), 1329–1344.
- 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.
- Wilson, B. M., Seale-Carlisle, T. M., & Mickes, L. (2018). The effects of verbal descriptions on performance in lineups and showups. *Journal of Experimental Psychology: General*, 147(1), 113–124.
- 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.

# 25 Associative memory illusions

*Henry L. Roediger, III, and David A. Gallo*

We usually trust our memories. If we get into a discussion with a friend or family members about “what really happened” in some disputed past event, we believe our own memories and not our friend’s. But what if we are wrong? Can we remember events differently from the way they really happened? Can we remember events that never happened at all? The answer turns out to be *yes* to both questions.

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 (see Chapter 23). 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 et al., 1998). 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 25.1 and discussed in the next section.

### **Text box 25.1 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.

#### **Material**

The material consists of four 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 one to two 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 participants how they are constructed. You can find many lists to use in a paper by Stadler et al. (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 participants 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 et al., 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 25.1 – 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 Roediger et al., 2001, for a set of 55 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 eight of the lists immediately after their presentation (with two 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 eight 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. That is, they saw words one at a time and said yes or no as to whether or not the item had been studied in the list. Because only 16 of the 24 lists had been studied, items from the other eight 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 procedure 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 Umanath & Coane, 2020, for a critical analysis of the distinction).

### **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 25.1. Shown in the two panels are data from the eight lists that were studied and recalled (the right panel) and from the eight lists that were studied but not previously recalled (the left panel). 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 eight 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 25.1 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 (words that were not actually presented) were practically the same! Participants recognized 81% as old and even judged 58% *remembered*. Thus, the participants recalled words that were never presented, 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. People recall and remember

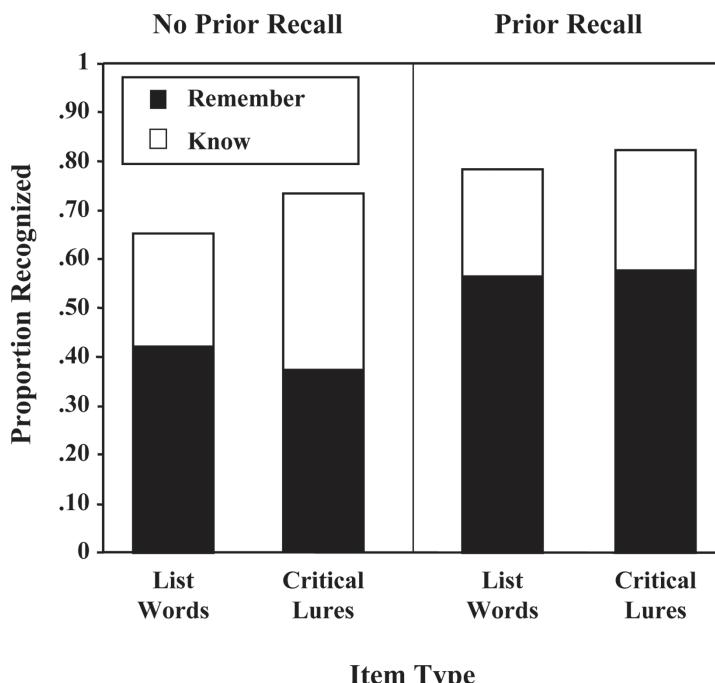


Figure 25.1 The DRM false-recognition effect (Roediger & McDermott, 1995, Exp. 2).

*Note:* 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.

events that never happened. 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 affected by prior false recall. Nonetheless, striking levels of false recognition and “remember” judgments were obtained. And note that the act of recall boosted both accurate and false recall, an example of the effect of retrieval practice on later retention (see McDermott, 2021). Practicing retrieval boosts recall of both veridical and false memories (see also McDermott, 2006).

### ***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 a liberal guessing strategy for related items), in the hopes of facilitating the correct recognition of 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 just 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 et al., 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 26 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 memory illusion. 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 ten years of DRM research, and 15 years of additional research have ensued since then (see Huff et al., 2015, and Pardilla-Delgado & Payne, 2017, for more recent reviews). 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 et al., 1993). That is, did the item really occur during the list presentation, or did I just think about it? 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 et al. (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 et al., 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 et al., 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., one to two seconds 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 et al., 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 et al., 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 et al., 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 based on meaning of the words. 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 et al., 2008). This last point poses important constraints on these theories. Without positing some sort of item-specific activation of the critical 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 et al., 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 et al., 1999). Illustrative data from Toglia et al. (1999) are presented in Figure 25.2. True recall declined rapidly over a three-week retention interval, whereas false recall persisted at high levels.

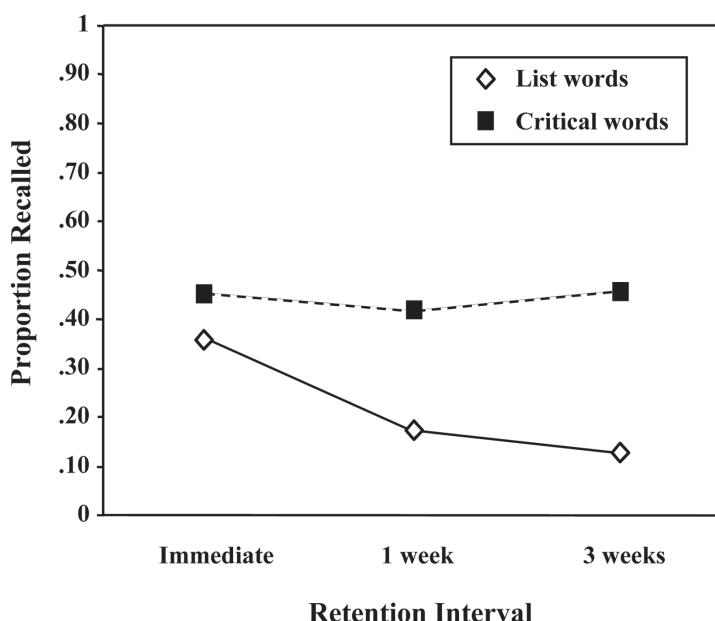


Figure 25.2 The effects of retention interval on true and false recall (Toglia et al., 1999, Exp. 2).

Note: 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.

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% v. 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.

Another line of evidence is relevant to the discussion. Activation theories postulate two different types of activation: lexical (based on the form and sound of the word) and semantic (based on word meaning). As already noted, McDermott's (1997) implicit memory results were credited to lexical activation rather than semantic activation. In addition, Sommers and Lewis (1999) showed that lists of words phonologically associated with a missing word leads to false recall and false recognition. For example, for a word like *speak*, the list words might include *sneak*, *sleek*, *peak*, *cheek*, etc. Sommers and Lewis found that such lists did create recall of the target word that was not presented in the list (*speak*, in this case). Because phonological associates are not related in meaning to the target word, the assumption is that they activated a lexical level of representation and that activation spreads through the lexical network. This idea is in accord with a theory called the Neighborhood Activation Model (Luce, 1987). The model postulates that phonological associates of a word are represented in space, with some closer and some further away in the "neighborhood" of the associates.

Watson et al. (2003) asked what would happen if hybrid lists were created, such that half the items converged on meaning of the missing word and half on its phonology or sound. So, for the word *sleep* (omitted from the list), the list words might include *bed*, *rest*, *awake*, and *tired* for the semantic associates and *sheep*, *peep*, *beep*, and *sleet* for the phonologically similar (rhyming) words. These researchers asked if the false memories that occurred from using such hybrid lists would be additive; that is, would the effects of a combined list with phonological and semantic associates be equal to the sum of presenting the phonological and semantic items when they are presented alone? Watson et al. (2003) expected to find additivity, but found that the hybrid lists created superadditive effects, meaning that the hybrid lists created false memories greater than the sum of the two independent lists that comprised the hybrid lists. That is, they found superadditivity! The activation from both the semantic and phonological levels zeroes in on the missing item and leads to greater false recall. In one study, semantic lists created .34 false recall of the missing item, whereas phonological lists created .14 false recall. By the additivity rule, false recall from hybrid lists should be around .48 (i.e., .34 + .14). However, actual false recall in the hybrid lists was .65, one of the largest false-memory effects ever obtained.

Finley et al. (2017) confirmed the overadditivity effect and plotted out the effects of various levels of hybridization (e.g., 13 phonological associates and three semantic associates). In general, adding only a few semantic associates to a list of mostly phonological associates produced a large increase in false recall; the same was true of adding a few phonological associates to a semantic list.

The phonological false memory effect and the powerful effect of using hybrid lists to create superadditive false memories are explainable in a straightforward manner by associative activation theories. For hybrid lists, activation of the critical missing item cumulates from spreading activation through both a lexical and a semantic network, converging on the nonpresented associate. On the other hand, fuzzy trace theory, with its exclusive focus on meaningful gist as causing false memory effects, seems unable to account for these effects. However, recently Chang and Brainerd (2021) have extended fuzzy trace theory to account for phonological effects, too.

### *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 v. visual, or different voices), participants are often willing to assign a source to critical items that are falsely recognized (Roediger et al., 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 et al., 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 11 on availability, Chapter 14 on the validity effect, Chapter 15 on the mere exposure effect, and Chapter 26 on the misinformation effect). If the attribution process occurs automatically, or nonconsciously, then the phenomenological experience would be one of remembering (cf. Jacoby et al., 1989).

In addition to imagination, events that were actually studied may provide another source of details for these kinds of false memories. Lampinen et al. (1999) used the term “content borrowing” to describe how features of studied items could be retrieved in a fragmented way, such that these free-floating features in memory could then be mistakenly bound into a false memory for the nonpresented associate. This process could cause a detailed yet false recollection of the nonstudied event, especially if the item is processed fluently. Indeed, more recent work on false memories for nonstudied pictures has demonstrated that the conceptual fluency of an item as well as the availability of fragmented perceptual features in memory can independently drive false recollection effects (Doss et al., 2016).

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 examples, see Franks et al., 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 earliest 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 et al., 1999) and presentation modality (e.g., switching presentation from auditory to visual, which reduces the DRM effect; Smith & Hunt, 1998). Although more recent work indicates these kinds of manipulations might impact activation in addition to monitoring processes in the DRM task (e.g., Smith & Hunt, 2020), these same manipulations also have been found to reduce false memory effects in source recollection tasks that do not involve the activation of semantic associations (for review, see Gallo, 2013). Thus, there is converging evidence across different tasks that both presentation format and presentation modality can impact monitoring processes.

Other evidence for monitoring processes in the DRM task 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, 1,000, 3,000, and 5,000 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 – often called recall-to-reject – 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 et al., 1998). This decrease implies that damage to medial temporal regions such as the hippocampus (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. Advanced aging also can increase susceptibility to the DRM illusion, especially in those older adults with poor frontal-lobe functioning (Butler et al., 2004). This effect has been attributed to a breakdown in the ability to monitor memory, coupled with a spared ability to process semantic associations. Considered as a whole, the data from these different populations nicely illustrate the opposing influences of activation/similarity and editing processes.

Developmental patterns also provide a unique window into these opposing influences. Young children have underdeveloped frontal lobes, and in general, they tend to have difficulty monitoring memory and avoiding false-recognition errors in many memory tasks (e.g., Moore et al., 2020). However, in the DRM task, young children tend to be less susceptible to false memories compared to older children and adults. This effect has been attributed to underdeveloped semantic processing in young children, such as difficulties in connecting and comprehending the semantic associations in DRM lists (e.g., Brainerd et al., 2018). Analogous to patients with damage to medial temporal lobes, children are less likely to “get the gist” than adults, leading to reduced relatedness effects on false memory.

Data from neuroimaging techniques, such as fMRI or EEG, have provided further insights into the neural mechanisms of the DRM effect. Unlike lesion studies, these techniques can separate neural involvement at encoding from retrieval, although findings tend to be highly specific to nuances of the experimental design and analytical techniques. Despite these challenges, a meta-analysis of 34 fMRI studies using the DRM task and related tasks by Kurkela and Dennis (2016) identified two persistent findings that deserve

highlighting. The first finding was that brain regions typically associated with the meaningful processing of language (e.g., left prefrontal and left middle temporal gyrus) tend to be activated during the encoding phase, potentially owing to the processing of semantic associations (see McDermott et al., 2005). The second finding was elevated activity in numerous prefrontal regions during false memory retrieval, which were typically attributed to monitoring demands. When considered with the aforementioned studies of neuropsychological populations, these neuroimaging studies provide converging evidence for the distinction between activation/similarity and editing processes.

## 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 had 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 et al. (1998) and Mitchell and Johnson (2009). Finally, while intriguing new findings are coming out every day (Wang et al., 2021), 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, CA: 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., Reyna, V. F., & Holliday, R. E. (2018). Developmental reversals in false memory: Development is complementary, not compensatory. *Developmental Psychology*, 54, 1773–1784.
- 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.
- Butler, K. M., McDaniel, M. A., Dornburg, C. C., Price, A. L., & Roediger, H. L., III. (2004). Age differences in veridical and false recall are not inevitable: The role of frontal lobe function. *Psychonomic Bulletin & Review*, 11, 921–925.
- Carneiro, P. & Fernandez, A. (2013). Retrieval dynamics in free recall: Revelations from identifiability manipulations. *Psychonomic Bulletin and Review*, 20, 488–495.
- Chang, M., & Brainerd, C. J. (2021). Semantic and phonological false memory: A review of theory and data. *Journal of Memory and Language*, 119, article 104210.
- 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.
- Doss, M. K., Bluestone, M. R., & Gallo, D. A. (2016). Two mechanisms of constructive recollection: Perceptual recombination and conceptual fluency. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 42, 1747–1758.
- Finley, J. S., Sungkhasette, V. W., Balota, D. A., & Roediger, H. L. (2017). Relative contributions of semantic and phonological associates to over-additive false recall in hybrid DRM lists. *Journal of Memory and Language*, 93, 154–168.
- 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. (2013). Retrieval expectations affect false recollection: Insights from a criterial recollection task. *Current Directions in Psychological Science*, 22, 316–323.
- 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.
- 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, and 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 origin 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.
- Kurkela, K. A., & Dennis, N. A. (2016). Event-related fMRI studies of false memory: An activation likelihood estimation meta-analysis. *Neuropsychologia, 81*, 149–167.
- Lampinen, J. M., Neuschatz, J. S., & Payne, D. G. (1999). Source attributions and false memories: A test of the demand characteristics account. *Psychonomic Bulletin & Review, 6*, 130–135.
- Luce, P. A. (1987). The neighborhood activation model of auditory word recognition. *Journal of the Acoustical Society of America, 81*, S1.
- 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. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition, 34*(2), 261–267.
- McDermott, K. B. (2021). Practicing retrieval facilitates learning. *Annual Review of Psychology, 72*, 609–633.
- 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., Hitchison, 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 Language, 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.
- Moore, K. N., Lampinen, J. M., Bridges, A. J., & Gallo, D. A. (2020). Developmental trends in children's use of different monitoring processes to avoid false memories. *Cognitive Development, 55*, 100911.
- 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.
- Pardilla-Delgado, E., & Payne, J. (2017). The Deese-Roediger-McDermott (DRM) task: A simple cognitive paradigm to investigate false memories in the laboratory. *Journal of Visualized Experiments, 119*, 54793.
- 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.

- 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, UK: 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., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362, 773–786.
- 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., 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.
- Smith, R. E., & Hunt, R. R. (2020). When do pictures reduce false memory? *Memory & Cognition*, 48, 623–644.
- Sommers, M. S., & Lewis, B. P. (1999). Who really lives next door: Creating false memories with phonological neighbors. *Journal of Memory and Language*, 40, 83–108.
- Stadler, M. A., Roediger, H. L., & 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.
- Umanath, S., & Coane, J. H. (2020). Face validity of remembering and knowing: Empirical consensus and disagreement between participants and researchers. *Perspectives on Psychological Science*, 15(6), 1400–1422.
- Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology*, 70, 122–129.
- Wang, J., Otgaar, H., Santtila, P., Shen, X., & Zhou, C. (2021). How culture shapes collective false memory. *Journal of Applied Research in Memory and Cognition*, 10, 24–32.
- Watson, J. M., Balota, D. A., & Roediger, H. L. (2003). Creating false memories with hybrid lists of semantic and phonological associates: Over-additive false memories produced by converging associative networks. *Journal of Memory and Language*, 49, 95–118.

## 26 Misinformation effect

*Emma PeConga, Jacqueline E. Pickrell, Daniel M. Bernstein,  
and Elizabeth F. Loftus*

Our memory is an amazing entity with the ability to provide us with the continuous sense of *self* described by William James (James, 1890). 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. But memory involves a reconstructive process: It is vulnerable to interference from other experiences, especially from experiences that occur after a to-be-remembered event. Thus, memory is subject to distortion. The study of memory distortion has historically focused on the notion of “interference”, the idea that we forget some experiences because other memories in long-term memory impair our ability to retrieve the target memory (McGeoch, 1932). In this chapter, we describe contemporary studies of memory distortion, focusing on the “misinformation effect”.

The misinformation effect refers to the tendency for post-event misleading information to reduce memory accuracy for the original event. In this chapter, we discuss various laboratory techniques that demonstrate the misinformation effect. We begin our discussion with a brief overview of the basic processes of memory, such as encoding and retrieval, and the ways in which errors in these steps can lead to forgetting. Next we describe the history of studies demonstrating the misinformation effect and how the science examining this effect has grown over time.

We then discuss the moderators of the misinformation effect, including who might be susceptible to misinformation or what kind of post-event information is necessary to distort a memory. 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 then consider the myriad positive and negative consequences of creating a false memory, and provide a simple methodology for demonstrating the misinformation effect in a classroom.

Finally, we discuss unanswered questions regarding the misinformation effect, such as potential underlying mechanisms, what the misinformation effect says about memory’s permanence and the difficulty in distinguishing true from false memories. We point to several research areas, including nascent work in the neural mechanisms of the misinformation effect, as a means to inform these questions.

