frontiers | Research Topics

Highlights in psychology: Cognitive bias

Edited by
Sergio Da Silva, Rashmi Gupta and Dario Monzani

Published in
Frontiers in Psychology
Frontiers in Psychiatry

ISSN 1664-8714
ISBN 978-2-8325-3022-1
DOI 10.3389/978-2-8325-3022-1

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Highlights in psychology: Cognitive bias

Topic editors

Sergio Da Silva — Federal University of Santa Catarina, Brazil
Rashmi Gupta — Indian Institute of Technology Bombay, India
Dario Monzani — University of Palermo, Italy

Citation

Da Silva, S., Gupta, R., Monzani, D., eds. (2023). *Highlights in psychology: Cognitive bias*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3022-1

Table of contents

frontiers | Frontiers in Psychology

TYPE Editorial
PUBLISHED 03 July 2023
DOI 10.3389/fpsyg.2023.1242809

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Eddy J. Davelaar,
Birkbeck, University of London, United Kingdom

*CORRESPONDENCE
Sergio Da Silva
✉ professorsergiodasilva@gmail.com

RECEIVED 19 June 2023
ACCEPTED 22 June 2023
PUBLISHED 03 July 2023

CITATION
Da Silva S, Gupta R and Monzani D (2023) Editorial: Highlights in psychology: cognitive bias. *Front. Psychol.* 14:1242809.
doi: 10.3389/fpsyg.2023.1242809

Editorial: Highlights in psychology: cognitive bias

Sergio Da Silva[1]*, Rashmi Gupta[2] and Dario Monzani[3]

[1]Department of Economics, Federal University of Santa Catarina, Florianopolis, Brazil, [2]Cognitive and Behavioural Neuroscience Laboratory, Indian Institute of Technology Bombay, Mumbai, India, [3]Department of Psychology, University of Palermo, Palermo, Italy

KEYWORDS

cognitive bias, cognitive biases and heuristics, cognitive psychology, social psychology, psychopathology

Editorial on the Research Topic
Highlights in psychology: cognitive bias

Cognitive biases are unconscious and systematic errors in thinking that occur when people process and interpret information in their surroundings and influence their decisions and judgments (Kahneman et al., 1982). These biases can distort an individual's perception of reality, resulting in inaccurate information interpretation and rationally bounded decision-making (Kahneman, 2011). Cognitive biases may also contribute to psychotic symptoms (Garety et al., 2007). This Research Topic brings together 13 articles that address these issues.

Two papers are reviews. In the first, Berthet and de Gardelle conduct a systematic review of heuristics and biases tasks that measure individual differences and reliability in this Research Topic and provide a heuristics and biases inventory, an open-source catalog of over 40 previously published individual difference measures. This is useful because it takes time to find measures and determine their reliability. The second review on this Research Topic by Liu et al. is about negative-biased implicit memory. According to the literature, patients with current major depressive disorder have abnormal implicit memory. However, its function in current and remitted major depressive disorder patients when processing stimuli with positive, neutral, and negative emotions is unknown. The authors review and elaborate on the role of implicit memory in these patients found in meta-analyses in the Web of Science, PubMed, and EMBASE databases between 1990 and 2022. They report a general deficit in implicit memory in current patients. Furthermore, current patients' implicit memory performance to neutral stimuli is lower than controls', but recovered in remitted patients. Furthermore, both current and remitted patients have an implicit memory deficit to positive stimuli, and the implicit memory response to negative stimuli in current patients is similar to controls but worse in remitted patients. As a result, current patients' negative bias compensates for the general implicit memory deficit. With remission, the implicit memory of neutral stimuli recovers, but it remains abnormal in processing positive and negative stimuli. Therefore, abnormal implicit memory of positive and negative stimuli is relevant to the pathogenesis of depression.

There are four works in cognitive psychology. (1) The study on this Research Topic by Melnik-Leroy et al. is about the intriguing exponential bias, which is the tendency to underestimate exponential growth systematically and perceive it in linear terms. Attempts to reduce this bias in graphical representations using a logarithmic scale rather than a linear scale produce more perceptual errors. The authors show that the log scale induces more errors in graph description tasks, whereas the linear scale misleads people when predicting the future trajectory of exponential growth. However, a brief mathematical educational intervention can mitigate both scales' difficulties. (2) Polyanskaya's article on this Research Topic addresses the overconfidence bias or awareness of what one knows vs. does not know. In particular, it focuses on individuals' ability to monitor their cognitive performance and decisions. Retrospective confidence ratings are used to assess metacognitive monitoring, in which individuals are asked to report how certain they are in response or their performance in high-level cognitive or low-level perceptual tasks. Polyanskaya contends that the reliability of this measure is affected by factors such as what is being evaluated, how the confidence response is elicited, and the overall proportion of different trial types within one experimental session. It is important to consider how questions are posed and whether individuals are asked to evaluate what they know rather than what they do not know. When individuals are asked to assess positive evidence and the absence of positive evidence, retrospective confidence ratings are unreliable. (3) People frequently misestimate the probability of an event based on uncertain evidence. Various explanations for these judgment errors have been proposed. Some studies attribute the errors to underweighting the event's base rate or overweighting the evidence for the individual event. The paper on this Research Topic by Branch and Hegdé examines the contributions of potential explanatory variables to probability judgments under four different problem scenarios. They discovered that the explanatory variables accounted for ∼30–45% of the overall variance of responses, depending on the problem scenario. No single factor can explain more than 53% of the explainable variance, let alone all of it. They conclude that attributing probabilistic judgment errors to any cause, including base rate neglect, is statistically untenable. A more nuanced explanation is that actual biases result from a weighted combination of multiple contributing factors, the exact mix of which depends on the problem scenario. (4) In this Research Topic, Suomala and Kauttonen define computational meaningfulness as the ability of humans to make a situation understandable to respond optimally. Computational meaningfulness takes into account multidimensional and changing settings. As a result, computational meaningfulness should moderate biases. Using the confirmation bias and the framing effect as examples, the authors argue that computational meaningfulness implies that these biases are necessary for optimal decision-making and can thus be deemed rational from this standpoint. The authors propose using naturalistic stimuli, such as vignettes, to build more realistic decision-making study environments and evaluate the resulting data with machine learning to improve behavior modeling.

Four articles on this Research Topic are concerned with social psychology. (1) Meng and Feng's paper on this Research Topic investigates the optimistic bias in young online taxi users who must choose between convenience and privacy in digital travel platforms. Using a model of protective motivation theory, the authors investigate the moderating effect of user knowledge of privacy settings on the relationship between privacy concerns and protective behavior. They discovered that increased privacy-protective behavior is associated with privacy concerns and positively related to perceived threats, self-efficacy, and response efficacy. As a result, there is an optimistic bias in privacy management. Previous research on the impact of intuitive-deliberative cognitive style and risk style on risky choice framing has yielded conflicting results. (2) Wyszynski and Diederich consider a psychophysical data collection approach in this Research Topic and discover that framing effects strength, cognitive style, and risk style are related. They conduct two studies, one of which counts the number of frame-inconsistent choices, and the other compares the proportions of risky choices on gain-loss frames. They vary the number of people affected, the chances of surviving or dying from an unusual disease, the type of disease, and the response deadlines. They find that risk style moderates the framing effect on the proportion of risky choices, while cognitive style in one of the studies moderates the framing effect. However, they find no link between the number of frame-inconsistent choices and cognitive or risk styles. (3) The article on this Research Topic by Korteling et al. contends that perceptions of sustainability problems, such as climate change, do not lead to sustainable choices due to cognitive biases. They list social-psychological dimensions common to most sustainability issues, such as experiential ambiguity, long-term consequences, complexity and uncertainty, threats to the status quo and social status, social dilemmas, and group pressure. They match corresponding cognitive biases using a neuro-evolutionary perspective for each characteristic. These are evolved biases that influence our preferences and behavior. They then propose interventions such as incentives and nudges to help people make more sustainable choices. (4) Naive realism is the tendency to believe that our subjective experience of reality is objective and that others should naturally perceive the world as we do. The bias that others share one's knowledge entails the curse of knowledge, which is the inability to fathom the reasoning of those who do not share one's knowledge. In this Research Topic, Beattie and Beattie apply the curse of knowledge to political cognition. They argue that overestimating the knowledge of political opponents is associated with more negative evaluations. As a result, opponents take on the character of someone who understands why an opposing viewpoint is correct but continues to oppose it. This results in political polarization. Fortunately, in a debiasing experiment, the authors discover that participants who receive an epistemological treatment evaluate those with whom they disagree more favorably.

The final three papers focus on psychopathology. (1) The paper on this Research Topic by Blauth and Iffland addresses the attentional biases associated with maltreatment and victimization experiences in childhood and adolescence. Using an online version of the facial dot-probe task, they investigate attentional processes for emotional facial expressions (anger, disgust, happiness, and sadness) in an adult sample. They discovered that attentional biases in child maltreatment are associated with angry facial expressions, which can be interpreted as threat-related biases. In contrast, biases in the context of peer victimization are related to sad facial expressions, indicating a mood-congruent bias. (2) Overcoming persistent negative thoughts is one obstacle in encouraging depressed individuals to seek assistance. In this Research Topic, Keeler et al. conducted two randomized pre-post trials to determine whether a novel online intervention employing mental contrasting and implementation intentions could increase actual help-seeking or the intent to seek help for depression. The study takes into account self-reports from individuals in the United States. These trials show the viability and preliminary success of such an intervention to encourage help-seeking, which may be helpful to clinicians. (3) According to the cognitive model of psychosis, psychotic symptoms may originate from biased information processing. The study in this Research Topic by Sanchez-gistau et al. focuses on the differences in selected cognitive biases (intentionalizing, catastrophizing, dichotomous thinking, jumping to conclusions, and emotional reasoning) between individuals experiencing first-episode psychosis (FEP) with and without comorbid attention deficit and hyperactivity disorder (ADHD). The researchers use the Cognitive Biases Questionnaire for Psychosis to assess the severity and types of cognitive biases in FEP-ADHD+, FEP-ADHD-, and healthy controls (HCs). According to the findings, FEP-ADHD+ participants have considerably greater cognitive biases than FEP-ADHD- individuals and HCs. In particular, the FEP-ADHD+ group is more strongly related to intentionalizing and emotional reasoning biases. Cognitive biases are associated with positive psychotic symptoms in both groups but only with depressive symptoms in the FEP-ADHD- group and impaired functioning in the FEP-ADHD+ group. These findings imply that FEP-ADHD+ individuals may require focused metacognitive interventions. The study emphasizes the necessity of treating FEP with ADHD and recommends more research to develop individualized pharmacological and psychological interventions for specific FEP subpopulations.

In conclusion, the publications on this Research Topic demonstrate the rich diversity of theoretical and empirical findings across a broad spectrum of contemporary cognitive bias research. Cognitive biases are significant because they can influence how individuals perceive and interpret information, which may or may not lead to judgment and decision-making errors. Researchers must comprehend cognitive biases to develop interventions designed to improve decision-making and mental health. We are grateful to all those who took the time to provide insightful input. We hope that their research will inspire future investigation.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Garety, P.A., Bebbington, P., Fowler, D., Freeman, D., and Kuipers, E. (2007). Implications for neurobiological research of cognitive models of psychosis: a theoretical paper. *Psychol. Med.* 37, 1377–1391. doi: 10.1017/S003329170700013X

Kahneman, (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D., Slovic, P., and Tversky, A. (eds.). (1982). *Judgment Under Uncertainty: Heuristics and Biases*. New York, NY: Cambridge University Press.

frontiers | Frontiers in Psychology

TYPE Methods
PUBLISHED 22 November 2022
DOI 10.3389/fpsyg.2022.1049925

Online taxi users' optimistic bias: China youths' digital travel and information privacy protection

Xiaoyang Meng and Bobo Feng*

School of Journalism and Communication, Southwest University of Political Science and Law, Chongqing, China

Digital travel platforms not only provided people with convenient travel but also raised a series of problems regarding information privacy protection. In order to analyze privacy protection behavior, this study surveyed 441 subjects aged 18–35 who utilized digital travel platforms based on a structural model of protective motivation theory. The results indicated that a perceived threat, self-efficacy, and response efficacy positively and significantly impacted youths' privacy concerns. Furthermore, privacy concerns were positively related to privacy protection behavior and were an intermediate variable between the relationships among perceived threat, self-efficacy, response efficacy, and privacy protection behavior. This study identified the moderating effect of youths' knowledge of platform privacy settings on the relationship between privacy concerns and protection behavior. In addition, the results confirmed that an optimistic bias did exist among talented youth with high privacy knowledge in terms of a practical level of privacy management. These unique findings represent the exceptional contributions and innovation points of this study.

KEYWORDS

privacy practice, privacy concern, privacy knowledge, optimistic bias, digital travel

OPEN ACCESS

EDITED BY
Sergio Da Silva,
Federal University of Santa Catarina, Brazil

REVIEWED BY
Doug Twitchell,
Boise State University, United States
Valentina Marinescu,
University of Bucharest, Romania

*CORRESPONDENCE
Bobo Feng
milk_bravo@163.com

SPECIALTY SECTION
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

RECEIVED 21 September 2022
ACCEPTED 03 November 2022
PUBLISHED 22 November 2022

CITATION
Meng X and Feng B (2022) Online taxi users' optimistic bias: China youths' digital travel and information privacy protection.
Front. Psychol. 13:1049925.
doi: 10.3389/fpsyg.2022.1049925

Introduction

Digital travel platforms (DTP) have gradually permeated our daily lives in various fields due to information technology's rapid development and evolution. According to a statistical report of the 49th China Internet Development Status, there were over 4.53 hundred million online taxi users in China, which accounted for 43.9% of Chinese Internet users. However, digital travel platforms not only provided them with convenient travel but also raised a series of problems regarding information privacy protection. Due to various incidents of serious illegal collection of people's personal information known to the public, the National Network Information Office officially shut down 25 digital travel platforms on July 4, 2021. The practical levels of this phenomenon illustrated that online taxi users' personal information was collected unreasonably and illegally, which reflected a tremendous threat to privacy loopholes. According to studies of digital travel platforms, youths account for the majority of customers of online taxis, and college students prefer carpooling. The data in DTP include users' ID card numbers, names, ages, and other information for privacy purposes such as audio/video records during the ride, facial IDs, travel routes, call logs, and sensitive sites. Leakage of information privacy would seriously damage an individual's personal safety, property, and dignity. Thus, this study focused on young users' attitudes toward online travel platforms' information privacy concerns and protection. This study aimed to explore privacy protection theories among youths and offer practical guidance regarding information privacy protection.

Few quantitative studies on youths' attitudes toward digital travel platform information privacy protection were identified, but they were very helpful. Protection motivation theory and social cognitive theory have become significant theories investigating the relationship between perceived online threats and online behavior (Milne et al., 2009). Studies argue that the enhancement of privacy concerns leads to an increase in protection behavior and a decrease in online privacy disclosure (Chen and Chen, 2015). From the perspective of privacy protection, youths' personal information safety behavior in social networks was significantly affected by their perceived threat, self-efficacy, and response efficacy (Wang et al., 2018). Certain studies showed that college students' perceived risk of the WeChat application process significantly triggered their privacy concerns in social networking (Shen, 2017), which illustrated their concerns about information privacy. In addition, the increase in online privacy concerns among youths directly affects their privacy protection behavior and disclosure of information privacy (Jia et al., 2021), which indicates that the privacy protection behavior of the youths was affected by multiple predisposing factors and variables. In contrast to the traditional privacy framework structure, a study asserted that privacy knowledge level was an intermediary factor in the relationship between privacy concerns and self-disclosure behavior (Qiang and Xiao, 2021). Based on current studies of privacy protection, we expanded privacy protection issues among youths to the information system of digital travel platforms. We intended to explore the relationships between protection motivation, privacy concern, and privacy protection behavior among young online taxi users. Thus, we listed the following research questions:

> Research Question 1: What is the status of protection motivation and privacy protection behavior among young people?
> Research Question 2: Does protection motivation affect privacy concerns and privacy protection behavior among young people?
> Research Question 3: What is the level of privacy knowledge of young users? Will it affect the relationship between privacy concerns and privacy protection behavior?