### **Memory primer: encoding, storage, and retrieval**

Memory consists primarily of information encoded or 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 influences 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 first 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 stronger 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 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 can involve elaborations, omissions, or misperceptions, thus resulting in an incomplete or distorted memory. For instance, if you tell the same story to different friends, you may choose to focus on or ignore certain details and each friend may ask about different aspects of the memory. 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 schematic memory. Schema is an organized pattern of thought about some aspect of the world, such as events, situations, objects, or groups 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. 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 re-encode 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 recall something, we usually remember it 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 promoted it to the public

as well as to the scientific community. 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. Studies have since been unable to show that memories leave a lasting trace in the brain. Nevertheless, the belief that memory was permanently and physically stored remained strongly held by the scientific community and the public.

Given that history, when Elizabeth Loftus and her colleagues (Loftus, 1975; Loftus et al., 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 24). 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 were later 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 participants' 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, and half 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 the basis for a classroom demonstration given in Text box 26.1. In sum, these studies demonstrate that misleading post-event information affects what people erroneously report about the past.

### Text box 26.1 Misinformation study

As a room 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  $\chi^2$ -test can be used (or a binomial test, if expected frequencies are below 10).

## Method

### Materials and procedure

A series of eight 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://osf.io/24egy/?view\\_only=e608f86378f04216b7a7c0cc93ffbb92](https://osf.io/24egy/?view_only=e608f86378f04216b7a7c0cc93ffbb92).

The set includes the following eight slides (plus eight additional distractor slides):

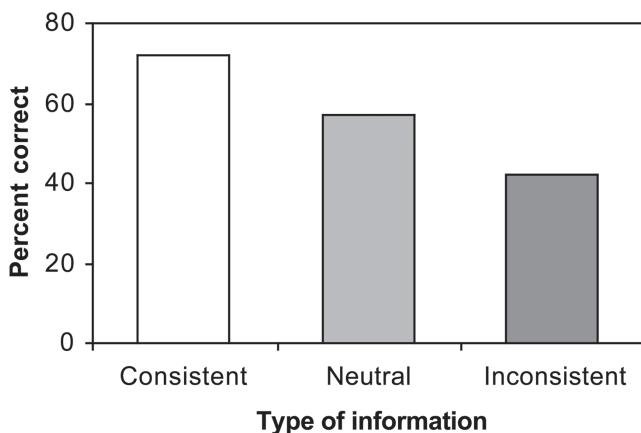
- A black car approaching on a neighborhood street.
- The black car passing a red car driving in the opposite direction.
- The black car approaching an intersection.
- The black car stopped at a stop sign with a pedestrian on the right.
- A female pedestrian, walking her dogs, crossing the street in front of the car.
- The pedestrian lying on the road in front of the black car.
- The driver of the black car gets out of the car.
- The driver of the black car speaks 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



*Figure 26.1* 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.

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 eight 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 26.1) 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 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 models of the misinformation effect*

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 hypotheses 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 questioned the procedure used in these studies. They argued that the acceptance of misinformation occurred because:

- Participants never encoded the original 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 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:

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 & Zaragoza, 1985, p. 7)

Zaragoza and colleagues (1987) tested the modified and original procedures with recall rather than recognition, as recall is more difficult as it involves fewer memory cues. That study supported their belief that differences between the misled and control conditions in the original procedure were due to factors other than memory-impairment, including task demand or other methodological effects. 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 methodological effects have demonstrated memory-impairment effects (see Ayers & Reder, 1998, for a review). Moreover, the modified test may be insensitive to memory impairment. Specifically, when overall memory performance is low, the effect may need the additional strength of the exposure to the misleading information once again at test. Studies using false images and videos (Wright et al., 2013) supported the impact of repetition of the post-event

information on memory strength. Moreover, 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 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 et al., 2011). In other words, a misinformation response can arise for many 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.

### ***Moderators 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. In fact, recent studies show that when the misinformation is detected (i.e., the participant remembers the difference between the original event and the post-event) memory accuracy actually increases (Putnam et al., 2017). Third, the more ignorant one is of the potentially misleading effects of post-event information, the more susceptible one will be to the misinformation effect (Szpitak et al., 2021). Finally, manipulations to the content of the information and individual differences in personality and mood of the participant may impact susceptibility to misinformation.

Who is susceptible to the misinformation effect? One somewhat unusual misinformation study provides insight into just *who* is most likely to accept misinformation

(Loftus et al., 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 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 et al., 1987), the youngest participants showed large misinformation effects. Additionally, the elderly showed large misinformation effects, 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 24). Such memory distortion can be seen even in “flashbulb memories”, memories named for their highly vivid, emotional, and meaningful nature (e.g., recollections for the circumstances under which they first learned about an upsetting public event like the assassination of a president). 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 et al., 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 5 (Loftus & Pickrell, 1995). Likewise, Hyman and colleagues (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 et al., 1999).

Another form of suggestion used to inflate confidence in childhood events involves imagining an event in detail that never occurred. For example, Garry and colleagues (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 increased participants' confidence that they experienced this event in their childhood. Additional research on "imagination inflation" has found that perceptual elaboration may play an important role in individual differences in false memory creation across other types of events as well (e.g., a burglary; see Drivdahl & Zaragoza, 2001).

As noted above, both "flashbulb memories" and entire memories of childhood events have been shown to be malleable. At a greater extreme, even the most emotionally salient events of one's life may be susceptible to misinformation (e.g., difficult medical examinations and childhood sexual assault; Otgaar et al., 2010, 2017). Further, the likelihood that potentially traumatic or highly stressful life memories are manipulated might even be increased by an individual's distress at the time or later. In one example, Lommen and colleagues (2013) explored susceptibility to long-term misinformation effects in soldiers deployed to Afghanistan. In this study, soldiers were interviewed two months after deployment regarding deployment-related stressors. Immediately after the interview, participants were given subtle misinformation about a missile attack at the base on New Year's Eve. After seven months, the participants were asked to complete a questionnaire about events that they had experienced on deployment, which included an item about that fake attack. Analyses revealed that a combination of high arousal and more stressors on deployment were related to higher endorsement of experiencing the implanted made-up attack. Of course, the conclusions of studies examining the relationship between distress and susceptibility to misinformation should be drawn with caution, because common misinformation tasks may only be rough approximations of real-world scenarios.

One proposed mediator of the relationship between increased distress during an event or when it is remembered is that emotional suppression affects the ability to remember highly distressing personal events. Moore and Zoellner (2012) examined the effects of expressive suppression (i.e., concealing visible displays of emotion), experiential suppression (i.e., suppressing subjective emotional experiences), and controls on memory accuracy and distortion. In this study, trauma-exposed individuals with PTSD, without PTSD, and psychologically healthy controls were shown a trauma-related film (e.g., regarding gender-based violence), a misinformation narrative, and finally tested. Results showed that expressive and experiential suppression led to poorer memory accuracy and memory distortion in participants with and without PTSD. Taken together, studies show that numerous factors may impact susceptibility to the misinformation effect, including the circumstances surrounding and the nature of the misinformation, as well as individual factors like age and distress. These moderators make it possible for a large range of events to be susceptible to misinformation.

### ***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 2 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? Given the evidence that memories of traumatic events

are both malleable and implantable, the consequences of the misinformation effect may negatively affect long-term mental health (Otgaar et al., 2019).

At the same time, the misinformation effect and its implications for the reconstructive and malleable nature of memory creates opportunities for improving individuals' lives. For instance, new research illustrates just how the malleability of memory in general, and the generation of false memories specifically, may be used in behavior modification therapy. Using an imagination paradigm, Clifasefi and colleagues (2013) successfully planted a memory of participants becoming ill after consuming a specific alcohol before the age of 16 years. After this false memory was planted, there was a decrease in participants' self-reported preference ratings for that specific alcohol. 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 we move toward a desire for healthier living.

Further, given that we can manipulate positive events resulting in positive consequences, there may be therapeutic implications to memories for negative events. Specifically, changes in memory for negative events might help us alter negative beliefs and associated negative behaviors. In fact, several evidence-based treatments for post-traumatic stress disorder (PTSD) accomplish therapeutic gains through restructuring maladaptive trauma memory related beliefs. For instance, Cognitive Processing Therapy (Resick & Schnicke, 1992) challenges individuals' potential inflated negativity or metacommentary about their traumatic memories regarding their subjective safety, blameworthiness, or sense of control.

Previous research on the misinformation effect primarily focused on the impact misinformation has in the legal arena, eyewitness testimony, and false memories created during therapy. We are aware of misinformation in our daily lives, but until recently, the consequences seemed manageable. From politics to science, information is now being communicated through social media at a rate unseen in previous decades. Disseminating information through social media includes getting the most "likes", the most "shares", or the most "followers". Sexy, salacious headlines are used to capture attention. The article itself is frequently only skimmed or skipped altogether, and the salacious headline is all that gets stored in memory. This new communication milieu results in misinformation and disinformation becoming a more common phenomenon (see Chapter 20 on fake news).

As we experience the rates of communication increasing, we see the increase and spread of misinformation. Science and the scientific process take time. They involve gathering empirical evidence, and then filtering and interpreting that evidence. The peer-review process serves as guardian for scientific accuracy. However, the speed of social media has little patience. Today we see that preprint (not peer-reviewed) findings are often taken as truth when, in fact, that has not yet been confirmed (West & Bergstrom, 2021). Thus, while we have easier access to that information, our ability to learn and check facts for accuracy at times seems unmanageable due to the sheer volume of information. This spread of misinformation is particularly concerning given that studies have shown that when individuals are confronted with any form of misinformation, correcting such errors is challenging (i.e., the continued-influence effect; Lewandowsky et al., 2012).

### **Possible mechanisms**

What is not clear from the work we have described is the underlying mechanism responsible for memory distortion. One possible mechanism that might explain memory

distortion after different forms of misinformation is that of familiarity (Garry et al., 1996). Jacoby and colleagues (e.g., Jacoby et al., 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 11, 14, and 15 in this volume).

Familiarity misattribution may help explain why people accept misinformation and why they become more confident about 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 have processed it otherwise. 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. While these may be promising leads as to what cognitive mechanisms underly the misinformation effect, more research should explore key mediators.

### **Distinguishing true from false memories**

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 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). In some studies participants used more words when describing their true memories, and more highly rated the clarity and confidence of their true memories than they rated the clarity of their false memories. However, many studies have shown that false memories can also be expressed with a lot of detail and confidence, and even emotion (Otgaar, 2017).

Roediger and McDermott (1995) created false memories for words not presented in lists (see Chapter 25). 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 and colleagues (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. 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. Additionally, 40% of participants recalled the real memory from a participant perspective, or "their own eyes" (p. 28). The remaining 60% of participants viewed the real memory from the observer perspective (i.e., like 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, a finding backed by other studies (Blandon-Gitlin, 2009).

Heaps and Nash (2001) examined differences between true and false memories. These researchers used a variation of the Loftus and Pickrell paradigm (see Text box 26.1) to plant false memories in participants. They used information from relatives to suggest to participants that they had undergone certain experiences. On first pass, the true memories appeared different from the false memories, because they were rated as being more important, detailed, emotionally intense 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. However, even after controlling for rehearsal frequencies, false memories contained less information about any consequences of the event. Thus, differences in clarity and emotional valence of true and false memories may be tied to rehearsal effects of true memories, while a key differentiator may be in their contextualization. That conclusion has been supported by more recent research on rehearsal effects (Foster et al., 2012).

Despite the relevance of the misinformation effect to testimony, the move into the legal arena often lags behind 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. Further, the existence of co-witnesses may further increase confidence in false testimony (Jack et al., 2014). Currently, external corroboration remains the only reliable way to determine the veracity of a memory.

## **Neuroscientific evidence**

Much is known about the behavioral aspects of the misinformation effect. Relatively little is known about the neural underpinnings of this effect. Most work on the cognitive neuroscience of the misinformation effect involves functional magnetic resonance imaging (fMRI) and event-related potentials (ERPs). Researchers have used these techniques to try to distinguish true from false memories (e.g., Curran et al., 2001; Karanian et al., 2020; see Schacter & Slotnick, 2004, and Yu et al., 2019). After decades of studying behavioral aspects of the misinformation effect, Loftus (2005b) deplored the paucity of physiological measures of this effect. Sadly, little has changed since 2005. Although neuroscientific work on the misinformation effect continues, we are far from having a robust understanding of the effect’s neurophysiology. We briefly discuss fMRI and ERP studies in turn.

Okado and Stark (2005) were the first to use neuroimaging techniques to investigate the misinformation effect. They used detailed vignettes 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. Participants viewed the vignettes while in the fMRI scanner. A short time later, participants viewed the vignettes again. Unbeknownst to them, several changes had been made 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. Two days later, 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 correlated with later

reporting of true and false memories on the recognition test (see also Baym & Gonsalves, 2010; Stark et al., 2010).

Turning to work involving ERPs, there is evidence that the Late Positive Component (LPC) is more positive for true than false memories (Curran, 2000). The LPC is believed to underlie recollection-based processes during recognition memory. Using ERPs during the recognition portion of a misinformation paradigm, Kiat and Belli (2017) showed that the LPC was more positive for true than false memories. Others have failed to replicate this effect (Volz et al., 2019). A few common problems in cognitive neuroscience, which also plague neuroscientific studies of the misinformation effect, include the use of small samples and the lack of direct replication studies. Another problem is overlapping brain activation patterns for true and false memories that make it particularly difficult to extract neural “signatures” for false memories.

In the misinformation studies, participants are mostly trying to be accurate. But what would happen in studies in which people are deliberately lying? What if they lied about the details of an event? Would the neural signals detect such activity (Meek et al., 2012)? Could individuals lie about an event when they do not know all the details? In one study, participants observed a slide show depicting a crime, 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. A recent meta-analysis of fMRI data revealed that deception and false memory tasks activate similar frontoparietal regions, including the left superior frontal gyrus. However, the two types of tasks also activate different regions, prompting the authors to conclude that fMRI might prove useful in distinguishing between deception and false memory (Yu et al., 2019). Despite some advances, at this point, there is simply too little neuroscientific work on the misinformation effect to draw any firm conclusions about its neurophysiology.

## 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 25). 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). 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.

Along with technological advances in brain imaging techniques comes the potential to distinguish true from false memories at a physiological level. However, it is important to use caution when drawing conclusions from these results. The data from fMRI and ERP studies reflect an average of each individual participant's results averaged across participants (see Frenda et al., 2011). While providing important information about how false memories evolve, averages do not tell us whether a particular memory is true (see Bernstein & Loftus, 2009). Certainly, more work is needed on the neurophysiology of the misinformation effect (see Schacter & Loftus, 2013, for possible links between the misinformation effect and reconsolidation).

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, given the growing literature base on potential negative consequences of the misinformation (Otgaar, 2019). 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 (West & Bergstrom, 2021).

## **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 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 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 consequences of memory distortions have important implications for several applied problems.
- At this point, there is simply too little neuroscientific work on the misinformation effect to draw any firm conclusion.
- We still 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**

- 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.
- Blandón-Gitlin, I., Pezdek, K., Lindsay, D. S., & Hagen, L. (2009). Criteria-based content analysis of true and suggested accounts of events. *Applied Cognitive Psychology, 23*(7), 901–917.
- 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.
- Ceci, S. J., Ross, D. F., & Toglia, M. P. (1987). Suggestibility of children's memory: Psycholegal implications. *Journal of Experimental Psychology: General, 116*(1), 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. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition, 28*(6), 923–938.
- 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.
- Foster, J. L., Huthwaite, T., Yesberg, J. A., Garry, M., & Loftus, E. F. (2012). Repetition, not number of sources, increases both susceptibility to misinformation and confidence in the accuracy of eyewitnesses. *Acta Psychologica, 139*(2), 320–326.
- 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.
- 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.
- 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.
- Jack, F., Zydervelt, S., & Zajac, R. (2014). Are co-witnesses special? Comparing the influence of co-witness and interviewer misinformation on eyewitness reports. *Memory, 22*(3), 243–255.
- 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.
- James, W. (1890). *The principles of psychology*. New York: Henry Holt.
- Karanian, J. M., Rabb, N., Wulff, A. N., Torrance, M. G., Thomas, A. K., & Race, E. (2020). Protecting memory from misinformation: Warnings modulate cortical reinstatement during memory retrieval. *Proceedings of the National Academy of Sciences, 117*(37), 22771–22779.
- 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.
- Kiat, J. E., & Belli, R. F. (2017). An exploratory high-density EEG investigation of the misinformation effect: Attentional and recollective differences between true and false perceptual memories. *Neurobiology of Learning and Memory, 141*, 199–208.

- Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3), 106–131.
- Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, 7, 560–572.
- 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.
- Lommen, M. J., Engelhard, I. M., & van den Hout, M. A. (2013). Susceptibility to long-term misinformation effect outside of the laboratory. *European Journal of Psychotraumatology*, 4(1), 19864.
- 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.
- McGeoch, J. A. (1932). Forgetting and the law of disuse. *Psychological Review*, 39(4), 352–370.
- 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(2), 145–160.
- Moore, S. A., & Zoellner, L. A. (2012). The effects of expressive and experiential suppression on memory accuracy and memory distortion in women with and without PTSD. *Journal of Experimental Psychopathology*, 3(3), 368–392.
- 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.
- Otgaar, H., Candel, I., Scoboria, A., & Merckelbach, H. (2010). Script knowledge enhances the development of children’s false memories. *Acta Psychologica*, 133(1), 57–63.
- Otgaar, H., Howe, M. L., Patihis, L., Merckelbach, H., Lynn, S. J., Lilienfeld, S. O., & Loftus, E. F. (2019). The return of the repressed: The persistent and problematic claims of long-forgotten trauma. *Perspectives on Psychological Science*, 14(6), 1072–1095.
- Otgaar, H., Merckelbach, H., Jelicic, M., & Smeets, T. (2017). The potential for false memories is bigger than what Brewin and Andrews suggest. *Applied Cognitive Psychology*, 31(1), 24–25.
- 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.
- Putnam, A. L., Sungkhasettee, V. W., & Roediger, H. L., III. (2017). When misinformation improves memory: The effects of recollecting change. *Psychological Science*, 28(1), 36–46.
- Resick, P. A., & Schnicke, M. K. (1992). Cognitive processing therapy for sexual assault victims. *Journal of Consulting and Clinical Psychology*, 60(5), 748–756.
- 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.
- Schacter, D. L., & Slotnick, S. D. (2004). The cognitive neuroscience of memory distortion. *Neuron*, 44, 149–160.
- Stark, C. E. L., Okado, Y., & Loftus, E. F. (2010). Imaging the reconstruction of true and false memories using sensory reactivation and the misinformation paradigm. *Learning & Memory*, 17, 485–488.
- Szpitak, M., Woltmann, A., Polczyk, R., & Kekus, M. (2021). Memory training as a method for reducing the misinformation effect. *Current Psychology*, 40, 5410–5419. <https://doi.org/10.1007/s12144-019-00490-9>
- Tulving, E., & Thompson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80, 352–373.
- Volz, K., Stark, R., Vaitl, D., & Ambach, W. (2019). Event-related potentials differ between true and false memories in the misinformation paradigm. *International Journal of Psychophysiology*, 135, 95–105.
- West, J. D., & Bergstrom, C. T. (2021). Misinformation in and about science. *Proceedings of the National Academy of Sciences*, 118(15), e1912444117.
- 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.
- Yu, J., Tao, Q., Zhang, R., Chan, C. C. H., Lee, T. M. C. (2019). Can fMRI discriminate between deception and false memory? A meta-analytic comparison between deception and false memory studies. *Neuroscience and Biobehavioral Reviews*, 104, 43–55.
- 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(1), 36–44.

## 27 Hindsight bias

*Rüdiger F. Pohl and Edgar Erdfelder*

Whenever in hindsight, we tend to exaggerate what we had known 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” and has been observed in numerous studies up to date. Hindsight bias was the focus of two meta-analyses (Christensen-Szalanski & Willham, 1991; Guilbault et al., 2004) and several overviews (e.g., Bernstein et al., 2016; Hawkins & Hastie, 1990; Pezzo, 2011; Pohl & Erdfelder, 2019; 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), which is 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 so that they are left to reconstruct their earlier given estimate, however, 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 first describe the phenomenon of hindsight bias in more detail and how it is typically assessed. Next, we present an overview of empirical findings and 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.

### **Text box 27.1 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 et al., 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 et al., 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.

## Assessment and definition

As the examples in Text box 27.1 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 bias, and the definition of hindsight bias.

### *Experimental designs*

Most importantly, two different general experimental procedures were employed. In the *memory* design (Example 1 in Text box 27.1), 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 correct judgment?” 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 13).

### *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). We distinguish here more traditional from model-based measures that both contain measures of accuracy and distortion.

#### *Traditional measures*

A widely used measure is Pohl’s (1992) difference score (see also Fischer & Budescu, 1995)

$$\Delta HB := |OJ - CJ| - |ROJ - CJ| \quad (1)$$

which would give us  $\Delta HB = |1710 - 1756| - |1740 - 1756| = 30$  in the Mozart example as the difference of absolute deviations from the correct judgment. Positive values indicate hindsight bias in the sense that, compared to *OJ*, *ROJ* is closer to *CJ*. In general, the larger  $\Delta HB$ , 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)  $\Delta z$  measure

$$\Delta z = |z_{OJ} - z_{CJ}| - |z_{ROJ} - z_{CJ}|, \quad (2)$$

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  $\Delta z$  measure has a number of advantages compared to previously used measures (see, e.g., Hell et al., 1988), making it the measure of choice whenever an overall measure of hindsight bias is needed (cf. Pohl, 2007).

Overall measures of hindsight bias 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.

### *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,  $ROJ$ 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  $c$ . 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  $OJ$ ). To illustrate, for control items, retrieval leads to a perfect recollection ( $ROJ = OJ$ ) with probability  $r_C$  or to a recollection failure (and thus an unbiased reconstruction) with probability  $1 - r_C$ . For experimental items, retrieval leads to a perfect recollection with probability  $r_E$  or to a recollection failure with probability  $1 - r_E$ . This failure could lead to a biased reconstruction with probability  $b$  or to an unbiased one with probability  $1 - b$ . In the former case, the biased reconstruction could lead to a source confusion ( $ROJ = CJ$ ) with probability  $c$  or not (with probability  $1 - c$ ).

Parameters of the HB13-model are then estimated using the maximum likelihood method (see Erdfelder & Buchner, 1998, for the details). To test how well 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 for testing parameter differences, namely “multiTree” (Moshagen, 2010).<sup>1</sup>

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 et al., 2015; Groß & Pachur, 2019).

### **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  $\Delta z$  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**

Hindsight bias is a robust and widespread phenomenon (see Bernstein et al., 2016; Hawkins & Hastie, 1990; Pezzo, 2011; Pohl, 2007; Pohl & Erdfelder, 2019; Roese & Vohs, 2012, for overviews). It has been observed with many different experimental procedures in many different content domains. The materials could be assorted into three categories (and tasks): (1) judge the truth of assertions (like “Dallas is the capital of Texas”); (2) rate the likelihood of several potential outcomes of an event or episode (like a political election); and (3) generate numerical estimates in response to knowledge questions (like “How many books did Agatha Christie write?”).

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 that it did not occur. (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. 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).

### ***Variants of hindsight bias***

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 et al., 2012, 2018; Higham et al., 2017), or even gustatory materials (e.g., Pohl et al., 2003b). As a special case of visual materials, Giroux et al. (2022) reported hindsight bias for emotional faces. Other studies found hindsight bias in the recall of metacognitive judgments (Ackermann et al., 2020; Zimdahl & Undorf, 2021). And even conjectures (i.e., no explicit outcome knowledge) were sufficient to elicit hindsight bias (von der Beck et al., 2019).

### ***Individual differences***

Several researchers have addressed the question of individual differences. For example, Musch (2003) found that hindsight bias (in the hypothetical design) was positively related to field dependence, to the tendency for a favorable self-presentation, to the participants' conscientiousness, and to their need for predictability and control. 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 so far 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.

More recently, Kausel et al. (2013) reported a positive relation with individuals' narcissism, and Lamberty et al. (2018) with their "conspiracy mentality" (cf. Chapter 20). Hindsight bias was also positively related to states of dysphoria and induced negative mood (Groß & Bayen, 2017) as well as depression (Groß et al., 2017).

Another individual difference 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 (Guilbaut et al., 2004) or even found positive effects of expertise (Musch & Wagner, 2007). Thus, the role of expertise remains unclear and presumably depends on further factors.

Another area of research is age-related changes in hindsight bias. 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 (Bernstein et al., 2011; Pohl et al., 2018). In their meta-analysis across nine studies comparing younger and older adults, Groß and Pachur (2019) found a small and similar recollection bias in both age groups, whereas older adults showed a larger reconstruction bias than young adults did. Some studies (e.g., Coolin et al., 2016; Groß & Bayen, 2015) found empirical support for explaining age-related differences in hindsight bias with corresponding differences in episodic memory and inhibitory capacities, which are both known to follow inverted U-shaped functions across the lifespan (but see Pohl et al., 2018).

### ***Debiasing attempts***

Generally, participants are not even aware of hindsight bias, let alone being able to avoid it. Many studies nevertheless tried to debias their participants, but most have failed (e.g., Pohl & Hell, 1998), thus supporting the view of a robust phenomenon. However, a few attempts succeeded. For example, Hasher et al. (1981) simply discredited the given CJs as "being wrong" and subsequently found no hindsight bias (replicated by Erdfelder & Buchner, 1998), suggesting that avoiding the bias is possible even after encoding the CJs. Accordingly, Dietvorst and Simonsohn (2019) assumed a motivational explanation of hindsight bias and argued that participants typically want to use the outcome information. In one of their studies, they successfully persuaded their participants (with a number of arguments) to not use to-be-ignored information in their judgments. Finally, Van Boekel et al. (2017) used distinctive cues in the ROJ phase (i.e., announcing recall of both the CJ and the OJ) and thereby eliminated hindsight bias.

### **Hindsight-bias experiment**

As a typical experiment, we focus here on studies using numerical almanac-questions in the memory design (e.g., Hell et al., 1988; Pohl, 1998; Pohl et al., 2003a; Pohl & Hell, 1996, Exp. 1). Text box 27.2 provides all necessary details to set up an adapted classroom demonstration of hindsight bias. In the following, we describe the main results from these studies.

#### **Text box 27.2 A classroom demonstration of hindsight bias**

##### **Method**

###### **Participants**

Setting  $\alpha$  and  $\beta$  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). 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  $\Delta z$  (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 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 (or erroneous inputs), 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 (OJs, CJs, and ROJs) are then transformed into  $z$ -scores separately for each question.

The absolute frequencies of correct recollections (i.e.,  $ROJ = OJ$ ) in the two conditions (experimental v. control) can be compared with a paired  $t$ -test (with  $df = N - 1$ ). For the remaining pairs of OJs and ROJs, the shift indices  $\Delta z$  are computed and then averaged across all items in the same condition for each participant. The resulting two mean shift indices (experimental v. 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, 1998). 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.

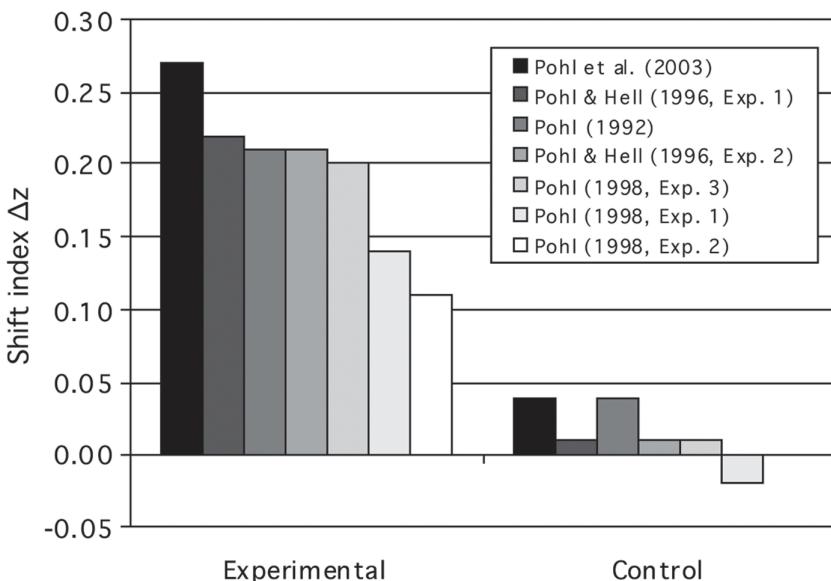
Differences between experimental and control condition are typically absent or only small when these conditions are manipulated within-subjects (see Erdfelder et al., 2007, for an overview). Only a few studies reported that knowing the CJs can impair memory for one's OJs and thus lead to recollection bias.

### *Hindsight bias*

Comparing the mean shift indices for experimental and control items yielded significant differences in all of the cited studies (see Figure 27.1). 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). On average, 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 27.1).

### *Model-based measures*

To illustrate model-based measures of hindsight bias, we refer to Erdfelder et al. (2007), who analyzed nine data sets. First, 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 two of the nine conditions. 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 was 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).



*Figure 27.1* Mean shift indices  $\Delta z$  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 ROJs were on average closer to the CJs than the OJs had been).

### Discussion

First, 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 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. In contrast, judgments in the control condition were not recalled with a systematic bias. 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 & Buchner, 1998; Erdfelder et al., 2007).

### 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 these approaches in more detail, though, we focus on a framework that distinguishes different types or components of hindsight bias.

### ***Components view***

Blank et al. (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 et al. (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 focused on separate components of hindsight bias. Roese and Vohs (2012) suggested viewing 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 focused 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; Pohl et al., 2003a):

- 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.
- Focusing on learning and efficient use of cue knowledge, the CJ may also be used to update hitherto uncertain or unknown cue knowledge (see Hoffrage et al., 2000), thus again irreversibly altering one's knowledge.

- In terms of selective activation during encoding, the CJ may activate – by way of spreading activation – knowledge that is most similar to it (cf. the explanation of anchoring effects in Chapter 13), 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.

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 focused hindsight-bias component) and the experimental procedure (like memory versus hypothetical design).

### *Computational models*

Pohl et al. (2003a) 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) 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 18). 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 (Bernstein & Harley, 2007; see also Birch et al., 2017).

Pohl (1998) found that CJs that were labeled 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. Support for Pezzo’s view can be taken from a recent study (Sleegers et al., 2021) that found a positive relation between hindsight bias and participants’ pupil size at the time of presenting the CJ, reflecting general arousal (presumably based on surprise). If sense-making then 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 et al., 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***

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 8). The second factor, self-presentational concerns, has received quite some attention, mainly because when informed about the hindsight-bias phenomenon, most lay persons 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” (Tykocynski & 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 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 et al., 2006). Giroux et al. (2016) gave an overview on “hindsight bias and law” and also discussed several debiasing strategies (see also Strohmaier et al., 2021). Louie et al. (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 in hindsight could also bias us towards a too optimistic evaluation of our prior knowledge state. If we consider ourselves more knowledgeable than we really are, 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; 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 (as exemplified in hindsight bias) is probably more than outweighed by the benefits of adaptive learning.

## 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.
- 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.
- 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.
- The phenomenon has important implications for several applied problems.

## Note

- 1 The software can be downloaded free of charge for Windows, MacOS, and Linux platforms (from [www.sowi.uni-mannheim.de/erdfelder/forschung/software/multitree/](http://www.sowi.uni-mannheim.de/erdfelder/forschung/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.

## Further reading

Hawkins and Hastie (1990) gave an excellent early overview. Christensen-Szalanski and Willham (1991) and Guilbault et al. (2004) reported instructive meta-analyses. Two special issues are devoted to hindsight bias, namely in *Memory* (Hoffrage & Pohl, 2003) and in *Social Cognition* (Blank et al., 2007), both covering theoretical, empirical, as well as applied aspects of hindsight bias. Later, Bernstein et al. (2016), Pezzo (2011), Pohl and Erdfelder (2019), and Roese and Vohs (2012) provided further comprehensive overviews. As an example for an applied context, we suggest the review paper by Giroux et al. (2016) on hindsight bias in legal settings.

## Acknowledgment

The cited studies of the authors have been supported by various grants from the Deutsche Forschungsgemeinschaft.

## References

- Ackerman, R., Bernstein, D. M., & Kumar, R. (2020). Metacognitive hindsight bias. *Memory & Cognition*, 48(5), 731–744.
- 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.
- 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., Kumar, R., Masson, M. E. J., & Levitin, D. J. (2018). Fluency misattribution and auditory hindsight bias. *Memory & Cognition*, 46(8), 1331–1343.
- 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.
- Birch, S. A. J., Brosseau-Liard, P. E., Haddock, T., & Ghrear, S. E. (2017). A “curse of knowledge” in the absence of knowledge? People misattribute fluency when judging how common knowledge is among their peers. *Cognition*, 166, 447–458.
- 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.
- Bukkszar, 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., 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.
- Coolin, A., Erdfelder, E., Bernstein, D. M., Thornton, A. E., & Thornton, W. L. (2016). Inhibitory control underlies individual differences in older adults’ hindsight bias. *Psychology and Aging*, 31(3), 224–238.
- Dietvorst, B. J., & Simonsohn, U. (2019). Intentionally “biased”: People purposely use to-be-ignored information, but can be persuaded not to. *Journal of Experimental Psychology: General*, 148(7), 1228–1238.
- 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.
- Giroux, M. E., Coburn, P. I., Harley, E. M., Connolly, D. A., & Bernstein, D. M. (2016). Hindsight bias and law. *Zeitschrift für Psychologie/Journal of Psychology*, 224(3), 190–203.
- Giroux, M. E., Hunsche, M. C., Erdfelder, E., Kumar, R., & Bernstein, D. M. (2022). Hindsight bias for emotional faces. *Emotion*. DOI: 10.1037/emo0001068.

- Groß, J., & Bayen, U. J. (2015). Hindsight bias in younger and older adults: The role of access control. *Aging, Neuropsychology, and Cognition*, 22(2), 183–200.
- Groß, J., & Bayen, U. J. (2017). Effects of dysphoria and induced negative mood on the processes underlying hindsight bias. *Cognition and Emotion*, 31(8), 1715–1724.
- Groß, J., Blank, H., & Bayen, U. J. (2017). Hindsight bias in depression. *Clinical Psychological Science*, 5(5), 771–788.
- Groß, J., & Pachur, T. (2019). Age differences in hindsight bias: A meta-analysis. *Psychology and Aging*, 34(2), 294–310.
- 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.
- Hasher, L., Attig, M. S., & Alba, J. W. (1981). I knew it all along: Or, did I? *Journal of Verbal Learning and Verbal Behavior*, 20(1), 86–96.
- 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.
- Higham, P. A., Neil, G. J., & Bernstein, D. M. (2017). Auditory hindsight bias: Fluency misattribution versus memory reconstruction. *Journal of Experimental Psychology: Human Perception and Performance*, 43(6), 1144–1159.
- 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).
- Kausel, E. E., Culbertson, S. S., Jackson, A. T., Leiva, P. I., & Reb, J. (2013). The role of narcissism and should counterfactual thinking in the hindsight bias. *Academy of Management Annual Meeting Proceedings*, 2013(1), 1491–1496.
- Lamberty, P. K., Hellmann, J. H., & Oeberst, A. (2018). The winner knew it all? Conspiracy beliefs and hindsight perspective after the 2016 US general election. *Personality and Individual Differences*, 123, 236–240.
- 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 foreseeability 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. (2007). Ways to assess hindsight bias. *Social Cognition*, 25(1), 14–31.
- Pohl, R. F., Bayen, U. J., Arnold, N., Auer, T.-S., & Martin, C. (2018). Age differences in processes underlying hindsight bias: A lifespan study. *Journal of Cognition and Development*, 19(3), 278–300.
- 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., & Erdfelder, E. (2019). Hindsight bias in political decision making. In D. P. Redlawsk (Ed.), *Oxford encyclopedia of political decision making*. Oxford: Oxford University Press.
- 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., Szczesny, 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 meta-cognitions. *Memory*, 11, 395–410.
- Sleegers, W. W. A., Proulx, T., & van Beest, I. (2021). Pupillometry and hindsight bias: Physiological arousal predicts compensatory behavior. *Social Psychological and Personality Science*, 12(7), 1146–1154.
- 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.
- Strohmaier, N., Pluut, H., den Bos, K., Adriaanse, J., & Vriesendorp, R. (2021). Hindsight bias and outcome bias in judging directors' liability and the role of free will beliefs. *Journal of Applied Social Psychology*, 51(3), 141–158.
- 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.
- Van Boekel, M., Varma, K., & Varma, S. (2017). A retrieval-based approach to eliminating hindsight bias. *Memory*, 25(3), 377–390.