Literature review

Theories and hypotheses

Protection Motivation Theory (PMT) uses the social cognition perspective to examine an individual's behavior when faced with threats (Rogers, 1975). Following a series of research, PMT described its coping strategies in detail and categorized the motivation to self-protect from threats into two cognitive assessment processes: threat assessment (including perceived susceptibility and perceived severity) and coping assessment (including self-efficacy and response efficacy). Based on the assessment results of its cognitive threat, individuals may choose to engage in protection behavior (Rippetoe and Rogers, 1987). In terms of the PMT cognitive assessment processes, Witte argued that perceptual susceptibility and perceived severity described an individual's cognition of severity and possibility of a threatening occurrence, i.e., a perceived threat (Witte, 1992). Technology Threat Avoidance Theory (TTAT) further proposed that perceived threat was determined by predisposed variables of perceptual susceptibility and perceived severity. The perceived degree of potential threat initiated by technology would affect subjects' attitudes and behavior (Liang and Xue, 2009). This study combined PMT, TTAT, and other related research and intended to investigate the relationships between perceived threat, self-efficacy, response efficacy, privacy concern, and privacy protection behavior among youths who utilized DTP.

Perceived threat, self-efficacy, response efficacy, and privacy concern

The predisposed variables of privacy concern, i.e., perceived threat, self-efficacy, and response efficacy, were used to measure the information privacy concerns. "Perceived threat" was defined as an individual's expected negative consequence of a certain technique, product, or even behavior, which affects the desire and motivation to take protective behavior (De Zwart et al., 2009). Therefore, this study used it to measure the perceived threat to personal information privacy among youths who utilized digital travel platforms. Self-efficacy is defined as an individual's capability to carry out expected behavior. Bandura asserted that self-efficacy was the perceived belief in individuals' capability to organize and execute the action process of established achievements (Bandura, 1977). Self-efficacy was the core concept of social psychology, which illustrated the belief in individuals' ability to execute behaviors successfully, and was critical to the explanation of subjective motivation. This study used self-efficacy as an attribute of youths' capability and confidence in protecting personal privacy from intrusion. Response efficacy was identified as the perceptual ability to reduce the risk effectively. The higher the belief that individuals benefit from protective behavior, the greater the motivation for engaging in such behavior (Maddux and Rogers, 1983), and an adaptive response to engaging in such protective behavior is capable of protecting themselves and others (Hanus and Wu, 2016).

The terminology and concept of privacy concern gradually appeared in academic fields due to the rapid development of information technology, which raised the issue of privacy protection and related research. Culnan argued that when an individual releases personal information to a certain organization, the issue of privacy concerns arises regarding how it will use and protect the information (Culnan, 1993). Information privacy concern refers to an inherent worry of information privacy loss, which was often applied to the research of predicting users' privacy protection behavior (Smith et al., 1996). Privacy concerns echoed the awareness of how service providers collect, restore, and use personal information obtained from customers (Sheng et al., 2008). Previous studies revealed that the worry about information privacy leakage significantly influenced the attitude and behavior of social media platforms (Adhikari and Panda, 2018). In addition, studies delineated that potential variables of protection motivation, such as perceived threat, self-efficacy, and response efficacy, tended to affect an individual's information privacy concern. Youn identified the perception of threat as a decisive factor in the internet privacy concern among youths (Youn, 2009). According to an empirical study of users' self-disclosure on the socialized internet, the greater the perceived risk, the higher the privacy concern (Chen, 2013).

Self-efficacy was another significant predisposing factor of privacy concern, which predicted the intention of taking protective behavior. Another study of accurate advertising push and consumers' privacy concerns found a positive correlation between self-efficacy in preventing privacy leakage from accurate advertising and privacy concern (Yu and Yang, 2019). Finally, a medical big data cloud study confirmed the significant positive relationship between self-efficacy and privacy concerns (Wu, 2020). Based on the above evidence, the following hypotheses were proposed:

H1. Self-efficacy has a positive influence on personal information privacy concerns.
H2. Response efficacy has a positive influence on personal information privacy concerns.
H3. A perceived threat has a positive influence on personal information privacy concerns.

Privacy concern and privacy protection behavior

An empirical study of internet fraud confirmed that an increase in victims' predicted online privacy concerns tended to amplify privacy protection behavior (Chen et al., 2017). In addition, a related study of privacy protection delineated that users of socialized media tended to employ various modes of privacy protection behavior due to a high level of privacy concern (Feng and Xie, 2014). A similar Singapore study based on broadened planned behavior theory also found that the level of privacy doubt magnified the intention of online privacy protection (Ho et al., 2017). Another study on college students' privacy protection behavior verified that their privacy concerns about the WeChat APP influenced their privacy protection behavior significantly and positively (Xie and Karan, 2019). In order to examine the relationship between information privacy concerns and privacy protection behavior among youths, the following hypothesis was proposed:

H4. Privacy concern has a positive influence on privacy protection behavior.

Indirect effect of privacy concern

In order to explore the predisposing factors of youths' privacy concerns, which affect privacy protection behavior among socialized internet users, a pragmatic study demonstrated that an indirect effect of privacy concern did exist in the relationships between perceived threat, self-efficacy, and privacy protection behavior (Hanus and Wu, 2016). Another study on the privacy protection behavior among Sina MicroBlog users also verified this indirect effect between perceived threat and privacy protection behavior; however, no indirect effect was found between the relationships of self-efficacy/response efficacy and privacy safety protection behavior (Wang et al., 2019). In addition, a Malaysian study of young socialized media users validated that perceived threat, self-efficacy, and response efficacy indirectly affect privacy protection behavior through privacy concerns (Adhikari and Panda, 2018). In order to verify the indirect effect of privacy concerns, the following hypotheses were proposed:

H5. Privacy concern mediates the relationship between perceived threats and privacy protection behavior.
H6. Privacy concern mediates the relationship between self-efficacy and privacy protection behavior.
H7. Privacy concern mediates the relationship between response efficacy and privacy protection behavior.

Moderating effect of privacy knowledge

Privacy knowledge is a latent variable that could be flexibly elevated with refinement and training, thus reflecting its moderating characteristic. The results of a quasi-experimental study on the development of intelligent mobile phone APP software for privacy knowledge showed that APP users tended to pay more attention to their private personal information and use active protection means (Gerber et al., 2018). Similar research on children's digital literacy training revealed that the boost in training cost led to a decline in their personal information disclosure, which means children paid more attention to protecting their personal information privacy after training and tended to acquire protective actions (Desimpelaere et al., 2020). Knowledge regulated the relationship between privacy concerns and privacy protection behavior to a certain degree. How do young online taxi users comprehend the extent of privacy and information safety settings in the digital travel software they are using in their daily lives? Will it affect their protective manners? In order to verify these questions, the following hypotheses were proposed:

> H8a. Privacy knowledge moderates the relationship between privacy concerns and privacy protection behavior.
> H8b. Privacy knowledge groups moderate the relationship between privacy concerns and privacy protection behavior.

Figure 1 summarizes the research model of the study.

Research design

Data collection and implementation

According to the regulation of *the "Medium and Long Term Youth Development Plan (2016-2035)"* released by the CPC Central Committee and the State Council, these study subjects were limited to Chinese youths aged 18 to 35 who employ DTP. Questionnaire Star was utilized to sketch the questionnaire and distribute it *via* WeChat Moments on August 17 and 30, 2021. A total of 507 subjects responded to the survey, excluding 66 invalid subjects and responses. A total of 441 subjects remained, with a sample qualification rate of 86.9%. This study adopted SPSS v23.0 for descriptive analyses, and AMOS v23.0 was used for confirmatory factor analyses and research hypotheses testing.

This study consisted of six dimensions, i.e., perceived threat, self-efficacy, response efficacy, privacy concern, privacy protection behavior, and privacy knowledge. Except for privacy, knowledge was segregated by dichotomized categories (yes, no, don't know), and a Likert 7-point scale was used for measuring the other variables (1 = totally disagree, 7 = totally agree). In order to ensure the reliability and validity of the questionnaire, a small-scale pilot test was conducted, and the tested subjects' opinions on questioning, sentencing, and wording were for modifications. In addition, several experts and scholars were invited for content validity checks and revision. The final version of the questionnaire consisted of six dimensions and 27 measurement indicators. The Appendix A shows the detailed questionnaire measurement items. The structure of the survey is shown in Table 1.

Statistical analysis and hypothesis test

Descriptive analysis

Female respondents accounted for 61.2 vs. 38.8% of males. Regarding age allocation, respondents aged 18–25 accounted for 46.7%, 26–30 34.9%, and 31–35 18.4%. The majority of respondents were students (39.9%), enterprise employees (35.1%), personnel of public institutions, and other occupations accounted for 25%. Education level of an undergraduate degree accounted for the majority of 49%. Regarding monthly income, 69.8% reported less than 8,000 RMB, and 30.2% over 8,000 RMB.

SEM-AMOS was used for the confirmatory factor analysis of the research model. All standardized factor loadings (STD) were greater than 0.6, Cronbach's α and composite reliability (CR) were higher than 0.7, and the convergence effect (AVE) was higher than 0.5, which illustrated the excellent reliability and validity of the research model. In addition, Table 2 identified all the AVE square roots as being greater than the correlation coefficients between the variables, which indicated outstanding discriminant validity among the variables. The Appendix B shows the detailed measurement model reliability.

Structural model

Based on the calculations of AMOS, related indices of model fitness were as follows: Normed Chi-square (χ^2/DF) = 2.947, GFI = 0.902, NFI = 0.922, IFI = 0.947, TLI (NNFI) = 0.937, CFI = 0.947, RMSEA = 0.067. All the indexes were in a reasonable range, which confirmed that the fitness of the research model was acceptable.

Path analysis and hypothesis test

Figure 2 illustrates the regression coefficients as follows: perceived threat ($\beta = 0.533$, $p < 0.001$), self-efficacy ($\beta = 0.144$, $p < 0.05$), and response efficacy ($\beta = 0.150$, $p < 0.05$), which significantly affect privacy concern (R^2=0.441). In addition, privacy concern ($\beta = 0.586$, p< 0.001) significantly affects privacy protection behavior ($R^2 = 0.344$). Therefore, hypotheses 1 4 were accepted to various degrees.

Indirect effect of privacy concern

Bootstrapping 5,000 times was utilized to check the indirect effect, with the bias-corrected 95% CI and percentile 95% CI not including 0. Table 3 delineated the significant total effect and total indirect effect of perceived threat, self-efficacy, and response efficacy on privacy protection behavior ($p < 0.05$),

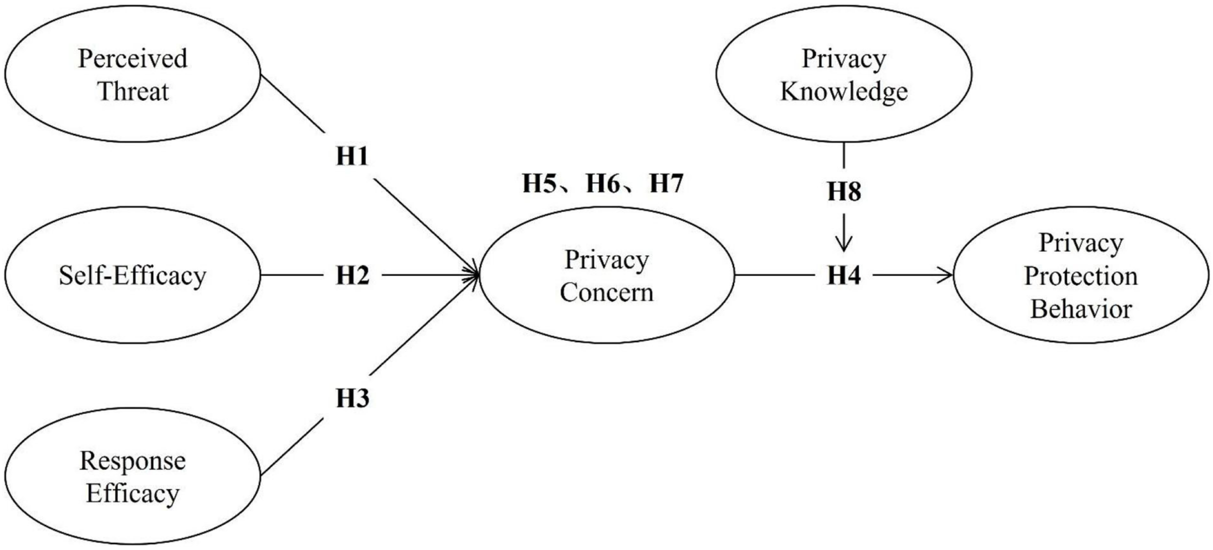


FIGURE 1
The framework of the research model.

TABLE 1 Research variables and sources.

Variables	Sources	No. of items
Perceived threat	(Johnston and Warkentin, 2010; Qi and Li, 2018)	3
Self-efficacy	(Schwarzer et al., 1999; Youn, 2009)	5
Response efficacy	(Workman et al., 2008)	3
Privacy concern	(Taylor et al., 2009; Adhikari and Panda, 2018)	4
Privacy protection behavior	(Hanus and Wu, 2016)	4
Privacy knowledge	(Park and Jang, 2014; Masur et al., 2017; Rosenthal et al., 2020)	8

TABLE 2 Reliability, convergent and discriminant validities of the research model.

Variable	FL	CR	AVE	PT	SEEF	REEF	PC	PPB
PT	0.733∼0.857	0.694	0.639	**0.799**				
SEEF	0.765∼0.880	0.700	0.694	0.254	**0.833**			
REEF	0.775∼0.881	0.639	0.700	0.268	0.745	**0.837**		
PC	0.824∼0.887	0.705	0.705	0.603	0.366	0.379	**0.840**	
PPB	0.627∼0.796	0.525	0.525	0.428	0.563	0.499	0.560	**0.725**

FL, factor loadings; CR, composite reliability; AVE, average variance extracted. SEEF, self-efficacy; REEF, response efficacy; PT, perceived threat; PC, privacy concern; PPB, privacy protection behavior.

which confirmed the indirect effect of privacy concern. Thus, hypothesis 5/6/7 were supported.

Moderating effect of privacy knowledge

The moderating effect of privacy knowledge was one of the key interpretations of this study. The following specific procedures followed:

Step 1: The scores of eight items were summed up for a total score. All the total scores were divided into three groups: high (top 27 percentile), medium, and low (bottom 27 percentile) scores, based on Cureton (1957) proposal. In order to maintain statistical power, the difference only between high- and low-score groups (120 subjects each) was calculated. The independent *t*-test identified significant differences between high- and low-privacy knowledge level groups ($t = -30.933$).