- von der Beck, I., Cress, U., & Oeberst, A. (2019). Is there hindsight bias without real hindsight? Conjectures are sufficient to elicit hindsight bias. *Journal of Experimental Psychology: Applied*, 25(1), 88–99.
- Yopchick, J. E., & Kim, N. S. (2012). Hindsight bias and causal reasoning: A minimalist approach. *Cognitive Processing*, 13, 63–72.
- ZimdaHL, M. F., & Undorf, M. (2021). Hindsight bias in metamemory: Outcome knowledge influences the recollection of judgments of learning. *Memory*, 29(5), 559–572.

## 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 kinds 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 minutes]
19. What is the radius of the earth (in kilometers)? [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 centimeters)? [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 the Mount Everest (in meters)? [8,848 m]
28. In what year did the Roman emperor Nero commit suicide? [AD 68]
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 kilometers)? [2,450 km]
39. How many islands make up Hawaii? [132]
40. How long is the Panama canal (in kilometers)? [81.6 km]

# Author index

- Aaker, J. L. 72  
Abele, A. E. 278  
Abelson, R. P. 128  
Aboukoumin, G. 241  
Abramson, L. Y. 92, 108, 111, 115, 124–125, 127–130, 132, 135–136  
Acitelli, L. K. 272  
Ackerman, R. 164, 187, 311, 315, 440  
Aczel, B. 28–29  
Adams, J. K. 287–288  
Adams, P. A. 287–288  
Agnoli, F. 27–28, 36, 93  
Ahmad, S. 373  
Ajzen, I. 14  
Alba, J. W. 290, 293, 441  
Alexander, D. N. 247  
Allan, L. G. 109, 112  
Allen, J. 333  
Allen, J. L. 159  
Allison, R. I. 393  
Alloy, L. B. 92, 108, 111, 115, 124–125, 127–130, 132–133, 135–136  
Allwood, C. M. 294  
Alogna, V. K. 395  
Altarriba, J. 372–373  
Altay, S. 325–326, 332–333  
Alter, A. L. 186, 201, 266, 268–269  
Amazeen, M. A. 334  
Ammirati, R. 89  
Andreou, C. 135  
Andrews, S. O. 204  
Anes, M. D. 414  
Angle, S. T. 129  
Anglim, J. 278  
Applebaum, A. 330  
Apra, F. 35  
Ariely, D. 211  
Arif, A. 330  
Aristotle 192  
Arkes, H. R. 13, 225, 228–229, 231–233, 448  
Armstrong, W. 128  
Arndt, J. 410, 412  
Aron, A. 277  
Arp, R. 5  
Asch, S. 259, 262, 266, 268  
Ashmore, R. D. 263  
Ashton, M. C. 276  
Aslan, A. 372  
Aslett, K. 327  
Asmolov, G. 328, 330–331  
Aßfalg, A. 19, 235  
Åstebro, T. 302  
Atalay, N. B. 379  
Attanasio, J. S. 200  
Attig, M. S. 441  
Ayeroff, F. 128  
Ayers, M. S. 424–425  
Ayton, P. 196  
Back, M. D. 17, 275  
Bacon, A. M. 167  
Bacon, F. 78  
Bacon, L. P. 372  
Bago, B. 28–29, 46, 57, 164  
Bahník, Š. 216–218, 220  
Baker, M. 301  
Bakhti, R. 36  
Bakir, V. 325  
Ball, L. J. 84, 143, 159, 162  
Ballard, I. C. 252  
Balliet, D. 281  
Balota, D. A. 409, 411–412  
Balzan, R. P. 115  
Bandura, A. 124  
Banks, A. P. 57  
Banks, D. M. 343  
Bar-Hillel, M. 50–51, 192  
Bar-Tal, D. 248  
Barbaranelli, C. 124  
Barber, B. C. 129  
Barber, B. M. 301  
Barberia, I. 113, 119  
Barbey, A. K. 44, 46, 52, 58  
Barbieri-Hermitté, P. 214  
Barbone, S. 5  
Barefoot, J. C. 343

- Barlas, S. 300  
 Baron, J. 5  
 Barr, D. J. 212  
 Barr, N. 36, 56, 327  
 Barron, F. 250  
 Barston, J. L. 158–159  
 Bartlett, F. C. 12, 420  
 Bassok, M. 80–81  
 Bauer, R. M. 251  
 Baumeister, R. F. 345, 376  
 Bäuml, K.-H. T. 18, 372, 395, 398  
 Baxter, B. 243  
 Bayen, U. J. 440–441, 443, 446  
 Baym, C. L. 431  
 Beck, A. T. 341  
 Beck, J. R. 31  
 Becker, M. L. 133  
 Beer, A. 276–277  
 Bell, R. 376, 378–379  
 Belli, R. F. 392, 431  
 Ben-David, I. 302  
 Bénabou, R. 299  
 Bender, M. 7  
 Beneteau, J. L. 111  
 Benevenuto, F. 335  
 Benjamin, A. S. 176, 187, 309, 311, 413  
 Benoît, J. P. 287, 298  
 Benson, B. 4, 13–15  
 Benvenuti, M. F. L. 130, 132  
 Berg, C. A. 254  
 Berger, J. 325  
 Bergquist, M. 344  
 Bergstrom, C. T. 428, 432  
 Berk, M. 328–331  
 Bernoster, I. 302  
 Bernstein, D. M. 429, 432, 436, 439–440, 444, 446  
 Berscheid, E. 262–264, 266  
 Besson, T. 266  
 Bhattacharyya, G. K. 204  
 Biais, B. 301, 448  
 Biddle, J. E. 267  
 Biesanz, J. C. 272, 275–278  
 Binder, N. 241  
 Biner, P. M. 129–130  
 Birch, S. A. J. 446  
 Birnbaum, M. H. 46, 50  
 Bishop, J. 413  
 Biswas, A. 212  
 Bizarro, L. 130, 132  
 Bjärehed, J. 341  
 Bjork, R. A. 310, 317–319  
 Bjorklund, D. F. 300  
 Björkman, M. 297  
 Black, P. M. 414  
 Blanco, F. 110–111, 114–115, 118  
 Blandon-Gitlin, I. 429  
 Blank, H. 436, 440, 445–446  
 Bless, H. 183, 185–186  
 Blitz-Miller, T. 134  
 Block, R. A. 295  
 Bluestone, M. R. 412  
 Blunt, J. R. 379  
 Boca, S. 242  
 Bodnar, G. E. 408  
 Boehm, L. E. 225, 228–229, 231, 233, 235  
 Boer, D. 278  
 Bolger, F. 293, 295  
 Bond, G. D. 350  
 Bond, R. M. 328  
 Bonneau, R. 327  
 Bonner, S. E. 368  
 Borah, P. 62  
 Borgida, E. 31  
 Bornstein, B. H. 212  
 Bornstein, M. H. 386, 391–392  
 Bornstein, R. F. 241–255  
 Bottoms, H. C. 367  
 Botvinick, M. M. 135  
 Bouts, P. 130  
 Bowden-Jones, H. 134  
 boyd, d. 325, 329  
 Boyer-Kassem, T. 33  
 Boyer, P. 12, 17  
 Brady, W. J. 325, 329  
 Braga, J. N. 179  
 Brainerd, C. J. 37, 71–72, 409–410, 412, 414  
 Bramley, N. R. 38  
 Brandimonte, M. A. 395, 398  
 Brashier, N. M. 231, 236, 327  
 Bratslavsky, E. 376  
 Braun, K. A. 392  
 Bredart, S. 361, 365–366  
 Bregman, N. J. 393  
 Brekke, N. 214  
 Brewer, N. T. 214, 216–217  
 Bridges, A. J. 414  
 Brier, G. W. 288  
 Brigham, J. C. 395  
 Brighton, 181  
 Briley, D. A. 134  
 Bröder, A. 309, 316, 379  
 Brody, N. 246  
 Brooks, R. A. 350  
 Brown, D. 134  
 Brown, J. D. 133, 344, 346–347  
 Brown, N. R. 187  
 Bruce, M. 5  
 Bruchmann, K. 212  
 Bruno, D. 379  
 Brunswik, E. 295, 298  
 Brydges, C. 327  
 Buchanan, E. M. 372  
 Buchner, A. 333, 376–379, 438, 442–445  
 Budescu, D. V. 296–297, 437

- Budson, A. E. 414  
 Buehler, R. 287  
 Buehner, M. J. 112  
 Bugg, J. M. 372–373  
 Bukszar, E. W. 448  
 Burns, D. J. 376–377, 379–380  
 Burns, S. A. 376, 379  
 Burton, S. 212  
 Busemeyer, J. R. 32  
 Buss, D. 269  
 Butler, A. C. 377  
 Butler, K. M. 413–414  
 Byrne, R. M. J. 154, 159, 167  
 Bystranowski, P. 211  
 Cabral, K. 387  
 Cacciapaglia, H. 387  
 Cacioppo, J. T. 233  
 Cadinu, M. R. 274  
 Cain, D. M. 293, 297  
 Calderwood, K. 134  
 Calio, F. 234–235  
 Calvillo, D. P. 225, 231, 235  
 Camerer, C. 301–302  
 Cameron, C. D. 329  
 Campbell, J. D. 440, 447–448  
 Campbell, J. I. D. 145  
 Campbell, W. K. 347  
 Cantor, A. D. 367  
 Caprara, G.V. 124  
 Cara, F. 83, 149  
 Cardwell, B. A. 312  
 Carlson, B. W. 27–28, 31, 34–35  
 Carmichael, L. 386–390  
 Carneiro, P. 414  
 Carpenter, S. K. 311  
 Carroll, M. 397  
 Carstensen, L. L. 343  
 Caruso, E. G. 243, 254  
 Carver, C. S. 348  
 Castel, A. D. 311, 315, 317–318  
 Caverni, J.-P. 5, 12, 16  
 Ceci, S. J. 426  
 Cervone, D. 211  
 Cesarin, D. 294  
 Chaiken, S. 62, 65  
 Chan, J. C. K. 425  
 Chang, M. 412  
 Chapman, G. B. 211–212, 214, 217–218  
 Chapman, J. P. 92  
 Chapman, L. J. 92  
 Chase, V. M. 28, 36–37  
 Chassin, L. 267  
 Chater, N. 144, 147, 166  
 Cheek, N. N. 211  
 Chen, J. S. 302  
 Cheng, P. W. 112, 148  
 Chernev, A. 220  
 Cheyne, J. A. 36, 56, 327  
 Chiesi, F. 37  
 Chong, D. 62, 66  
 Chow, J.Y. L. 108, 118  
 Christandl, F. 394  
 Christensen-Szalanski, J.J. J. 436, 439–441, 445  
 Chuang, W. I. 301  
 Chudzynski, E. N. 130  
 Chung, J. 344  
 Cikara, M. 329  
 Clare, J. 395  
 Clark, D. P. 379  
 Clarke, C. E. 333  
 Claypool, H. M. 244  
 Cleary, A. M. 373  
 Clements, C. M. 133  
 Clifasefi, S. L. 428  
 Clinton, H. 324, 330  
 Coane, J. H. 379, 405  
 Coe-Odess, S. 211  
 Cohen, G. L. 395  
 Cohen, J. 27  
 Cohen, J. D. 345  
 Cohen, L. J. 149–150  
 Cohen, M. S. 317  
 Cohen, S. M. 372  
 Cohen, T. R. 277–278  
 Collisson, B. 277  
 Compton, R. J. 252  
 Connelly, B. S. 279  
 Connolly, T. 448  
 Cook, J. 324, 331  
 Coolin, A. 439, 441, 446  
 Cooper, W. H. 259  
 Corneille, O. 236  
 Cornell, K. R. 243  
 Corredor, J. 308  
 Cosmides, L. 148, 371  
 Costa, P.T. 278  
 Costello, F. 28, 31, 38–40, 97  
 Cowley, E. 134  
 Cox, J. R. 145  
 Cox, J. W. 324  
 Coyle, J. T. 5  
 Craik, F. I. M. 377–378  
 Craver-Lemley, C. 247  
 Creswell, K. G. 397  
 Crick, N. R. 343  
 Critcher, C. R. 214, 216–217  
 Crocker, J. 92  
 Crockett, M. J. 325, 329  
 Cromheeke, S. 36, 163  
 Cronbach, L. J. 272, 275  
 Cronk, N. 350  
 Croson, D. C. 302  
 Croson, R. 194

- Crupi, V. 28, 35, 37, 39  
 Cueto, J. 302  
 Curran, T. 430–431  
 Curtis-Holmes, J. 161
- D'Agostino, P. R. 246–247, 253, 255  
 D'Argembeau, A. 376  
 Daffner, K. R. 414  
 Dagnall, N. 373  
 Dalgleish, T. 342  
 Daneman, M. 366–367  
 Daniel, T. C. 390–391  
 Davey, S. J. 5  
 Davis, E. 38  
 Davis, M. H. 277  
 de Araujo, E. 325  
 de Freitas Melo, P. 335  
 de Keersmaecker, J. 233–234  
 De Melo, P. O.V. 335  
 De Mul, S. 361, 363–365  
 De Neys, W. 13–14, 20, 36, 46, 57, 163–164  
 de Toledo, T. F. N. 130, 132  
 de Vries, R. E. 275, 277  
 Dearden, T. E. 37  
 Dearnaley, E. J. 27  
 DeCapito, A. 391  
 Dechêne, A. 226, 229, 231–234  
 Deese, J. 404, 408–409  
 Delton, A. W. 375  
 Demarest, I. H. 388, 390  
 DeMartino, B. 72  
 Denison, S. 50  
 Dennett, D. C. 17  
 Dennis, N. A. 414–415  
 Derbish, M. H. 372–373  
 deStefano, J. 329  
 Dewey, C. 13–14  
 Dickins, T. E. 374–375  
 Dickinson, A. 108  
 Dietvorst, B. J. 441  
 diFonzo, N. 232  
 Dimoka, A. 211  
 Dion, K. K. 262–264, 266–269  
 Directorate-General for Communication 324  
 Ditto, P. H. 344  
 Dixit, P. 335  
 Dixon, M. J. 134  
 Dobelli, R. 5  
 Dobroth, K. M. 279  
 Dodson, C. S. 16  
 Doerksen, S. 376  
 Doherty, M. E. 194  
 Donati, C. 36  
 Donovan, S. 31, 36  
 Donovan, W. L. 133  
 Dooren, W. V. 197  
 Dorner, W. W. 96
- Doss, M. K. 412  
 Dougherty, M. R. P. 297  
 Dragonetti, R. 134  
 Drivdahl, S. B. 427  
 Druckman, J. N. 61–62, 64, 66  
 Dshemuchadse, M. 74  
 Dube, C. 162  
 Dubra, J. 287, 298  
 Duchene, S. 33  
 Duffy, A. 326  
 Dunlosky, J. 307–308, 314–315, 317, 373–374  
 Dunn, D. S. 128  
 Dunn, J. C. 162  
 Dywan, J. 412
- Eagly, A. H. 263, 266, 268–269  
 Ecker, U. K. H. 236, 324, 327–328  
 Eddy, D. M. 46, 58  
 Edgington, D. 166  
 Edwards, K. 84–87  
 Edwards, W. 5  
 Efran, M. G. 264, 267  
 Egloff, B. 17, 445, 447  
 Einstein, G. O. 379  
 Eisenegger, C. 133  
 Eisenhauer, M. 18, 445–446  
 Elfénbein, D. W. 302  
 Elia, F. 35  
 Ellison, B. 266  
 Elqayam, S. 12, 149–150  
 Emery, C. 50  
 Endo, Y. 345  
 Englich, B. 209, 211–212, 216  
 Engstler-Schooler, T.Y. 395, 397–398  
 Enke, B. 211, 215  
 Enli, G. 329  
 Epley, N. 201, 214–215  
 Epstein, S. 31, 36  
 Epstein, Z. 334  
 Erdfelder, E. 10–11, 142, 227, 231–232, 235, 255, 316, 333, 372, 375, 377–379, 436, 438–439, 442–446  
 Erev, I. 74, 296–297  
 Erickson, T. D. 359–361, 363–365  
 Ericsson, K. A. 409  
 Ernst, H. M. 97  
 Etling, K. M. 214  
 Evans, J. St. B.T. 5, 12–14, 20, 35, 56–57, 78, 83, 88, 140–145, 149–150, 154–162, 164–169  
 Eyre, R. N. 201
- Fabre, J. M. 5, 12  
 Fang, K. 329  
 Farrell, C. 134  
 Fast, N. J. 130, 133  
 Faul, F. 333  
 Fawcett, J. M. 408