Step 2: Grouping regression and identity tests were conducted using AMOS software. In order to examine the significant difference, the combination of high- and low-score groups was designated as the constraint model (all the parameters are equal) and compared with the default model

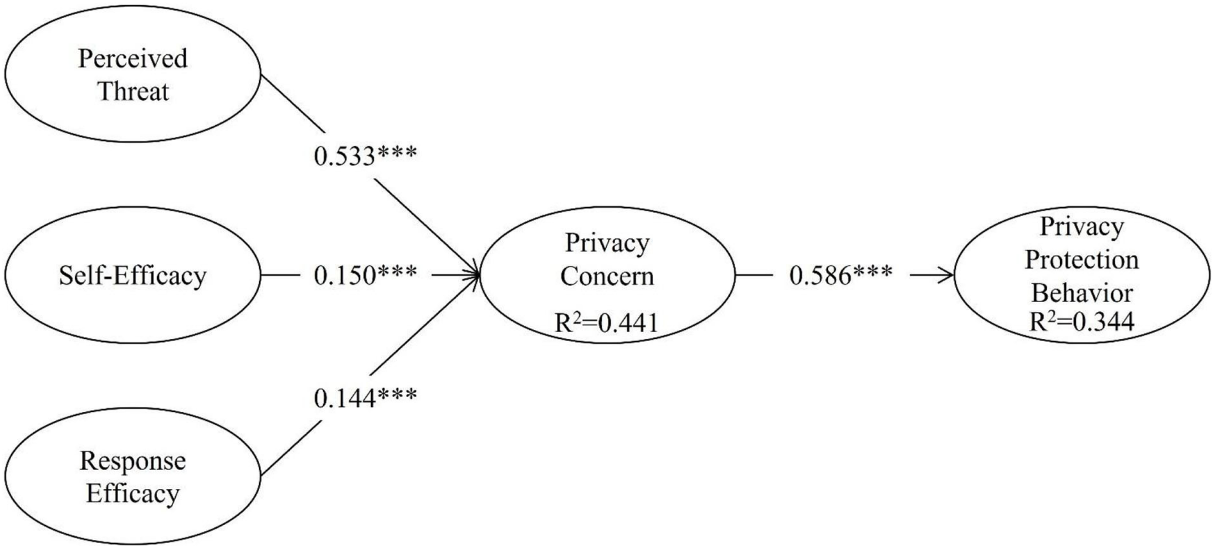


FIGURE 2
Path coefficients of the structural equation model. *$p < 0.05$; **$p < 0.01$, ***$p < 0.001$.

TABLE 3 Indirect effects of privacy concerns.

Hypothesis	Effect	P. E.	C.P.			Bias-corrected 95% CI		Percentile 95% CI		Result
			SE	Z	P	LL	UL	LL	UL	
H5	Total effect SEEF→ PPB	0.200	0.042	4.762	0.000	0.127	0.291	0.121	0.285	Accept support
	TIE SEEF→ PPB	0.200	0.042	4.762	0.000	0.127	0.291	0.121	0.285	
H6	Total effect REEF→ PPB	0.223	0.054	4.130	0.000	0.131	0.342	0.125	0.337	Accept support
	TIE REEF→ PPB	0.223	0.054	4.130	0.000	0.131	0.342	0.125	0.337	
H7	Total effect PT→ PPB	0.452	0.073	6.192	0.000	0.311	0.600	0.311	0.600	Accept support
	TIE PT→ PPB	0.452	0.073	6.192	0.000	0.311	0.600	0.311	0.600	

(without any restriction) (Wen et al., 2012). Table 4 delineated significant results of grouping regression: the Chi-square value change of the constraint model ($\chi^2_{95\%,1df} = 8.941 > 3.84$) with $p = 0.003$, which concluded that significant privacy knowledge moderates the relationship between privacy concern and privacy protection behavior. In order to consolidate the credibility of the findings, the following outcomes were identified:

The *p* values of both models were less than 0.001, and CMIN/*df* values were < 3.

The baseline comparison found significant differences in the NFI, RFI, IFI, TLI, and CFI values.

RMSEA indexes of the models were unequal (0.059 vs. 0.061).

Thus, the default and constraint models were not matched, i.e., hypothesis 8a should be accepted.

Step 3: Data from grouping regression demonstrated greater mean values of privacy concern and privacy protection behavior in the high privacy knowledge group than that of the low privacy knowledge group, with regression coefficients of 0.371 (high privacy knowledge group) vs. 0.620 (low privacy knowledge group), which means the impact of the moderating effect among the high privacy knowledge group was significantly lower than that of their counterparts (data not shown in Table 4). Therefore, hypothesis 8b should be rejected.

Conclusion and discussion

Conclusion

This study expanded privacy protection theory and context to digital travel platforms that youths employ in their daily lives, work, and social contact. Based on a comprehensive of understanding the privacy protection behavior of contemporary youth online taxi users, this study offered coping strategies from subjective and objective dimensions of youths' privacy

TABLE 4 Grouping regression of the constraint model vs. the default model.

Model	NPAR	CMIN	*df*	*P*	CMIN/*df*
Default	51	144.159	57	<0.001	2.529
Constraint	50	153.100	58	<0.001	2.640
	NFI	**RFI**	**IFI**	**TLI**	**CFI**
Default	0.926	0.891	0.954	0.931	0.953
Constraint	0.921	0.886	0.950	0.926	0.949
	RMSEA	**Lo 90**		**Hi 90**	**PCLOSE**
Default	0.059	0.047		0.071	0.102
Constraint	0.061	0.049		0.073	0.057

protection and hoped digital society could protect the personal information and privacy of youths. The conclusions are as follows:

A perceived threat, self-efficacy, and response efficacy positively affected privacy concerns;

Privacy concerns positively affected privacy protection behavior. Youths tended to have a higher level of privacy concern (with a mean value of 5.187 over 7) and used countermeasures to protect their privacy, such as fake names and shutting off location services;

Privacy concern was an intermediate factor in the relationships between perceived threat, self-efficacy, response efficacy, and privacy protection behavior;

Privacy knowledge moderates the relationship between privacy concerns and privacy protection behavior. The mean values of privacy concern and privacy protection behavior in the high privacy knowledge group were significantly greater than those of their counterparts. However, the predictive power of privacy concern on privacy protection behavior in the high privacy knowledge group was significantly less than that of their counterparts.

Discussion

Perceived threat, self-efficacy, and response efficacy were significant variables in predicting the relationship between privacy concerns and privacy protection behavior among youths utilizing DTP. Of which, the perceived threat was identified as the main predictive factor of privacy concern, followed by response efficacy and self-efficacy. In addition, the mean values of these variables were greater than their average scores, which denoted that youth online taxi users did not trust digital travel platforms. The implications of this finding are 2 fold: on the one hand, at the level of the impact of perceived risk on privacy concern, the results of this study echo previous studies on Internet use and the privacy concern of social media use among youths (Youn, 2009; Ho et al., 2017). Although youths of internet aborigines handled digital travel platforms in their daily lives constantly, they still sharply noticed the threat of digital technology to personal information, data, and privacy.

On the other hand, previous studies have suggested that self-efficacy is unrelated to privacy concerns (Yao et al., 2007). Contrary to previous studies, the statistical results of the two kinds of efficacy reported in our study indicate that self-efficacy and response efficacy have significant effects on privacy concerns. It is precisely because the youths are technologically proficient and thus believe that they are able to effectively protect their private information. These findings exposed self-confidence in information technology among contemporary youths, i.e., they are capable of employing cutting-edge technological gadgets to protect their privacy.

Are youths concerned about their privacy? Youths are the most active and vital force in society. In the era of privacy transparency, the entire society is questioning privacy concerns among youths. It is valuable and meaningful to examine whether youths pay attention to the information privacy of DTP or not. This study found an average score of privacy concern of 5.187 out of 7, which revealed a high level of privacy concern about digital travel platforms among the youth of online taxi users. It is noteworthy that privacy concerns not only directly influenced the privacy protection behavior of the youths but also functioned as an indirect factor between the relationships of perceived threat, self-efficacy, response efficacy, and privacy protection behavior. This finding is in line with the study of Lee et al. (2017). They suggest that privacy concerns have a positive impact on online privacy protection behavior among young people, which means that privacy concerns are an important element of privacy management for youth that cannot be ignored.

One of the imperative findings of this study was that the predictive power of privacy concern on privacy protection behavior among the high privacy knowledge group was significantly less than that of the low privacy knowledge group. Schwarzer et al. (1999) suggest that self-efficacy pertains to optimistic beliefs about coping with a large variety of stressors. However, excessive optimism can lead individuals to develop "optimism bias." Weinstein asserted that individuals tended to believe in having a greater opportunity to encounter active events than inactive ones, and negative experience with privacy protection might depress an individual's enthusiasm for acquiring protective action (Weinstein, 1980), which explained the logic of this finding. Sharot (2012) demonstrated the existence of optimism bias in human society through an experimental study and argued that optimism bias is a result of the evolution of the human brain, which can subconsciously change the subject's behavior and enhance individual wellbeing, but optimism bias may also cause blind optimism due to a lack of crisis awareness and reduce the individual's sense of prevention.