- Fay, N. 331  
 Fazio, L. K. 227, 231–233  
 Fedeli, S. 324  
 Feist, M. I. 391  
 Fenton-O'Creavy, M. 134  
 Fernandez, A. 414  
 Fessler, D. M. T. 327  
 Fetchenhauer, D. 394  
 Fiddick, L. 148  
 Fiedler, K. 4, 11, 17, 27, 31, 92, 96–98, 101–104  
 Finkelstein, M. O. 290  
 Finkenauer, C. 376  
 Finley, J. S. 412  
 Finn, S. 324  
 Fiore, S. M. 395  
 Fischbach, G. D. 5  
 Fischer, I. 437  
 Fischer, R. 278  
 Fischhoff, B. 50, 288, 445  
 Fisher, A. V. 386  
 Fisher, M. 324  
 Fisk, J. E. 28, 31–32, 36, 39–40  
 Fiske, S. T. 198, 272  
 Flammimi, A. 327  
 Flavell, J. H. 307  
 Fleig, H. 98  
 Flores, A. R. 242, 254  
 Fodor, J. 148  
 Fong, G. T. 44  
 Fong, N. 212, 217  
 Fong, N. M. 211  
 Forester, G. 379  
 Forgas, J. P. 260, 263–264, 267, 269  
 Forsberg, A. 396  
 Forster, J. 241  
 Foster, J. L. 225, 430  
 Fournier, M. 348  
 Fournier, P. 325  
 Fowers, B. J. 345  
 Franco, R. 32  
 Frankovic, K. 325  
 Franks, B. A. 413  
 Franssens, S. 57  
 Freckelton, I. 108  
 Frederick, S. W. 36, 56, 203–204, 214, 216,  
     218–220  
 Freeman, D. 115  
 Freeman, H. R. 245  
 Frenda, S. J. 425, 432  
 Fresco, D. M. 115  
 Freund, T. 248  
 Freytag, P. 96, 98  
 Friedland, L. 325  
 Friedland, N. 130  
 Friedman, M. 11  
 Friendmeyer, M. T. 267  
 Frisch, D. 201  
 Fugelsang, J. A. 36, 45, 56–57, 327  
 Funder, D. C. 5, 279  
 Fung, H. H. 343–344  
 Furr, R. M. 275  
 Gaertig, C. 215  
 Gaeth, G. J. 62, 66  
 Galam, S. 334  
 Gale, A. G. 84, 143  
 Galinsky, A. D. 130, 211–212  
 Gallagher, F. M. 346  
 Gallagher, K. M. 66–67  
 Galley, D. J. 244  
 Gallo, D. A. 344, 404–405, 407–414  
 Galton, F. 204  
 Garb, H. N. 199  
 Garcia-Marques, T. 229–231, 233  
 Garcia, A. D. 372  
 Gardner, J. 4, 13, 15  
 Garety, P. A. 115  
 Garimella, K. 335  
 Garnham, A. 161  
 Garrett, N. 347  
 Garrett, R. K. 328  
 Garrido, M. V. 373  
 Garry, M. 426, 429  
 Garza, A. A. 252  
 Gavanski, I. 200  
 Gaviria, C. 308  
 Gawlik, B. 392  
 Geil, M. 83, 89  
 Geniole, S. N. 133  
 Gennaioli, N. 199  
 Gentner, D. 391  
 George, C. 167  
 Gerbino, G. 124  
 Getz, S. J. 135  
 Gibb, M. 57, 163  
 Gibbert, M. 326  
 Gibbons, F. X. 278  
 Gibbons, J. A. 348  
 Gifford, R. K. 93–94  
 Gigerenzer, G. 4, 6–9, 11, 13–14, 16–17, 46, 50,  
     52, 58, 145, 147, 151, 179, 181, 195, 201, 225,  
     232, 291–292, 294–296  
 Giglietto, F. 330  
 Gignac, G. E. 327, 331  
 Gillard, E. 197  
 Gillebaart, M. 241, 243, 245  
 Gilovich, T. 4–6, 9, 13, 212, 214–217  
 Giroto, V. 83, 148–149  
 Giroux, M. E. 440, 448  
 Gladwell, M. 151  
 Glaser, M. 290, 301  
 Gleser, G. C. 275  
 Glöckner, A. 216  
 Glumicic, T. 46, 163

- Goel, A. M. 302  
 Goetz, T. 133  
 Gold, E. 194  
 Goldberg, L. R. 278  
 Goldsmith, M. 311, 315  
 Golebiewski, M. 329  
 Gong, J. 66  
 Gonsalves, B. D. 431  
 Gonzalez-Vallejo, C. 300  
 Gonzalez, M. 5, 12  
 Goodwin, H. 342  
 Goodwin, K. A. 409  
 Goschke, T. 74  
 Gosling, C. J. 72  
 Gothib, I. H. 341  
 Gottfried, J. 324  
 Gradl, P. 69, 71  
 Graham, J. R. 302  
 Graham, S. A. 386  
 Gramm, K. 97  
 Gramzow, R. H. 345  
 Gregory, K. J. 372  
 Gregory, R. L. 248, 250  
 Greifeneder, R. 183, 185–186, 366  
 Greve, K. W. 96, 251  
 Grice, H. P. 213, 215, 229, 360  
 Griffin, D. 9, 13, 287  
 Griggs, R. A. 145  
 Grinberg, N. 325–328, 333  
 Gronchi, G. 36  
 Groß, J. 439–441  
 Gross, R. 266  
 Gruenfeld, D. H. 130  
 Gualtieri, S. 50  
 Guazzini, A. 36  
 Guerci, E. 33  
 Guess, A. 325, 327, 329, 333–334  
 Guess, A. M. 327  
 Guilbault, R. L. 436, 439, 441, 445  
 Guinote, A. 185  
 Guo, L. 334  
 Gureckis, T. M. 38  
 Gwozdecki, J. 8  
 Gyoba, J. 245
- Ha, Y. 79, 97, 141, 144, 217  
 Hacquin, A.-S. 326  
 Haddock, G. 185  
 Hadjichristidis, C. 166  
 Hahn, U. 12, 17, 203  
 Hall, L. 85  
 Hallam, M. 269  
 Halpern, A. R. 247  
 Hamilton, D. L. 92–94, 97  
 Hammermesh, D. S. 267  
 Hammerton, M. 46  
 Hanawalt, N. G. 388, 390
- Handley, S. J. 57, 143, 154–155, 157, 162–163, 165–167  
 Hannah, S. D. 111  
 Hannon, B. 366–367  
 Hansel, C. E. M. 27  
 Hansen, J. 184, 229  
 Hansson, G. 28, 38–39  
 Harari, H. 266–267  
 Hards, E. 341  
 Hardt, O. 18  
 Hargis, M. B. 317–318  
 Harley, E. M. 440, 446, 448  
 Harper, C. 157, 162  
 Harper, D. R. 295  
 Harrigan, K. 134  
 Harris, A. J. 220  
 Harris, A. J. L. 12, 17  
 Harris, C. M. 397  
 Hart, J. 376–377  
 Hart, J. T. 314  
 Harvey, C. R. 302  
 Hasher, L. 92, 180, 225–226, 232, 441  
 Hashimoto, Y. 347  
 Hashtroodi, S. 408  
 Hastie, R. 212, 436, 439, 445  
 Haubrich, J. 281  
 Hausman, H. 319  
 Hawkins, S. A. 212, 225, 232–233, 436, 439, 445  
 Hayes, B. K. 162  
 Healy, P. J. 287, 297  
 Heaps, C. M. 430–431  
 Hearst, E. 96  
 Heider, F. 277  
 Heit, E. 162  
 Hell, W. 4, 7–8, 16, 441, 443  
 Heller, W. 252  
 Henderson, D. X. 350  
 Henkel, L. A. 412  
 Hennigan, K. 252  
 Hermann, P. 324  
 Hertwig, R. 10, 17, 28, 36–37, 74, 176, 186–187, 201, 291, 324–325  
 Herz, H. 302  
 Herz, R. S. 386, 394  
 Hester, G. 194  
 Hewstone, M. 98  
 Heyman, J. E. 219–220  
 Higgins, E. T. 63, 71–72, 131, 217, 387  
 Higham, P. A. 440  
 Hilbig, B. E. 10–11, 255, 274  
 Hill, P. L. 348  
 Hills, T. T. 326  
 Hilton, D. 301  
 Hinsz, V. B. 212  
 Hirshman, E. 410  
 Hisasaka, T. 307  
 Hitchison, K. A. 409

- Hoch, S. J. 225, 233, 446, 448  
 Hodgson, R. 5  
 Hoffrage, U. 11, 17, 46, 50, 52, 58, 291, 445–446, 448  
 Hogan, H. P. 386–390  
 Hogarth, R. M. 180  
 Holbrook, C. 327  
 Holliday, R. E. 414  
 Holmes, D. S. 273  
 Holtrop, D. 278  
 Holtzworth-Munroe, A. 343  
 Holyoak, K. J. 148  
 Holzworth, R. J. 194  
 Homer 198  
 Hong, R. Y. 274, 277  
 Hope, C. 57  
 Hou, C. 379  
 Houston, C. 214  
 Hövel, P. 326  
 Howard, D. V. 379  
 Howard, J. 5  
 Howard, J. H. 379  
 Howe, M. L. 372–373, 380  
 Howell, J. L. 277  
 Howland, B. 333  
 Hribar, P. 303  
 Hsee, C. K. 72  
 Hsieh, P. J. 246  
 Huang, Y. F. 246  
 Hudgens-Haney, M. E. 134–135  
 Huff, M. J. 408  
 Hug, K. 147  
 Hughes, B. T. 280–281  
 Human, L. J. 272, 275–278  
 Hunt, C. 342  
 Hunt, L. J. 397  
 Hunt, R. R. 379, 413  
 Hutchinson, G. 343  
 Hutchinson, J. W. 183–184, 290, 293  
 Hwang, A. J. 379  
 Hyman, I. E. 425–426  
 Iannelli, L. 330  
 Indahl, K. E. 212  
 Inoue, K. 241–242  
 Isenberg, D. J. 87  
 Israel, L. 413  
 Jack, F. 430  
 Jacobson, G. C. 329  
 Jacoby, L. L. 412, 425, 429  
 Jacowitz, K. E. 211, 215  
 Jalbert, M. 234–235  
 James, W. 419  
 Janik, B. 211  
 Janiszewski, C. 216  
 Jenkins, H. M. 112  
 Jensen, J. D. 66–67  
 Jetter, M. 328  
 Johannesson, M. 294  
 Johansson, P. 85  
 Johnson-Laird, P. N. 140–142, 145, 154, 157, 161  
 Johnson, C. A. 372  
 Johnson, E. J. 211, 217–218  
 Johnson, M. 412  
 Johnson, M. K. 408  
 Johnson, R. A. 204  
 Johnson, S. C. 57–58  
 Johnson, W. 396  
 Johnston, B. C. 130  
 Johnston, L. 267  
 Jokela, M. 269  
 Jones, C. R. 387  
 Jones, I. F. 244  
 Jones, K. T. 201  
 Jones, S. K. 201  
 Joormann, J. 341  
 Joseph, K. 325  
 Jost, J. T. 325, 327  
 Juanchich, M. 204  
 Judd, C. M. 366  
 Judisch, J. M. 8  
 Jung, M. H. 212  
 Juslin, P. 28, 38–39, 291, 293–295, 297  
 Juul, J. S. 334  
 Kahneman, D. 3–4, 8–9, 17, 27–28, 32, 34, 36–37, 44, 46–50, 52, 56–57, 61–65, 144, 175–183, 186–187, 191, 193–194, 196–201, 203–204, 209–216, 295  
 Kail, R. V. 245  
 Kalakanis, L. 269  
 Kale, A. R. 243  
 Kamal-Smith, E. 162  
 Kamas, E. N. 365–366  
 Kamp, S.-M. 378–379  
 Kang, S. H. K. 372, 376–377  
 Kao, S.-F. 93, 96  
 Kapur, N. 251, 254  
 Karanian, J. M. 430  
 Kareev, Y. 104  
 Katz, R. C. 387  
 Kauff, D. M. 246  
 Kaufman, M. 133  
 Kausel, E. E. 440  
 Kawakami, N. 244, 247  
 Kazanas, S. A. 372–373  
 Keane, M. T. 30  
 Keinan, G. 130  
 Kelemen, W. L. 372  
 Kellen, D. 142, 162  
 Keller, J. 185  
 Kelley, C. M. 315–316, 412, 425  
 Kelman, M. 9

- Kennedy, Q. 344  
 Kenny, D. A. 272, 275–278, 280  
 Kensinger, E. A. 376  
 Kent, D. R. 134  
 Keren, G. 14, 73, 145, 204, 288, 293  
 Keysar, B. 212  
 Kiat, J. E. 431  
 Kiesler, D. J. 280  
 Kim, N. 326  
 Kim, N. S. 447  
 King, G. 330  
 Kirk, R. 329  
 Kirkpatrick, E. A. 402–403  
 Klaczynski, P. A. 168  
 Klauer, K. C. 97, 142, 159, 162, 164  
 Klayman, J. 79, 97, 141, 144, 217, 289, 300  
 Klein, R. A. 67, 214, 216–217  
 Klein, S. B. 373, 375–376  
 Kliegl, O. 18  
 Klusowski, J. 128–129, 132  
 Knecht, S. 133  
 Koehler, D. J. 36, 45, 56–57, 327  
 Kok, I. 365  
 Kola, I. 142  
 Kongthong, N. 252  
 Kool, W. 135  
 Koretelting, J. E. 5  
 Koriat, A. 307–312, 314–316, 318–319  
 Korn, C. W. 347  
 Kornell, N. 310, 317  
 Kort, J. 269  
 Kosinski, M. 329  
 Kostic, B. 372–373  
 Koutstaal, W. 16  
 Koval, P. 266, 268–269  
 Kozyreva, A. 324–326, 333–334  
 Kramer, M. E. 377  
 Krantz, D. H. 27–28, 36, 44  
 Kriegelbauer, M. 378  
 Kroneisen, M. 372–373, 375, 377–379  
 Krosnick, J. A. 50  
 Krouwel, A. P. M. 327  
 Krueger, J. I. 5, 11, 273  
 Kruger, J. 344  
 Krüger, N. 379  
 Kruglanski, A. W. 14, 145, 234, 248, 254  
 Krupic, D. 342  
 Küfner, A. C. P. 17  
 Kühberger, A. 66–67, 69–72  
 Kuhlmann, B. G. 97  
 Kuhn, D. 84, 89  
 Kunda, Z. 87–89  
 Kunst-Wilson, W. R. 246  
 Kurkela, K. A. 414–415  
 Kurzenhäuser, S. 46, 52  
 Kusbit, G. W. 363–364  
 Kutzner, F. 98  
 Kuzmanovic, B. 347  
 Kwong, J.Y.Y. 216–217  
 La Voie, L. 275  
 Lachmann, G. 7  
 Ladouceur, R. 134  
 Laham, S. M. 266, 268–269  
 Lamberty, P. K. 440  
 Lampinen, J. M. 407, 412–414  
 Landfield, K. 89  
 Landy, D. 263–265  
 Lane, D. J. 278  
 Langens, T. A. 129, 131–132  
 Langer, E. J. 110, 124, 127–129, 131–132, 136  
 Langlois, J. H. 269  
 Lanska, M. 376  
 LaPaglia, J. A. 425  
 Laplace, P. S. 192  
 Lapsley, D. K. 348  
 Larimer, M. E. 135  
 Larson, A. 269  
 Layman, E. 200  
 Lazer, D. 325  
 Leary, M. R. 345  
 Leavitt, L. A. 133  
 LeBoeuf, R. A. 215–216  
 Lee, A.Y. 72  
 Lee, B. S. 301  
 Lee, D. S. 366  
 Lee, J. 272, 277, 329  
 Lee, K. 276, 278–279  
 Lehman, D. R. 50  
 Lehmann, S. 326  
 Leibowitz, H. W. 8  
 Lem, S. 197  
 Lemay, E. P. 345  
 Leone, D. R. 244  
 Levin, B. 290  
 Levin, I. P. 62, 65–66  
 Levy-Sadot, R. 315  
 Lewandowski, G. 277  
 Lewandowsky, S. 236, 324–325, 328, 331,  
     334–335, 395, 428  
 Lewis, B. P. 411  
 Lewis, J. 215  
 Li, F. 50  
 Li, L. 379  
 Li, T. 343  
 Lichtenstein, S. 180, 212, 288–289, 293–294  
 Lilienfeld, S. O. 89, 117  
 Lim, A. 108  
 Lim, M. S. M. 134  
 Lim, Z. W. 330  
 Lindsay, D. S. 408  
 Ling, R. 326, 330  
 Link, B. G. 387  
 Lipnevich, A. A. 133

- Lipscomb, T. J. 393  
 Liu, J. 277, 281  
 Liu, Z. 379  
 Livingstone, C. 134  
 Lloyd-Jones, T. J. 395, 398  
 Lo, Y. F. 386  
 Locke, K. D. 277–280  
 Loewenstein, G. F. 72, 211, 446, 448  
 Loftus, E. F. 10, 386, 392–393, 421–423, 426,  
     429–432  
 Logan, J. M. 317  
 Logie, R. H. 396  
 Løhre, E. 204  
 Lommen, M. J. 427  
 Longo, L. C. 263  
 Lorenz-Spreen, P. 325–326  
 Lorenz, R. C. 136  
 Lorko, M. 211  
 Lotz, S. 394  
 Louie, T. A. 448  
 Lovallo, D. 301–302  
 Lu, J. G. 328  
 Lucas, E. J. 84, 143  
 Luce, P. A. 411  
 Lücking, A. 46, 52  
 Ludeke, S. G. 277, 281  
 Ludolph, R. 16  
 Ludwin-Peery, E. 38  
 Luipersbeck, S. M. 225  
 Luks, S. 330  
 Luna, K. 311  
 Luo, J. 162  
 Lupfer, M. B. 200  
 Lupyán, G. 391  
 Lynch, J. St. 83, 142–143  
 Lyons, B. A. 327
- Ma, L. 264  
 Ma'ayan, H. 316  
 Maass, A. 445  
 Mac, R. 335  
 MacFarlane, D. 108, 118  
 Machunsky, M. 277  
 MacKay, D. G. 365  
 McLaren, V. 134  
 Madsen, J. K. 334  
 Magnussen, S. 315  
 Maguire, P. 30  
 Maguire, R. 30  
 Maibach, E. W. 333  
 Makerud, S. E. 372  
 Makhijani, M. G. 263  
 Maki, R. H. 315  
 Malmendier, U. 302  
 Malmi, R. A. 92  
 Malouff, J. 212  
 Mandel, D. R. 70, 203
- Mandler, G. 244, 253  
 Manktelow, K. I. 146–148, 156  
 Mann, T. 63  
 Marin-Garcia, E. 247, 251, 254  
 Mark, M. M. 447  
 Markey, P. M. 280  
 Marks, G. 277  
 Marsh, E. J. 367  
 Marshall, D. A. 31  
 Marti, M. W. 212  
 Martin, P. J. 267  
 Mather, M. 412  
 Mathison, D. L. 278, 281  
 Matlin, M. W. 18  
 Matthes, J. 277  
 Matthews, M. 330  
 Mattson, M. E. 359–361, 363–365  
 Matute, H. 108, 110–111, 113–114, 118  
 Matz, S. C. 329  
 Mauser, G. A. 393–394  
 Mayer, E. 414  
 Mayr, S. 333  
 Mazurier, K. 301  
 Mazzoni, G. A. L. 426  
 Mcallister, H. A. 393  
 McCabe, D. P. 373, 376  
 McClelland, A. G. R. 293, 295  
 McClure, S. M. 252  
 McCrae, R. R. 278  
 McDaniel, M. A. 379  
 McDavid, J. W. 266–267  
 McDermott, K. B. 6, 16, 372, 402, 404–405, 407,  
     409, 411–413, 415, 429, 431  
 McDonald, K. P. 264  
 McFarlan, C. C. 373  
 McGahan, J. R. 95  
 McGeoch, J. A. 419  
 McKay, R. T. 17  
 McKenna, F. P. 125–128, 132, 134  
 McKenna, P. A. 241  
 McKenzie, C. R. M. 73–74, 95, 103  
 McMillan, D. 96  
 McPhetres, J. 328  
 McStay, A. 325  
 Meade, M. L. 409  
 Medvec, V. H. 212  
 Meek, S. W. 431  
 Meiser, T. 97–98  
 Meissner, C. A. 395, 398, 409  
 Melcher, J. M. 393, 396–397  
 Mellinger, A. E. 129  
 Mellor, S. 447  
 Memon, A. 395  
 Menczer, F. 327  
 Mercier, H. 83, 86–89, 308, 325–326, 332–333  
 Merckelbach, H. 376  
 Mesulam, M. M. 5

- Metaxas, P.T. 324, 329  
 Metcalfe, J. 307–308, 312, 314–315, 317, 425  
 Meyerowitz, B. E. 62, 65  
 Mezulis, A. H. 347  
 Michalkiewicz, M. 10–11  
 Mickes, L. 395–396  
 Mikels, J.A. 343  
 Miles, J. N. 84, 143  
 Milkman, K. L. 325  
 Mill, J. S. 199–200  
 Miller, M. B. 407  
 Miller, N. 277  
 Miller, P.J. E. 346  
 Minami, T. 252  
 Misirlisoy, M. 379  
 Mitchell, A. 324  
 Mobius, M. 333  
 Mochon, D. 214, 216, 218–220  
 Modolo, K. 361, 365–366  
 Mojardin, A. H. 409  
 Molz, G. 6–7, 17–18  
 Monahan, J. L. 241, 251, 254  
 Mønsted, B. M. 326  
 Montepare, J. 272  
 Montgomery, H. 294  
 Montgomery, J. M. 327  
 Montoya, R. M. 243–246, 248, 252–253  
 Moore, D. A. 287, 293, 297  
 Moore, J. E. 267  
 Moore, K. N. 414  
 Moore, M. T. 115  
 Moore, S. A. 427  
 Moreland, R. L. 241, 244, 252  
 Moreno-Fernández, M. M. 110, 114–116  
 Moritz, S. 16, 135  
 Morley, N. J. 162  
 Morrier, D. M. 31  
 Morris, M. W. 200  
 Morris, S. L. 263–264  
 Morrison, K. R. 277  
 Morsanyi, K. 37, 163  
 Moscovitch, M. 377  
 Moser, P. 30  
 Moshagen, M. 278, 439  
 Moshman, D. 83, 89  
 Moskowitz, D. S. 278  
 Mosleh, M. 333  
 Moss, S. A. 278  
 Moutier, S. 72  
 Moxey, L. M. 73  
 Moyens, E. 163  
 Msetfi, R. M. 115  
 Mueller, M. L. 316  
 Müller, P.A. 447  
 Munn, L. 329  
 Murayama, K. 347  
 Murphy, A. H. 288, 294  
 Murphy, R. A. 113  
 Murphy, S. G. 252  
 Murphy, S. T. 241, 244, 246, 254  
 Murray, S. L. 345–346  
 Musch, J. 159, 440–441, 447  
 Mussweiler, T. 209, 211–213, 215–219  
 Nabi, R. L. 66, 72  
 Nadarevic, L. 227, 230–235  
 Nagae, S. 391  
 Nagler, J. 325, 327  
 Nairne, J. S. 371–372, 375–380  
 Nakamura, K. 37  
 Nakamura, Y. 244  
 Nakauchi, S. 252  
 Nanda, R. 302  
 Napoleon, T. 267  
 Narens, L. 307  
 Nasby, W. 343  
 Nash, M. 430–431  
 Naumer, B. 159  
 Nave, G. 329  
 Neale, M. A. 211, 213, 215  
 Neely, J. H. 218  
 Nees, M. A. 344  
 Neilens, H. 167  
 Nelson, L. D. 212, 215  
 Nelson, T. O. 307, 309  
 Nestler, S. 17, 275, 445–447  
 Neumann, H. 395  
 Neuschatz, J. S. 407, 412  
 Newell, B. R. 18, 217, 247  
 Newman, E. J. 233–234  
 Newman, I. R. 57, 163  
 Newman, J. 96  
 Newstead, S. E. 159, 162–163  
 Nicholson, N. 134  
 Nickel, S. 97  
 Nickerson, R. S. 78  
 Nikolaisen, M. I. 73  
 Nikolov, D. 327–328  
 Nilsson, H. 28, 38–39  
 Nir, L. 325  
 Nisbett, R. E. 44, 181, 264, 269  
 Noguchi, T. 326  
 Northcraft, G. B. 211, 213, 215  
 Nouchi, R. 372, 379  
 Nyhan, B. 327, 329  
 O'Brien, E. G. 372  
 O'Connor, K. 263  
 O'Connor, M. 264, 293  
 O'Connor, M. G. 247  
 O'Keefe, D.J. 66–67  
 Oakhill, J. 161  
 Oaksford, M. 144, 147, 166  
 Obama, B. 330

- Odean, T. 301  
 Oeusoothornwattana, O. 255  
 Ofir, C. 185  
 Ohtani, K. 307  
 Ojemann, J. G. 415  
 Okado, Y. 430  
 Olds, J. M. 376  
 Olsson, H. 293, 295, 297  
 Ones, D. S. 279  
 Onslow, M. 200  
 Open Science Collaboration 372  
 Oppenheimer, D. M. 186, 201, 204, 216  
 Oreskes, N. 334  
 Ortmann, A. 10  
 Osman, M. 36, 163  
 Ost, J. 426  
 Ostrove, N. 263, 266–268  
 Otgaard, H. 372–373, 376, 380, 427–429, 432  
 Ott-Holland, C. J. 278  
 Over, D. E. 35, 146–149, 165–168  
 Ozanne, M. 265
- Pachur, T. 439, 441  
 Packman, A. 200  
 Pagin, A. 31  
 Paivio, A. 387, 397  
 Palmer, J. C. 10, 386, 392–393, 421  
 Palmore, C. C. 372, 377  
 Pan, J. 330  
 Pandeirada, J. N. S. 371–373, 375–380  
 Papp, L. M. 278  
 Pardilla-Delgado, E. 408  
 Park, H. 18, 364  
 Park, J. H. 129  
 Parker, A. 373  
 Parkin, A. 373  
 Parks, C. M. 229–230  
 Paschen, J. 325  
 Pastorelli, C. 124  
 Paul, C. 330  
 Paunonen, S. V. 274, 277  
 Payne, D. G. 407, 412  
 Payne, J. 408  
 Payne, J. W. 212  
 Peake, P. K. 211  
 Pekrun, R. 133  
 Penfield, W. 420–421  
 Pennycook, G. 36, 45–46, 52–58, 162–164, 225,  
   231–232, 235, 327–328, 330, 333  
 Pentland, J. 425  
 Perales, J. C. 112  
 Perfect, T. J. 397  
 Perfecto, H. 212  
 Perkins, D. N. 78, 84  
 Perner, J. 66–67, 70  
 Peters, M. J. V. 376  
 Peters, U. 17
- Petty, R. E. 233  
 Pezzo, M. V. 436, 439, 447–448  
 Pezzo, S. P. 447–448  
 Pfattheicher, S. 327, 395  
 Pfeiffer, P. E. 211, 296  
 Pham, M. T. 183  
 Phelan, J. C. 387  
 Phillips, L. D. 288  
 Phillips, P. 162  
 Piaget, J. 156  
 Piatelli-Palmarini, M. 5, 17  
 Pickrell, J. E. 426, 429, 431  
 Pidgeon, N. 28, 36, 40  
 Pierce, B. H. 410  
 Pieters, R. 448  
 Pilditch, T. D. 334  
 Pinon, A. 66  
 Pisoni, D. B. 412  
 Pisor, A. C. 327  
 Plassmann, H. 394  
 Plessner, H. 96  
 Plous, S. 211  
 Pohl, R. F. 6–8, 10–11, 16–18, 187, 255, 392, 394,  
   436–441, 443, 445–448  
 Polage, D. C. 235  
 Poletiek, F. H. 79, 141, 150  
 Pollard, P. 158–159  
 Polonioli, A. 9, 13, 17  
 Porter, M. A. 334  
 Porter, S. 429, 431  
 Posen, H. E. 302  
 Pothos, E. M. 32, 326  
 Pouget, S. 301  
 Prelec, D. 211  
 Prentice, W. C. H. 390–391, 393  
 Price, J. R. 252, 254  
 Primi, C. 37, 93  
 Próchnicki, M. 211  
 Profaci, C. P. 379  
 Pronin, E. 344  
 Prowse Turner, J. A. 163  
 Putnam, A. L. 425
- Quattrone, G. A. 214  
 Quayle, J. D. 162
- Racine, C. 413–414  
 Ragni, M. 142  
 Rand, D. G. 327–328, 333  
 Rand, K. M. 409  
 Raoelison, M. 164  
 Rau, R. 275, 280  
 Rawson, K. A. 317, 373–374  
 Raymaekers, L. 376  
 Read, J. D. 393  
 Ready, R. E. 272–273, 278  
 Reber, R. 186–187, 229, 368

- Reder, L. M. 18, 312, 315, 363–365, 424–425  
 Reed, A. E. 343  
 Regev, Y. 130  
 Reifler, J. 327, 329  
 Renner, C. H. 18, 255  
 Renner, M. J. 255  
 Repnikova, M. 329  
 Rescorla, R. A. 112–113  
 Resick, P. A. 428  
 Reyes, R. M. 181  
 Reyna, V. F. 71–72, 409, 414  
 Rhodes, M. G. 311, 315  
 Rholes, W. S. 387  
 Rietveld, C. A. 302  
 Riketta, M. 277  
 Rinck, M. 244  
 Risbey, J. S. 334  
 Ritchie, T. D. 348  
 Ritter, F. E. 312, 315  
 Roberts, M. E. 330  
 Robertson, T. E. 375  
 Robinson, B. 168  
 Robinson, K. 402, 408–409  
 Rodin, J. 124  
 Roediger, H. L., III 5–8, 16, 372–373, 377, 402, 404–405, 407–409, 411–413, 429, 431  
 Röer, J. P. 376, 378–379  
 Roese, N. J. 436, 439, 445, 447–448  
 Rogers, P. 28, 31, 36  
 Rogers, R. D. 134  
 Rom, S. C. 227, 230–231, 233  
 Roozenbeek, J. 326  
 Rose, R. L. 92  
 Röseler, L. 214, 217  
 Rosenthal, R. 67  
 Ross, L. 181, 274  
 Ross, M. 17, 175, 182, 287, 344  
 Rossi, L. 330  
 Rotello, C. 162  
 Roth, J. 129  
 Rothbart, M. 274  
 Rothman, A. J. 63, 185–186  
 Rothschild, D. 333  
 Rotteveel, M. 241  
 Rottman, B. M. 108  
 Rubenstein, A. J. 269  
 Rubin, D. B. 67  
 Ruder, M. 185  
 Ruggieri, S. 242  
 Ruiz-Vargas, J. M. 251, 254  
 Rummel, J. 98, 373, 377  
 Russer, S. 97  
 Russo, J. E. 292–293  
 Russo, S. 28  
 Sacramento, C. A. 277  
 Sadler, P. 280  
 Sailors, J. J. 219–220  
 Sajuria, J. 327  
 Sakaki, M. 347  
 Salovey, P. 63  
 Sanders, J. D. 252  
 Sandewall, Ö. 294  
 Sanford, A. J. 73  
 Sanft, H. 92  
 Sanna, L. J. 446  
 Santos-Pinto, L. 289, 299  
 Saraiva, M. 373  
 Sato, N. 241  
 Savary, J. 216  
 Savine, A. C. 373, 379  
 Savitsky, K. 212  
 Schacter, D. L. 5, 16–17, 327, 410, 413–414, 430, 432  
 Schaffner, B. F. 330  
 Scheier, M. F. 348  
 Scherbaum, S. 74  
 Schill, D. 329  
 Schindler, S. 327  
 Schkade, D. A. 212  
 Schlehofer, M. M. 134  
 Schneider, S. L. 62, 66, 71, 294  
 Schnicke, M. K. 428  
 Schnuerch, M. 234  
 Scholer, A. A. 72  
 Schooler, J. W. 393, 395–398  
 Schryer, E. 344  
 Schulz, Y. 14, 145, 204  
 Schulman, A. I. 377  
 Schulte-Mecklenbeck, M. 66–67, 70–71  
 Schulz, P. J. 16  
 Schulze, C. 18  
 Schutte, N. S. 212  
 Schütte, S. 379  
 Schwartz, B. L. 211, 312, 315  
 Schwartz, S. H. 279  
 Schwarz, N. 183–187, 215–216, 229–230, 324, 366, 440, 446  
 Schwarz, S. 446  
 Scofield, J. E. 372  
 Scott-Phillips, T. C. 374–375  
 Scott, J. 342  
 Scullin, M. K. 373  
 Seale-Carlisle, T. M. 395  
 Seamon, J. G. 241, 246  
 Sedikides, C. 347  
 Sedlmeier, P. 181, 187, 195  
 Seger, C. R. 346  
 Seifert, C. 324  
 Selfhout, M. 277, 279  
 Selvaggi, S. 244  
 Servátka, M. 211  
 Sévigny, S. 134  
 Shackle, G. L. S. 39–40

- Shafto, M. 365  
 Shah, A. K. 204  
 Shanks, D. R. 108, 112, 214, 217, 247, 255  
 Sharot, T. 347  
 Shaw, J. 6, 16  
 Shefrin, H. 199  
 Shepperd, J. A. 347  
 Sher, S. 73–74  
 Sherman, D. 63  
 Sherman, S. J. 97  
 Sherry, C. L. 231, 233  
 Shimamura, A. 376  
 Shleifer, A. 199  
 Shoemaker, P. J. H. 292–293  
 Sicolý, F. 175, 182  
 Sidi, Y. 318  
 Siegel, P. 244, 248  
 Sigall, H. 263–268  
 Silva, R. R. 230–232  
 Silvera, D. H. 346  
 Sim, D. L. H. 200  
 Simmons, J. P. 128–129, 132, 215  
 Simoes, R. A. G. 130, 132  
 Simonsohn, U. 441  
 Sims, V. 244  
 Singmann, H. 164  
 Sivanathan, N. 130  
 Skóra, Z. 396, 398–399  
 Skórská, P. 211  
 Skworonki, J. J. 348  
 Sleegers, W. W. A. 447  
 Sloman, S. A. 44, 46, 52, 58, 166  
 Sloof, R. 131  
 Slotnick, S. D. 430  
 Sloutsky, V. M. 386  
 Slovic, P. 4, 212  
 Small, D. A. 128–129, 132  
 Smeets, T. 372–373, 376  
 Smelter, T. J. 225, 231, 235  
 Smith, A. R. 212  
 Smith, E. E. 84–87  
 Smith, H. D. 193  
 Smith, P. K. 247  
 Smith, R. E. 413  
 Smoot, M. 269  
 Smorti, A. 36  
 Snyder, M. 80–81  
 Soane, E. 134  
 Sobel, J. 289, 299  
 Soderstrom, N. C. 373, 376  
 Soll, J. B. 289, 297, 300  
 Solomons, W. 394  
 Sommers, M. S. 411  
 Song, H. 366  
 Soroka, S. 325  
 Sörqvist, P. 266  
 Souza, A. S. 396, 398–399  
 Spacapan, S. 124  
 Speckmann, F. 361–363, 367–368  
 Speechley, W. 115  
 Speekenbrink, M. 220  
 Sperber, D. 83, 86–89, 148–149  
 Sporer, S. L. 396, 398  
 Spring, V. L. 329  
 Sprinkle, G. B. 368  
 Srivastava, S. 278, 280  
 Stadler, M. A. 404  
 Stahl, C. 142, 229  
 Stahlberg, D. 445–447  
 Stang, D. J. 245–246, 253  
 Stangor, C. 96  
 Stanovich, K. E. 13–14, 56–57, 89, 144–145, 150, 154, 167–168, 204  
 Stanton, C. H. 341  
 Starbird, K. 330–331  
 Stark, C. E. L. 430–431  
 Steiger, A. 66–67  
 Steinberg, N. 447  
 Stephens, R. G. 162  
 Sterling, J. 327  
 Sternberg, R. J. 254  
 Stevenson, R. J. 167  
 Stillman, C. M. 379  
 Stillwell, D. J. 329  
 Stock, R. 31  
 Stocking, G. 324  
 Strack, F. 209, 211, 213, 215–220  
 Strandberg, T. 85  
 Strecher, V. J. 348  
 Strohmaier, N. 448  
 Studer, B. 133  
 Stump, A. 232, 234  
 Stupple, E. J. N. 162–163  
 Sullivan, A. L. 414  
 Sullivan, L. E. 5  
 Summers, A. D. 130  
 Sundali, J. 194  
 Sungkhasette, V. W. 412  
 Sunstein, C. R. 325  
 Susa, K. J. 398  
 Sutton, J. 17  
 Suzuki, M. 245  
 Svenson, O. 289  
 Swann, W. B. 80–81  
 Swire-Thompson, B. 325  
 Szollosi, A. 28–29  
 Szpitak, M. 425  
 Szűcs, D. 37  
 Tandoc, E. C. 326, 330  
 Tanner, C. 71  
 Tanyas, H. 379  
 Tassoni, C. J. 199  
 Tate, G. A. 302