Xu (2011) confirmed the optimism bias of social network users. People usually believe they may be less vulnerable to privacy risks than others.

Similarly, another study also shows that users generally believe that negative events such as privacy leaks or information trafficking are less likely to happen to them (Campbell et al., 2007). In line with the above studies, our study also found the existence of so-called "optimistic bias" among the high privacy knowledge group. Due to the phenomenon of optimistic bias, individuals with high privacy knowledge tend to assume that they cannot confront threats more often than their counterparts. Therefore, they had a high level of privacy concern but a low level of privacy protection behavior. On the contrary, individuals with low privacy knowledge tended to lack IT awareness and skill, thus paying less attention to privacy and protective settings. Because they were unfamiliar with the degree of threat and its damage, which led to anxiety, they tended to enhance privacy concerns and adopt an aggressive protection mode when facing threats. This finding supports earlier research on the optimism bias of privacy risk (Kim and Hancock, 2015; Metzger and Suh, 2017). In addition, the results illustrated an inadequate understanding and familiarity with the privacy settings of digital travel platforms among youths, and approximately two-thirds were college students, meaning the knowledge of privacy settings was irrelevant to education level. The probable rationale was that youths tended to operate DTP when they needed online taxi-hailing but neglected the concern of privacy settings in their daily lives.

Youths should enhance their coping abilities with privacy risks. Firstly, intensifying the threat perception could effectively promote their concern for personal information and encourage them to adopt positive protective action on DTP. Secondly, individuals with extraordinary self-efficacy tended to adopt more active protective measures when applying digital travel platforms—for example, downloading travel software *via* an authorized APP store instead of a homepage link and avoiding clicking offensive websites to prevent possible intrusion of personal information. In addition, youths are able to promote response efficacy by paying more attention to related information about upholding privacy protection, awakening the coping ability of risky behavior, conducting adaptive training, such as conscious training on specific cases (i.e., party role-playing), and exercising prompt response aptitude. Finally, youths must recognize that enhancing their level of privacy knowledge is the most important method of preventing privacy threats. The fortification of skills and knowledge on privacy risk can improve privacy protection behavior and reduce the probability of infringement. Youths should improve their identification of various privacy risks and realize how to avoid them (Marcolin et al., 2000). Paying attention to the various elements of information safety, obtaining safety education and related training, enriching the knowledge of personal privacy protection, and keeping risk awareness of preparing for a rainy day are the required courses for youths to elevate personal information literacy.

From the perspective of platform self-discipline, a digital travel platform (an immediate information processor) is responsible for protecting users' information safety, particularly youths' information privacy. DTP must visibly declare the critical content of its privacy protection policy straightforwardly and clearly illustrate what kind of personal information was collected and how it was used. Thus, platform users are able to know fairly well how to raise awareness of privacy management. In addition, explicit, informed consent is the core principle of personal privacy protection and a basic maxim to comply with. Digital travel platforms should carefully respect youths' informational self-determination, exercise withdrawal, and obtain users' re-authorization as they employ the platform.

From the perspective of industry supervision, relevant government authorities should establish proprietary specifications for digital travel platform information privacy protection as soon as possible. Due to economies of scale and capital-seduced self-discipline failure, digital travel platforms tend to exhibit opportunistic motivations of so-called "management malfeasance." From the perspective of the legal guarantee, the legislations of the Civil Code, Personal Information Protection Law, and Data Security Law protected Chinese citizens' rights and interests in various information privacy matters effectively. When serious threats occur, youths should actively exercise their legal rights to defend personal information and privacy.

Nowadays, instead of sticking to a specific subject, communication research should focus on all walks of life (Schiller, 2018). Privacy is a multifaceted social problem, and youths are the backbone of society. Therefore, research on youths' consumption of DTP and privacy protection behaviors tended to have more academic value and space. This study explored the impact factors of digital travel platform utilization and privacy protection behavior among youths from a quantitative perspective. Further studies can examine the concerns and attitudes toward the privacy protection of DTP among youths. In addition, this study tried to examine youths' modes of acquiring knowledge of personal information privacy protection using test questions but failed to report the actual knowledge level objectively. Therefore, further studies are needed to measure multiple dimensions of knowledge regarding personal information protection.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

XM: guidance of research design, statistical analyses, and proof writing. BF: research designer and executor, conducting statistical analyses, and first draft writing. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by Chongqing Academy of Social Sciences (Grant No. 2022NDYB112).

Acknowledgments

The authors thank all the participants in this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.1049925/full#supplementary-material

References

Adhikari, K., and Panda, R. K. (2018). Users' information privacy concerns and privacy protection behaviors in social networks. *J. Global Market.* 31, 96–110. doi: 10.1080/08911762.2017.1412552

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychol. Rev.* 84, 191. doi: 10.1037/0033-295X.84.2.191

Campbell, J., Greenauer, N., Macaluso, K., and End, C. (2007). Unrealistic optimism in internet events. *Comput. Hum. Behav.* 23, 1273–1284. doi: 10.1016/j.chb.2004.12.005

Chen, H., Beaudoin, C. E., and Hong, T. (2017). Securing online privacy: An empirical test on Internet scam victimization, online privacy concerns, and privacy protection behaviors. *Comput. Hum. Behav.* 70, 291–302. doi: 10.1016/j.chb.2017.01.003

Chen, H. T., and Chen, W. (2015). Couldn't or wouldn't? The influence of privacy concerns and self-efficacy in privacy management on privacy protection. *Cyberpsychol. Behav. Soc. Netw.* 18, 13–19. doi: 10.1089/cyber.2014.0456

Chen, R. (2013). Living a private life in public social networks: An exploration of member self-disclosure. *Decis. Supp. Syst.* 55, 661–668. doi: 10.1016/j.dss.2012.12.003

Culnan, M. J. (1993). 'How did they get my name?': An exploratory investigation of consumer attitudes toward secondary information use. *MIS Quarterly* 17, 341. doi: 10.2307/249775

Cureton, E. E. (1957). The upper and lower twenty-seven per cent rule. *Psychometrika* 22, 293–96. doi: 10.1007/BF02289130

De Zwart, O., Veldhuijzen, I. K., Elam, G., Aro, A. R., Abraham, T., Bishop, G. D., and Brug, J. (2009). Perceived threat, risk perception, and efficacy beliefs related to SARS and other (emerging) infectious diseases: results of an international survey. *Int. J. Behav. Med.* 16, 30–40. doi: 10.1007/s12529-008-9008-2

Desimpelaere, L., Hudders, L., and Van de Sompel, D. (2020). Knowledge as a strategy for privacy protection: How a privacy literacy training affects children's online disclosure behavior. *Comput. Hum. Behav.* 110, 106382. doi: 10.1016/j.chb.2020.106382

Feng, Y., and Xie, W. (2014). Teens' concern for privacy when using social networking sites: An analysis of socialization agents and relationships with privacy-protecting behaviors. *Comput. Hum. Behav.* 33, 153–162. doi: 10.1016/j.chb.2014.01.009

Gerber, N., Gerber, P., Drews, H., Kirchner, E., Schlegel, N., Schmidt, T., and Scholz, L. (2018). FoxIT: Enhancing Mobile "Users' privacy behavior by increasing knowledge and awareness," in *Proceedings of the 7th Workshop on Socio-Technical Aspects in Security and Trust - STAST'17* (Orlando, Florida: ACM Press), 53–63. doi: 10.1145/3167996.3167999

Hanus, B., and Wu, Y. A. (2016). Impact of users' security awareness on desktop security behavior: A protection motivation theory perspective. *Inf. Syst. Manage.* 33, 2–16. doi: 10.1080/10580530.2015.1117842

Ho, S. S., Lwin, M. O., Yee, A. Z. H., and Lee, E. W. J. (2017). Understanding factors associated with singaporean adolescents' intention to adopt privacy protection behavior using an extended theory of planned behavior. *Cyberpsychol. Behav. Soc. Netw.* 20, 572–579. doi: 10.1089/cyber.2017.0061

Jia, R. N., Wang, X. W., and Fan, X. C. (2021). Empirical study on influencing factors of SNS user's personal information security and privacy protection behavior. *J. Modern Inf.* 41, 105–114.