- Tauber, S. K. 315–316  
 Taylor, H. A. 31  
 Taylor, S. E. 133, 198, 346  
 Teigen, K. 73  
 Teigen, K. H. 204  
 Tentori, K. 28, 34–35, 39  
 Tesser, A. 440, 447–448  
 Thakor, A.V. 302  
 Thaler, R. 17  
 Thielmann, I. 274, 277–279, 281, 395  
 Thomas, C. 128, 134  
 Thomas, D. R. 391  
 Thomas, G. 278  
 Thompson, C. P. 348  
 Thompson, D. M. 420  
 Thompson, S. C. 124, 127–129, 131–132,  
     134–136, 164  
 Thompson, S. R. 371–372, 375–377  
 Thompson, V.A. 52–55, 57–58, 145, 159,  
     162–164, 168–169  
 Thorndike, E. L. 259, 261–262  
 Thurik, A. R. 302  
 Tirole, J. 299  
 Titchener, E. B. 241  
 Toet, A. 5  
 Toglia, M. P. 407, 410  
 Toma, C. 277  
 Tombu, M. 70  
 Toneatto, T. 134  
 Tooby, J. 148, 371  
 Toplak, M. E. 89, 150, 204  
 Topolinski, S. 241, 244, 252, 313  
 Toppino, T. C. 225  
 Torres, M. 118  
 Torrès, O. 302  
 Toth, J. P. 229–230  
 Trippas, D. 57, 162–164  
 Trivers, R. 300  
 Trope, Y. 80–81  
 Trouche, E. 83, 87  
 Trueblood, J. S. 32  
 Trump, D. 329–330  
 Truong, L. 379  
 Tsanos, A. 134  
 Tse, C.-S. 373  
 Tsujii, T. 162  
 Tucker, J. 325, 327  
 Tucker, J. A. 325  
 Tulving, E. 377–378, 420  
 Tversky, A. 3–4, 8–9, 17, 27–28, 32, 34, 37, 44,  
     46–50, 52, 56, 61–65, 175–183, 186–187, 191,  
     193–194, 196–201, 209–214, 216, 295  
 Tykocinski, O. E. 447  
 Uhl, K. P. 393  
 Umanath, S. 367, 405  
 Underwood, B. J. 403, 408–409  
 Undorf, M. 307, 309, 313–317, 440  
 Unkelbach, C. 227, 229–231, 233, 361–363,  
     366–368  
 Updegraff, J. A. 63, 66–67  
 Urban, E. J. 342  
 US Food and Drug Administration 108  
 Uy, D. 216  
 Vadillo, M. A. 111, 214  
 Valeriani, A. 330  
 Van Arsdall, J. E. 372, 379  
 Van Avermaet, E. 130  
 Van Bavel, J. J. 325  
 Van Bergen, S. 372  
 Van Boekel, M. 441  
 Van Buiten, M. 73  
 Van der Linden, M. 376  
 Van der Linden, S. 326, 333–334  
 Van Hauwaert, S. M. 327  
 Van Kessel, S. 327  
 Van Lange, P.A. M. 281  
 van Oostendorp, H. 361, 363–365  
 van Overschelde, J. P. 373–374  
 Van Prooijen, J.-W. 277, 327  
 Van Valkenburg, K. M. 373  
 Van Zandt, B. J. 244  
 Vanags, T. 397  
 VandenBos, G. R. 279  
 Vansteenvagen, D. 163  
 Vargo, C. J. 334  
 Varma, K. 441  
 Varma, S. 441  
 Verde, M. F. 162  
 Verfaellie, M. 414  
 Verschaffel, L. 197  
 Vieira, C. C. 335  
 Villejoubert, G. 203  
 Viscusi, D. 130  
 Vishny, R. 199  
 Vitanova, I. 300  
 Vogel, T. 97–98, 230  
 Vohs, K. D. 243, 376, 436, 439, 445, 447–448  
 Voltaire 350–351  
 Volz, K. 431  
 von Clef, J. 386, 394  
 von Collani, G. 445  
 von der Beck, I. 440  
 von Hippel, W. 269  
 von Siemens, F.A. 131  
 von Winterfeldt, D. 5  
 Vrungos, S. 134  
 Wade, C. N. 4 162  
 Wagner, A. R. 113  
 Wagner, T. 440–441  
 Walker, M. 324  
 Walker, W. R. 348–349

- Wallsten, T. S. 296–297  
 Walsh, R. O. 133  
 Walster, E. 262–264, 266  
 Walters, A. A. 386–390  
 Walther, E. 97  
 Wänke, M. 184, 187, 229  
 Wanless, A. 328–331  
 Wardle, J. 394  
 Warren, P. A. 203  
 Waruwu, B. K. 326  
 Wason, P. C. 10, 79–80, 83, 140–145  
 Wasserman, E. A. 93, 96, 108, 112  
 Watanabee, S. 162  
 Watkins, L. M. 267  
 Watson, D. 276–278  
 Watson, J. M. 405, 409, 411, 413, 415  
 Watts, D. J. 333  
 Watts, P. 28, 31, 38–40, 97  
 Waytz, A. 243  
 Weber, E. U. 72  
 Weber, M. 290, 301, 448  
 Weber, R. A. 302  
 Webster, D. M. 234  
 Weick, M. 185  
 Weingarten, E. 183–184  
 Weinstein, N. D. 182, 290, 302, 347  
 Weinstein, Y. 372–373  
 Weiss, J. A. 410  
 Welch, N. 72  
 Welder, A. N. 386  
 Weller, J. 277  
 Wells, G. L. 27–28, 200  
 Welsh, G. S. 250  
 Wen, W. 268  
 Wenger, A. 345  
 West, J. D. 428, 432  
 West, R. F. 56–57, 89, 150, 168  
 West, S. A. 374–375  
 West, T. V. 276–278, 280  
 Westerman, D. L. 376  
 Wetzel, C. G. 269  
 White, R. W. 124  
 Whittlesea, B. W. A. 229, 252, 254, 429  
 Wiener, C. 72  
 Willard, G. 345  
 Willett, C. L. 108  
 Willham, C. F. 436, 439–441, 445  
 Williams, E. F. 312–313  
 Williams, J. M. G. 342  
 Williams, L. D. 229, 429  
 Williamson, J. D. 95  
 Williamson, S. 252  
 Willman, P. 134  
 Wills, J. A. 325  
 Wilson, A. E. 17  
 Wilson, B. M. 395  
 Wilson, P. R. 265–266, 269  
 Wilson, S. 378  
 Wilson, T. 330, 387  
 Wilson, T. D. 128, 214, 216–217, 264, 269  
 Wiltshire, D. 28, 36  
 Windshitl, P. 212  
 Winkelman, P. 184  
 Winkler, R. L. 294  
 Winman, A. 28, 38–39, 293, 295  
 Wissler, R. L. 212  
 Witthöft, M. 16  
 Wixted, J. T. 396  
 Wojcik, S. P. 344  
 Wojciszke, B. 278  
 Wolf, N. R. 345  
 Wolff, W. T. 96  
 Wolford, G. L. 31, 407  
 Woltin, K.-A. 185  
 Wong, K. F. E. 216–217  
 Wood, D. 275, 278, 280  
 Woody, E. 280  
 Woolfolk, A. 387  
 Woolfolk, R. 387  
 Wöstenfeld, F. O. 373  
 Wright, D. S. 424–425  
 Wright, G. 196  
 Wright, R. 409  
 Wu, S. 50  
 Wurgler, J. 301  
 Xu, W. 329  
 Yagi, Y. 241–242  
 Yan, V. X. 318  
 Yang, C. 316  
 Yang, H. 303  
 Yang, L. 379  
 Yarritu, I. 110, 114  
 Yates, J. F. 27–28, 31, 34–35, 293  
 Yates, M. C. 95  
 Yoon, S. 211–212, 217  
 Yopchick, J. E. 447  
 Yoshida, F. 244, 247  
 Yoshimura, S. 347  
 Young, D. 267  
 Young, S. G. 244  
 Yu, J. 430  
 Zacks, R. T. 92, 180  
 Zajac, R. 430  
 Zajonc, R. B. 241, 244, 246, 252, 254–255  
 Zalmanov, H. 187  
 Zaragoza, M. S. 427  
 Zarate, M. A. 252, 254  
 Zebowitz, L. A. 272  
 Zettler, I. 274, 277  
 Zhang, L. 211  
 Zhang, S. X. 302

- Zhang, Y. 328  
Zhang, Y. C. 216  
Ziano, I. 344  
Zimdahl, M. F. 313–314, 316, 440
- Zimmerman, C. A. 315–316  
Zoellner, L. A. 427  
Zupanek, N. 187  
Zydervelt, S. 430

# Subject index

Note: Page numbers in *italics* indicate figures and in **bold** indicate tables on the corresponding pages.

- abstract Wason selection task 141–145  
accuracy (self-other agreement): assumed similarity and 273, 276; of metacognitive judgments 308–309  
acrophobia 342–343  
activation theories of associative memory illusions 411  
actual similarity 273  
adjustment and anchoring 4  
adopting the perspective of others 181–182  
affective primacy model of mere exposure effect (MEE) 254  
age-related positivity 343–344  
amount of recall 182–186  
anchoring effect 4, 209–221; components of 209–214; conversational inferences in 215–216; experiment 210–211; insufficient adjustment and 214–215; numeric priming and 216–217; paradigms of 213–214; pervasiveness and robustness in 211–212; relevance of 212–213; scale distortion in 219–220; selective accessibility model of 217–219; theoretical accounts of 214–220  
anger 325, 341, 343  
anxiety 16, 342–343, 345, 348–350  
applied perspectives on cognitive illusions 16  
argumentative theory of reasoning 88–89  
Argument Evaluation Task (AET) 168  
artificiality of cognitive illusions 9–11  
associative memory illusions 402–415; association-based theories of 408–409; associative tradition and 402–403; classroom demonstration of 403–404; DRM paradigm and 404–415, 406, 410; fluency-based attributions of 412–413; neural mechanisms of 414–415; processes that cause 408–413; processes that reduce 413–414; similarity-based theories of 409–412, 410; theories and data on 408–415, 410  
assumed similarity 272–282; accuracy and 273, 276; characteristics of perceivers and 277–278; characteristics of targets and 277; characteristics to be judged and 278; classroom demonstration of 275–276; defined 272–274; dynamic perception model (DPM) of 280–281; global positivity and 280; lack-of-information account of 278–279; measures of 274–275; moderators of 277–278; practical relevance of 281; theoretical accounts of 278–281; value account of 279  
attribute framing 62, 65, 67–68, 73  
attributive projection 273–274  
availability 4, 175–188; in adopting the perspective of others 181–182; applications of 180–182; in biased encoding and retrieval of information 180–181; defined 175–176, 186; difference between representativeness and 179; ease or amount of recall and 182–186; famous-names experiment on 176–177; letter-frequency experiment on 178–181; retrieval fluency and 186–187; in vividness of information 181  
Backward Associative Strength (BAS) 409  
*Bad arguments* 5  
base-rate effect and overconfidence 294  
base-rate neglect 44–59, 197; dual-process theory and 56–58; many forms of 45–49, 49; theoretical accounts of 50–58, 54–55  
Battig and Montague norms 373–374  
Bayesian theory 12  
belief bias 154–170; in conditional inference 166–168; dual processes and 161–162; in informal reasoning 168; logical intuitions and 163–164; in syllogistic reasoning 158, 158–161  
belief bias in deductive reasoning, conditional inference and 164–168  
bias(es): artificial 10–11; belief (*see* belief bias); bias 6; causality (*see* causality bias); cognitive 5, 16, 165; confirmation (*see* confirmation bias); conjunction fallacy as judgmental 33–38; encoding and information retrieval 180–181;

- foresight 318; in intellectual tasks 5; matching 142–146; motivational 14; negativity 341–343, 342; as not irrational 202–203; as not universal 201–202; positive-testing 141; positivity (*see* positivity bias); and rationality in human reasoning 149–151; stability 310–311
- bivariate (true-false) logic 12
- boundary conditions and halo effects 266–267
- calibration and overconfidence 287–288, 291–292
- Candide* 350–351
- categories of cognitive illusions 3–4
- causality bias 108–120, 109; applications of 115–117, 117; in base-rate problems 51; empirical evidence in 110–111, 111; individual differences in 114–115; pseudoscience and 117–119, 118; theoretical background on 112–114, 113
- causality judgments 199–200
- ceiling rule 37
- CHARM theory 425
- childhood sexual abuse 16
- choice blindness 85
- classroom demonstrations: assumed similarity 275–276; confirmation bias 81–83; illusory correlations 98–103, 102; labeling and overshadowing effects 388–390; mere exposure effect (MEE) 248–249, 250; Moses illusion 361–363; overconfidence 290–293; survival processing effect 373–374; truth effect 228
- cognitive accounts: framing 70–71; hindsight bias 445–446
- cognitive biases 5, 16, 165
- cognitive-experiential self-theory (CEST) 36
- cognitive illusions 3–19; applied perspectives of 16; artificiality of 9–11; categories of 3–4; defining features of 7–8; deviating from reality 7; dual-process models of 13–14; explanations of 13–15; functional views of 16–18; implications of 16–18; as impossible to avoid 8; inadequate presentation format and material selection and 10–11; information-processing system 14–15; as involuntary 7–8; misleading information and 10; missing knowledge and 11; normative standards for 12; status of 6–13; summary of debate over 12–13; as systematic 7; true 11; as universal 8; wrong normative standards and 11
- cognitive miserliness 14
- Cognitive Processing Therapy 428
- color: labeling effects and memory for 391–392; verbal overshadowing on 396
- combined-cognitive-biases hypothesis 16
- combined error models of overconfidence 297
- conditional inference 164–168; belief bias in 166–168
- confidence-frequency effect and overconfidence 294
- confidence intervals 289–290
- confidence judgments 308
- configural weighted averaging (CWA) 38
- confirmation bias 16, 78–90; classroom demonstration 81–83; debiasing or making the best of myside bias and 89; demonstrations of 78–86; explaining the myside bias and 87–89; hypothesis testing 78–81; limits of the myside bias and 86–87; thought listing 84–86; Wason selection task 83–84
- congruity and survival processing effect 377
- conjunction fallacy 27–41, 197–198; configural weighted averaging and 38; consequences for human decision-making 37–38; conversational implicature of 31; different probabilistic rule applied to 31–32; dual-process theories of 35–37; example of 28–29; frequentist interpretations of 31; inductive confirmation and 34–35; joint probabilities and 39–40; as judgmental bias 33–38; laws of probability and 30–31; overstating the real-world consequences of 38–41; quantum probability of 32–33, 33; random noise or sampling errors and 38–39; representativeness and 34; signed summation and 34; source memory and 37
- conservatism 327–328
- consistency illusion 312–313
- conspiracy mentality 324–325, 440
- contaminated mindware 14
- content borrowing 412
- control heuristic 131
- conversational implicature of the conjunction fallacy 31
- conversational inferences in anchoring effect 215–216
- convincing others with overconfidence 300–301
- cooperative communication setting and Moses illusion 360–361
- counterfactual outcomes and representativeness 200
- COVID-19 pandemic 328, 334, 350
- cue-familiarity illusions 312
- death penalty 84, 86–87
- debiasing 158; hindsight bias and 441; truth effect and 235–236
- decision-making 5, 295; bivariate logic in 12; conjunction fallacy consequences for 37–38; medical 5; selection task as 145–149
- deduction paradigm 166
- deductive reasoning 154–156; belief bias in (*see* belief bias); conditional inference and 164–168; syllogistic reasoning 156–164

- depression 341–342; fading effect bias and 348, 350; hindsight bias and 440, 448; illusory control and 133–135
- deviation from reality, cognitive illusions and 7
- diagnosis and representativeness 195–197
- direct-access accounts of metacognition 314–315
- discrepancy-attribution hypothesis 229
- distinctiveness heuristic 413–414
- DRM paradigm 404–408, 406; neural mechanisms of 414–415; processes that produce 408–413, 410; processes that reduce 413–414
- dual-process theory 13–14; base-rate neglect and 56–58; belief bias and 161–162; conjunction fallacy and 35–37; Wason selection task and 144–145
- dying-to-remember (DTR) effect 376
- dynamic perception model (DPM) 280–281
- ease of recall 182–186
- ecocentrism 14, 300
- ecological models of overconfidence 295–296
- emotional arousal and survival processing effect 375–376
- encoding: biased 180–181; imperfect 363–364; survival processing effect and richness of 378–379
- error models of overconfidence 296–297
- errors, memory 5, 6
- Event Related Potential (ERP) 379, 430–431
- expectancy-based illusory correlation 96
- expertise 199, 211–212, 397, 440–441; Moses illusion and 367; overconfidence and 294
- external-validity problem 150
- eyewitness testimony 16
- face recognition, verbal overshadowing on 395–396
- Fading Affect Bias (FAB) 348–350
- false news 324–335; applied implications of 333–335; conservatism and 327–328; human attention and misinformation and 325; information load, cognitive capacity, and decision quality and 326–327; online reputation and sharing and 325–326; *see also* participatory propaganda
- false consensus effect 274
- false memories 429–432
- falseness, illusion of 230
- familiarity account of truth effect 228–229
- familiarity misattribution 429
- famous-names experiment 176–177, 182–183
- feeling of knowing (FOK) judgments 308, 311–312
- feelings in information retrieval 184–185
- fluency, retrieval 186–187
- fluency account of truth effect 229–230
- fluency-based attributions of associative memory illusions 412–413
- fluency-specificity hypothesis 230
- font-size illusion 311, 313–314
- foresight bias 318
- format dependence and overconfidence 294–295
- framing 61–75; attribute, goal, and risky-choice 62–64; broader construal of tasks in 73–74; class experiment in 68–70; cognitive accounts of 70–71; effect size 66–70; examples of 64–66; explaining the effect of 70–73; motivational and emotional accounts of 71–72; pragmatics of language and 73; varieties of 61–66
- frequencies versus probabilities 202
- frequentist interpretations of the conjunction fallacy 31
- functional views of cognitive illusions 16–18
- fuzzy-trace theory 71, 409, 411, 412, 413
- gambler's fallacy 194, 197
- Gestalt theories of perception 267–268
- goal framing 62–63
- half-block design 274–275
- halo effects 259–270; boundary conditions and 266–267; consequences and practical importance of 267; history of research on 261–262; moderators and bias reduction with 269; of other attributes 265–266; physical attractiveness 262–263, 262–265, 265; theoretical explanations and psychological mechanisms of 267–269; typical experiment and class illustration of 260–261
- halo error 259
- hard-easy effect 293
- heuristics 8–9
- “Heuristics and biases: The psychology of intuitive judgment” 4–5
- high accessibility of invalid partial information 312
- hindsight bias 8, 436–449; applied perspectives of 448; assessment of 437–439; cognitive processes in 445–446; components view of 445; computational models of 446; debiasing attempts and 441; defined 439; empirical evidence on 439–444, 444; examples of 436; experiment on 441–444, 444; individual differences in 440–441; metacognitions and 446–447; motivational processes in 447; theoretical accounts of 444–447; variants of 440
- Hostile Attribution Bias 343
- hybridization and associative memory illusions 411–412
- illusion of control 124–137; conditions for 128–131; implications of 132–135; intrusion

- of reality in 130; mood in 130; need or desire for the outcome in 129–130; negative consequences of 133–135; neural research on 135–136; overconfidence and 290; positive consequences of 133; power in 130–131; skill-related factors in 128–129; success or failure emphasis in 129; theories of 131–132
- illusion of validity 198
- illusions of memory 4, 5–6
- illusions of perception 6
- illusions of thinking 3–4
- illusory correlations 92–104; alignment of skewed base-rate distributions in 98; classroom demonstration of 98–103, 102; definitions in 94, 94–95; expectancy-based 96; experimental task and dependent measures in 95; phenomenon of 92–95; theoretical accounts of 96–98; unequal sample size and 97–98; unequal weighting of present versus absent information in 96–97
- illusory truth effect *see* truth effect
- imperfect encoding and Moses illusion 363–364
- imperfect matching and Moses illusion 364–365
- imperfect retrieval and Moses illusion 364
- implicit association response (IAR) 408–409
- impossibility of avoiding cognitive illusions 8
- inductive confirmation and conjunction fallacy 34–35
- inference, conditional 164–168
- informal reasoning, belief bias in 168
- information: biased retrieval of 180–181; feelings and retrieval of 184–185; high accessibility of invalid partial 312; processing motivation and processing opportunities with 185–186; retrieval fluency and 186–187; vividness of 180, 181, 312
- information load, online 326–327
- information overload 14
- information-processing system 14–15
- instructor-fluency illusion 311
- insufficient adjustment in anchoring effect 214–215
- integrative model of truth effect 230–231
- intellectual tasks, biases in 5
- interindividual differences and Moses illusion 366–367
- interleaving illusion 317
- interpretation problem 150
- intrusion of reality in illusions of control 130
- intuitions: about randomness 193–194; about sample sizes 194–195
- intuitive physics 5
- involuntary cognitive illusions 7–8
- irrationality and positive emotions 350–351
- joint probabilities 39–41
- Journal of Memory and Language* 5–6
- judgmental bias, conjunction fallacy as 33–38
- judgment(s) 3, 4; causality 199–200; conjunction fallacy and quasi-random adjustments to 39–41; expert 199; feeling of knowing (FOK) 308, 311–312; heuristics of 4–5; of learning (JOLs) 308–311; metacognitive 308–309; with missing knowledge 11; by representativeness (*see* representativeness); of taste and odor 393–394; under uncertainty 8–9
- just-world hypothesis 317
- kind environments 180
- knowing what needs to be remembered 15
- knowledge, missing 11, 14
- labeling and overshadowing effects 386–395; applied domains of 386–387; classroom experiment demonstrating 388–390; examples of 386; judgments of taste and odor and 393–394; memory for color and 391–392; memory for speed and 392–393; memory for visual objects and 387–391; in other areas of judgment 394–395; theoretical accounts of 387–395
- lack of meaning 14–15
- language, pragmatics of 73
- Late Positive Component (LPC) 431
- lawyer-engineer problem 46, 47–49, 49, 52
- letter-frequency experiment 178–181
- liberalism 327–328
- logical intuitions 163–164, 169
- logic and mental arithmetic 5
- malleability of judgment 185
- mammography problem 46, 51
- matching, Moses illusion and imperfect 364–365
- matching bias 142–146
- meaning: lack of 14–15; memory shaped to extract 12; subjective uncertainty and 30
- medical decision-making 5, 16
- memory 3, 4; for color 391–392; errors of 5, 6; for faces 395–396; false 429–432; illusions of 4, 5–6; reconstructive process of 419–420; research on 5–6; for speed 392–393; for visual objects 387–391
- Memory distortions: How minds, brains, and societies reconstruct the past* 5
- Memory illusion: Remembering, forgetting, and the science of false memory, The* 6
- mere exposure effect (MEE) 241–256; affective primacy model of 254; classroom demonstration of 248–249, 250; examples of 241–242, 250–251; in exposure and preference for novel types of art 242, 250–251, 252; in liking for music 241–242; moderating variables of 244–248; neurological correlates in 251–252; nonspecific activation model of 253;

- perceptual fluency/attributional model of 253–254; research methods on 242–244; theoretical accounts of 252–254; two-factor model of 253; unfamiliar people and 242
- metacognition 307; assessment of 308–309; direct-access versus inferential accounts of 314–315; hindsight bias and 446–447; theory- and experience-based processes of 315–317
- metacognitive illusions 307–320; defined 307; of knowledge assessment 311–312; of learning 310–311; mending 318–319; overview of 309, 310; in real life 308; real-life consequences of 317–318; theoretical accounts of 314–317; of thinking 312–313
- metacognitive monitoring 307, 317–319
- metacomprehension judgments 308
- mindware gaps 14
- misinformation effect 419–432; consequences of 427–428; defined 419; demonstration of 421–423; distinguished from false memories 429–430; early research on 420–421; encoding, storage, and retrieval of memories and 419–420; moderators of 425–427; neuroscientific evidence of 430–431; possible mechanisms of 428–429; theoretical models of 423–425
- misleading information 10
- missing knowledge 11, 14
- mood in illusions of control 130
- Moses illusion 359–369; classroom demonstration of 361–363; cooperative communication setting and 360–361; expertise and 367; explanation of 360–365; imperfect encoding and 363–364; imperfect matching and 364–365; imperfect retrieval and 364; implications of 368; interindividual differences and 366–367; moderators of 365–368; motivation and 367–368; paradigm of 359–360; semantic relatedness and 365; situational manipulations and 366; statements instead of questions and 365–366
- motivation: and emotional accounts of framing 71–72; hindsight bias and 447; Moses illusion and 367–368; overconfidence and 300; processing 185–186
- motivational biases 14
- myside bias 85; debiasing or making the best of 89; demonstration of 78–86; explanation of 87–89; limits of 86–87; *see also* confirmation bias
- Need for Cognitive Closure (NCC) 234
- need or desire for the outcome in illusions of control 129–130
- need to act fast 15
- negative consequences of illusions of control 133–135
- negativity bias 341–343, **342**
- Neighborhood Activation Model 411
- neural mechanisms: associative memory illusions 414–415; DRM paradigm 414–415; illusion of control 135–136; mere exposure effect (MEE) 251–252; misinformation effect 430–431
- new paradigms, development of 166, 167–168
- noise, random 38–39
- non-regressive predictions 198
- nonspecific activation model of mere exposure effect (MEE) 253
- normative-system problem 149
- norms 11, 12
- numeric priming and anchoring effect 216–217
- objective uncertainty 30
- obsessive-compulsive disorder (OCD) 16
- old and new paradigms 166, 167–168
- “On cognitive illusions and their implications” 5
- one-shot interventions 16
- Online-Tool for Assessing Perceiver Effects (O-TAPE) 275
- optical illusions 7, 13
- optimism 347–348; overconfidence and 290; unrealistic 182
- overconfidence 287–304; base-rate effect and 294; calibration and 287–288, 291–292; classroom demonstrations of 290–293; combined error models of 297; confidence-frequency effect and 294; convincing others with 300–301; defined 287; ecological models of 295–296; in entrepreneurship and market entry 301–302; error models of 296–297; expertise effect and 294; feeling good with 299–300; in finance and trading 301; format dependence and 294–295; functions of 299–301; hard-easy effect and 293; heuristics and biases and 295; in management and leadership 302–303; metacognitive judgments and 308–309; as motivation 300; overconfidence effect and 293; rational beliefs based on imperfect measurement and 298; relevance in applied settings 301–303; sampling procedure and 293–294; sensory sampling model of 297–298; subjective confidence intervals and 289–290; subjective multidimensional assessment and 299; theoretical accounts of 295–299; types, tasks, and measures of 287–288; typical findings on 293–295; underconfidence and error independence in sensory-discrimination tasks and 295
- override failure 14
- overshadowing effects *see* labeling and overshadowing effects
- Oxford handbook of memory* 6

- paradigms: anchoring effect 213–214; old versus new 166, 167–168; truth effect 226, 226–227
- paradox of rationality 149
- partial matching and Moses illusion 364–365
- participatory propaganda 324–335; applied implications of 333–335; conflict in perpetuity and 332; coveted authentic users of 331–332; demonstration experiment 332–333; digital tactics of strategic actors and 328–329; hallmarks of 330–331; involvement of the media in 329; obscuring origin and existence 331; when expressive responding trumps perception in 330; *see also* fake news
- perception, illusions of 6
- perceptual fluency/attributional model of mere exposure effect (MEE) 253–254
- pervasiveness of anchoring effect 211–212
- physical attractiveness halo effects **262–263**, 262–265, 265
- Pizzagate 324–325
- point of subjective equality (PSE) 391–392
- positive illusions 133; Fading Affect Bias (FAB) and 348–350; helping people attain life goals 343–346; reconciling negative and 135; self-protection and self-esteem with 346–348
- positive-testing bias 141
- positivity bias 144, 341–352; versus negativity bias of depression, anxiety, and anger misleading the mind 341–343, **342**; success and failure with 347
- post-traumatic stress disorder (PTSD) 427, 428
- potential surprise 39–41
- power in illusions of control 130–131
- pragmatic reasoning schemas 148
- pragmatics of language 73
- prediction and representativeness 195–197
- probabilistic mental models (PMMs) 295–298
- probability: applying different probabilistic rule to 31–32; conversational implicature and 31; versus frequencies 202; frequentist interpretations of 31; joint, conjunction fallacy and 39–41; laws of 30–31; not applying where there is subjective uncertainty 30–31; quantum 32–33, 33
- probability assessments and revision 5
- probability theory 11; violations of rules of 27–28
- processing motivation and opportunities 185–186
- profile-based approach and assumed similarity 275
- pronounceability illusion 313
- prospect-theory 64, 70–71
- proximate explanations of survival processing effect 375–380
- pseudocontingency heuristic 98
- pseudoscience 117–119, 118
- psychography 5
- qualitative likelihood index (QLI) 34
- quantum probability of the conjunction fallacy 32–33, 33
- quasi-random adjustment and conjunction fallacy 39–41
- random sequences: diagnosing 196–197; intuitions about 193–194
- rationality: bias in human reasoning and 149–151; overconfidence and 298
- realistic selection task 145–149
- reality monitoring 413
- Reconstruction After Feedback with Take-the-best (RAFT) 446
- referential theory of truth effect 230
- regulatory-focus theory 71–72; in illusions of control 131
- relational processing in survival processing effect 379–380
- representativeness 4, 34, 175, 191–206; alternative explanations for 203; biases as not irrational and 202–203; biases as not universal and 201–202; broadly defined 200–201; conceptual vagueness of 201; conjunctural fallacy and 34; criticisms of 201–203; demonstrations of 192–195, 204–205; difference between availability and 179; as general-purpose heuristic 197–201; in intuitions about random sequences 193–194; in intuitions about sample sizes 194–195; in prediction versus diagnosis 195–197; probabilities versus frequencies and 202; wider framework for 203–204
- Rescorla–Wagner model 113, 113
- resolution 309
- retrieval of information: biased 180–181; expectations at 413–414; fluency in 186–187; Moses illusion and imperfect 364
- risky-choice framing 63–64
- robustness of anchoring effect 211–212
- round-robin designs 274
- rule-discovery task 10, 79–80
- sample size: diagnoses based on 197; intuitions about 194–195
- sampling errors 38–39; artificial biases and 10–11; conjunction fallacy 38–39; overconfidence and 293–294
- scale distortion in anchoring 219–220
- screen inferiority 311
- search-fluency illusion 311
- selective accessibility model of anchoring 217–219
- Selective Activation and Reconstructive Anchoring (SARA) 446
- Selective Processing Model 162, 169
- Selective Scrutiny Model 161, 162
- self-anchoring 274

- self-based heuristic 273  
 semantic-fluency illusion 311  
 semantic relatedness and Moses illusion 365  
 sensory sampling model of overconfidence 297–298  
 sequential judgment paradigm 214  
 signed summation 34, 38  
 similarity, assumed *see* assumed similarity  
 similarity-based theories of associative memory illusions 409–412, 410  
 single-item and relational processing in survival processing effect 379–380  
 situational manipulations and Moses illusion 366 skewed base-rate distributions in illusory correlations 98  
 skill-related factors in illusions of control 128–129  
 social media 324–325; human attention and misinformation on 325; information load, cognitive capacity, and decision quality and 326–327; online reputation and sharing on 325–326  
 social projection 273  
 Social Relations Model (SRM) 275  
 solvability, judgments of 308  
 source dissociations 229  
 source memory and conjunction fallacy 37  
 source of activation confusion (SAC) 425  
 spacing illusion 317  
 speed, labeling effects and memory for 392–393  
 spurious similarity 274  
 stability bias 310–311  
 statements versus questions and Moses illusion 365–366  
 stimulus-size illusion 316–317  
 subjective uncertainty 30–31  
 success or failure emphasis in illusions of control 129  
 suppositional theory of the conditional 150  
 survival processing effect 371–381; affective explanations of 375–377; characteristics of 371–373, 372; classroom experiment demonstration of 373–374; item congruity and 377; planning and 375; possible mechanisms underlying 374–380; proximate explanations of 375–380; richness of encoding and 378–379; single-item and relational processing 379–380  
 syllogistic reasoning 156–158; belief bias in 158, 158–161; dual processes and 161–162; logical intuitions and 163–164  
 systematic fashion of cognitive illusions 7  
 systematic framing effect 61; *see also* framing taste and odor: labeling effects and judgments of 393–394; verbal overshadowing on 396–397 thinking 3–4; heuristics of 4–5; illusions of 3–4  
 thinking fast-and-slow 204  
 thought listing 84–86  
 trait-based approach and assumed similarity 275  
 true illusions 11  
 truth effect 225–237; classical paradigm of 226, 226; classroom demonstration of 228; context characteristics of 232–233; debiasing and 235–236; experimental designs and measures of 226, 226–227; exposure paradigm of 227; familiarity account of 228–229; fluency account of 229–230; integrative model of 230–231; memory paradigm of 227; moderators of 231–234; under naturalistic conditions 234–235; person characteristics in 233–234; real-world examples 225; referential theory of 230; relevance of 225; statement characteristics of 231–232; theoretical accounts of 228–231  
 Twitter *see* social media 2–4–6 problem 140–141  
 two-factor model of mere exposure effect (MEE) 253  
 two-response paradigm 164  
 Type 1 processes 144–145  
 Type 2 thinking 144–145  
 uncertainty, subjective 30–31  
 underconfidence 295, 308–309  
 underconfidence-with-practice effect 310  
 unequal sample size and illusory correlations 97–98  
 unequal weighting of present versus absent information 96–97  
 uninformative-picture illusion 312  
 universal nature of cognitive illusions 8  
 unrealistic optimism 182  
 validity, illusion of 198  
 verbal overshadowing 395–399; on face recognition 395–396; on other senses 396–397; on other visual materials 396; theoretical accounts of 397–399  
 visual objects, labeling effects and memory for 387–391  
 vividness of information 180, 181, 312  
 von-Restorff effect 97  
 wait-generate-validate strategy 319  
 Wason selection task 83–84, 140–151; abstract 141–145; bias and rationality in human reasoning and 149–151; dual-process theory and 144–145; realistic 145–149; Wason's early thinking and the 2–4–6 problem and 140–141  
 wicked environments 180  
 wrong normative standards 11



Taylor & Francis Group  
an informa business

# Taylor & Francis eBooks

[www.taylorfrancis.com](http://www.taylorfrancis.com)

A single destination for eBooks from Taylor & Francis with increased functionality and an improved user experience to meet the needs of our customers.

90,000+ eBooks of award-winning academic content in Humanities, Social Science, Science, Technology, Engineering, and Medical written by a global network of editors and authors.

## TAYLOR & FRANCIS EBOOKS OFFERS:

A streamlined experience for our library customers

A single point of discovery for all of our eBook content

Improved search and discovery of content at both book and chapter level

**REQUEST A FREE TRIAL**  
[support@taylorfrancis.com](mailto:support@taylorfrancis.com)

 Routledge  
Taylor & Francis Group

 CRC Press  
Taylor & Francis Group