Johnston, A. C., and Warkentin, M. (2010). Fear appeals and information security behaviors: An empirical study. *Mis Quarterly* 34, 549–66. doi: 10.2307/25750691

Kim, S. J., and Hancock, J. T. (2015). Optimistic bias and facebook use: Self–other discrepancies about potential risks and benefits of facebook use. *Cyberpsychol. Behav. Soc. Netw.*, 18, 214–220. doi: 10.1089/cyber.2014.0656

Lee, W. Y., Tan, C.-S., and Siah, P. C. (2017). The role of online privacy concern as a mediator between internet self-efficacy and online technical protection privacy behavior. *Sains Humanika.* 9, 1271. doi: 10.11113/sh.v9n3-2.1271

Liang, H., and Xue, Y. (2009). Avoidance of information technology threats: a theoretical perspective. *MIS Quart.* 33, 71. doi: 10.2307/20650279

Maddux, J. E., and Rogers, R. W. (1983). Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change. *J. Experim. Soc. Psychol.* 19, 469–479. doi: 10.1016/0022-1031(83)90023-9

Marcolin, B. L., Compeau, D. R., Munro, M. C., and Huff, S. L. (2000). Assessing user competence: conceptualization and measurement. *Inf. Syst. Res.* 11, 37–60. doi: 10.1287/isre.11.1.37.11782

Masur, P. K., Teutsch, D., and Trepte, S. (2017). Entwicklung und validierung der online-privatheitskompetenzskala (oplis). *Diagnostica*, 63, 256–268. doi: 10.1026/0012-1924/a000179

Metzger, M. J., and Suh, J. J. (2017). Comparative optimism about privacy risks on Facebook. *J. Commun.* 67, 203–232. doi: 10.1111/jcom.12290

Milne, G. R., Labrecque, L. I., and Cromer, C. (2009). Toward an understanding of the online consumer's risky behavior and protection practices. *J. Consumer Affairs*, 43, 449–473. doi: 10.1111/j.1745-6606.2009.01148.x

Park, Y. J., and Jang, S. M. (2014). Understanding privacy knowledge and skill in mobile communication. *Comput. Hum. Behav.*, 38, 296–303. doi: 10.1016/j.chb.2014.05.041

Qi, K. P., and Li, Z. Z. (2018). A study on privacy concerns of Chinese public and its influencing factors. *Sci. Soc.* 8, 36–58. doi: 10.19524/j.cnki.10-1009/g3.2018.02.036

Qiang, Y. X., and Xiao, D. (2021). Does overconfidence account for the Privacy Paradox? The impact of discrepancy between stated and actual privacy literacy on self-disclosure intention. *J. Mass Commun. Monthly* 6, 39–51. doi: 10.15897/j.cnki.cn51-1046/g2.20210507.002

Rippetoe, P. A., and Rogers, R. W. (1987). Effects of components of protection-motivation theory on adaptive and maladaptive coping with a health threat. *J. Person. Soc. Psychol.* 52, 596. doi: 10.1037/0022-3514.52.3.596

Rogers, R. W. (1975). A protection motivation theory of fear appeals and attitude change1. *J. Psychol.* 91, 93–114. doi: 10.1080/00223980.1975.9915803

Rosenthal, S., Wasenden, O. C., Gronnevet, G. A., and Ling, R. (2020). A tripartite model of trust in Facebook: acceptance of information personalization, privacy concern, and privacy literacy. *Media Psychol.* 23, 840–864. doi: 10.1080/15213269.2019.1648218

Schiller, D. (2018). *Networks and the Age of Nixon*. Beijing: Peking University Press.

Schwarzer, R., Mueller, J., and Greenglass, E. (1999). Assessment of perceived general self-efficacy on the Internet: data collection in cyberspace. *Anxiety, Stress Coping.* 12, 145–161. doi: 10.1080/10615809908248327

Sharot, T. (2012). The Optimism Bias: Why we're wired to look on the bright side. *The Psychiatrist*. (2012) 36, 439–40. doi: 10.1192/pb.bp.111.0 38182

Shen, Q. (2017). Risk and cost trade-offs: Privacy Paradox in social networks——An example of college students taking the WeChat mobile social application (platform) in Shanghai. *J. Commun.* 24, 55–69. Available online at: http://gfagzcadd5f6184ce4461snouvkuxuu6qo6uvu.fzfy.oca.swupl.edu.cn/kcms/detail/detail.aspx?FileName=YANJ201708004&DbName=CJFQ2017

Sheng, H., Nah, F. F. H., and Siau, K. (2008). An experimental study on ubiquitous commerce adoption: Impact of personalization and privacy concerns. *J. Assoc. Inf. Syst.* 9, 344–76. doi: 10.17705/1jais.00161

Smith, H. J., Milberg, S. J., and Burke, S. J. (1996). Information privacy: Measuring individuals' concerns about organizational practices. *MIS Quart.* 20, 167. doi: 10.2307/249477

Taylor, D. G., Davis, D. F., and Jillapalli, R. (2009). Privacy concern and online personalization: The moderating effects of information control and compensation. *Electr. Commerce Res.* 9, 203–223. doi: 10.1007/s10660-009-9036-2

Wang, L. Y., Li, Q., Qiao, Z. L., and Liu, S. (2019). Impact of protection motivation on privacy concerns and privacy security protection behaviors of SNS users. *J. Intell.* 38, 104–10. Available online at: http://gfagzcadd5f6184ce4461snouvkuxuu6qo6uvu.fzfy.oca.swupl.edu.cn/kcms/detail/61.1167.G3.20190712.1301.007.html

Wang, X. W., Wang, L., Jia, R. N., and Wang, D. (2018). An empirical study on the influencing factors of the security behavior in personal information in social networks. *Library Inf. Serv.* 62, 24–33. doi: 10.13266/j.issn.0252-3116.2018.18.003

Weinstein, N. D. (1980). Unrealistic Optimism about Future Life Events. *J. Person. Soc. Psychol.* 39, 806–20. doi: 10.1037/0022-3514.39.5.806

Wen, Z. L., Liu, H. Y., and Hau, K. T. (2012). *Analyses of Moderating and Mediating Effects*. Beijing: Educational Science Publishing House.

Witte, K. (1992). Putting the fear back into fear appeals: The extended parallel process model. *Commun. Monogr.* 59, 329–49. doi: 10.1080/03637759209376276

Workman, M., Bommer, W. H., and Straub, D. (2008). Security lapses and the omission of information security measures: A threat control model and empirical test. *Comput. Hum. Behav.* 24, 2799–2816. doi: 10.1016/j.chb.2008.04.005

Wu, D. J. (2020). Privacy concern of medical data and its influencing factors in the context of big data: An empirical study based on protection motivation theory. *J. Henan Normal Univ.* 47, 23–29. doi: 10.16366/j.cnki.1000-2359.2020.05.004

Xie, W., and Karan, K. (2019). Consumers' Privacy Concern and Privacy Protection on Social Network Sites in the Era of Big Data: Empirical Evidence from College Students. *J. Inter. Advert.* 19, 187–201. doi: 10.1080/15252019.2019.1651681

Xu, H. (2011). Reframing Privacy 2.0 in online social network symposium: privacy jurisprudence as an instrument of social change. *Univ. Pennsylvania J. Constit. Law*, 14, 1077–1102. Available online at: https://heinonline.org/HOL/P?h=hein.journals/upjcl14&i=1084

Yao, M. Z., Rice, R. E., and Wallis, K. (2007). Predicting user concerns about online privacy. *J. Am. Soc. Inf. Sci. Technol.* 58, 710–722. doi: 10.1002/asi.20530

Youn, S. (2009). Determinants of Online Privacy Concern and Its Influence on Privacy Protection Behaviors Among Young Adolescents. *J. Consumer Affairs* 43, 389–418. doi: 10.1111/j.1745-6606.2009.01146.x

Yu, T. T., and Yang, Y. H. (2019). Privacy concerns in online behavioral advertising. *J. Res.* 9, 101–116. Available online at: http://gfagzcadd5f6184ce4461snouvkuxuu6qo6uvu.fzfy.oca.swupl.edu.cn/kcms/detail/detail.aspx?FileName=XWDX201909010&DbName=CJFQ2019

frontiers | Frontiers in Psychiatry

TYPE Systematic Review
PUBLISHED 10 January 2023
DOI 10.3389/fpsyt.2022.1043987

OPEN ACCESS

EDITED BY
Dario Monzani,
University of Palermo, Italy

REVIEWED BY
Chun Wang,
Nanjing Brain Hospital Affiliated to Nanjing Medical University, China
Giulia Federica Perasso,
Fondazione Patrizio Paoletti, Italy
Giulia Ongaro,
University of Milan, Italy

*CORRESPONDENCE
Xiongzhao Zhu
✉ xiongzhaozhu@csu.edu.cn

SPECIALTY SECTION
This article was submitted to Psychopathology, a section of the journal Frontiers in Psychiatry

RECEIVED 14 September 2022
ACCEPTED 12 December 2022
PUBLISHED 10 January 2023

CITATION
Liu X, Wang X, Liu Y, Gao F, Xia J, Fan J and Zhu X (2023) The abnormal implicit memory to positive and negative stimuli in patients with current and remitted major depressive disorder: A systematic review and meta-analysis. *Front. Psychiatry* 13:1043987.
doi: 10.3389/fpsyt.2022.1043987

The abnormal implicit memory to positive and negative stimuli in patients with current and remitted major depressive disorder: A systematic review and meta-analysis

Xingze Liu[1,2,3], Xiang Wang[1,2,3], Yao Liu[1,2,3], Feng Gao[1,2,3], Jie Xia[1,2,3], Jie Fan[1,2,3] and Xiongzhao Zhu[1,2,3]*

[1]Medical Psychological Center, The Second Xiangya Hospital, Central South University, Changsha, Hunan, China, [2]Medical Psychological Institute of Central South University, Changsha, Hunan, China, [3]National Clinical Research Center for Mental Disorders, Changsha, Hunan, China

Introduction: In patients with current major depressive disorder (*c*MDD) a general abnormal implicit memory has been reported. However, the elaborate function of implicit memory when processing stimuli with different emotions (i.e., positive, neutral, and negative) in current and remitted (*r*MDD) patients is unclear. The present review examines implicit memory's general and elaborate in *c*MDD and *r*MDD patients.

Methods: We conducted meta-analyses based on published studies meeting criteria in Web of Science, PubMed, and EMBASE databases between 1990 and July 2022. The full sample patients included *c*MDD = 601 and *r*MDD = 143.

Results: Initial analysis of *c*MDD patients revealed a general implicit memory deficit. Subsequent subgroup analyses showed that the implicit memory performance to neutral stimuli is poorer in *c*MDD patients than controls, but recovered in *r*MDD patients; the deficient implicit memory to positive stimuli existed in *c*MDD and *r*MDD patients; the implicit memory performance to negative stimuli in *c*MDD patients is similar to controls but poorer in *r*MDD patients.

Conclusion: These findings indicate that the negative bias in *c*MDD patients might compensate for the general implicit memory deficit. Together, the implicit memory to neutral stimuli could recover with remission, whereas still abnormal in processing positive and negative stimuli. These results suggested that the abnormal implicit memory to positive and negative information might be relevant to depression pathogenesis.

Systematic review registration: https://www.crd.york.ac.uk/prospero, identifier CRD42020205003.

KEYWORDS

major depressive disorder, implicit memory, meta-analysis, stimuli types, memory bias

1. Introduction

Major depressive disorder (MDD) is a prevalent mental disorder (1, 2), accompanied by multiple cognitive abnormalities (3, 4). As one of the cognitive dysfunctions in MDD patients, the abnormality of implicit memory, which could be defined as unconscious or unintentional retrieval of past experience, has been broadly observed in patients who memorized more negative stimuli than healthy controls (5–7). This result is consistent with classical depression theories. For example, Beck et al. (8) suggested that MDD patients possessed stable and negative-biased representations of self-referential information like failure, loss, worthlessness, and hopelessness. Patients prefer to process various inputs toward negative experiences automatically once the negative representations stored in memory are activated (8–10). In other words, one activated negative memory node would automatically activate all the other associated negative nodes in memory (11–14). Such processing reflected a maladaptive memory pattern in MDD patients. As proposed in a study of Beevers (15), a cognitive vulnerability to depression is derived from the uncontrollable negative bias. If the stable and automatic bias cannot be controlled consciously, individuals may be more likely to develop depression. In contrast, if the implicit memory bias could be controlled consciously, it is possible to override the maladaptive pattern. However, the negative implicit memory bias was found in patients with current depression in most studies, whereas little was known whether the negative-bias was improved when they remitted from depression. Thus, a systematic and elaborate analysis of the implicit memory in patients with current and remitted major depressive disorder (i.e., *c*MDD and *r*MDD, correspondingly) could help recognize the stable pattern in patients' implicit memory, and provide possible diagnosis and interventions.

As Graf and Schacter (16) have stated, the implicit memory is revealed when performance on a task is facilitated in the absence of conscious recollection. This facilitation is usually measured through the repetition priming effect (17, 18), which is referring to the facilitation effect of a pre-exposed object to an identical object (19). Previous studies investigated the implicit memory in MDD patients by manipulating the emotional types of stimuli. Generally, the stimuli could be categorized into three types (i.e., positive, neutral, and negative). For studies that adopted stimuli with multiple emotional types, participants were presented a series of stimuli with different emotions, and they were told to respond to some stimuli characteristics (e.g., pronounce stimuli or judge the emotional type of each stimulus). After the presentations, they were asked to recall or recognize the stimuli presented before. The percentage or number of recalls in each emotional type is regarded as indicators to evaluate patients' implicit memory through comparisons with healthy controls; For studies only adopted neutral stimuli, there were always one or more regularities that were valid to improve performance but untold to participants. That is, the implicit memory occurs when the facilitation derived from the practice effect and benefits from the untold regularities. Importantly, these regularities are unconscious to participants after experiments (20). Therefore, the indicators to measure implicit memory are usually differences of reaction time or accuracy between trials with regularities and random. Most of these studies have observed that the implicit memory of *c*MDD patients was abnormal. However, this conclusion was inconsistent. For example, patients' implicit memory bias to negative stimuli was found in many studies but not observed in other studies [e.g., see (21–23)]; Likewise, the impaired implicit memory to neutral regularities was controversial (24, 25). These in conclusions made it difficult to tell whether the general function of implicit memory was abnormal.

In summary, one purpose of the present review is to examine the general function of implicit memory in *c*MDD patients. In addition, considering that previous studies were mainly focused on patients' negative-biased implicit memory, we would categorize implicit memory into three sub-function according to the emotional types of stimuli (i.e., positive, neutral, and negative) for elaborately examining the abnormalities of implicit memory when processing different stimuli. Lastly, as proposed in theories mentioned above, the implicit memory abnormality should be a stable cognition pattern in MDD patients. Thus, we will examine whether the general function and three sub-function of implicit memory was abnormal in patients remitted from major depressive disorder (*r*MDD). As depicted in **Figure 1**, the meta-analysis was conducted with two steps. Firstly, we will conduct two initial analyses of *c*MDD and *r*MDD patients separately to examine the general function of implicit memory in each patient group; then, we divided the data of included studies into three subgroups according to stimuli types (i.e., positive, neutral, and negative) and conducted subgroup analyses to each subgroup of stimuli types in *c*MDD and *r*MDD patients separately.

2. Materials and methods

The primary design of present review was registered on PROSPERO (CRD42020205003), and was conformed to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines regarding evidence selection, quality assessment, evidence synthesis, and research reporting [Hutton et al. (26)].

2.1. Literature search and inclusion criteria

Literatures that were written in English between 1990 and April 2022 were primarily sourced in three databases: PubMed, Web of Science and EMBASE, using the following combination of key words: ("major depressive disorder" [All

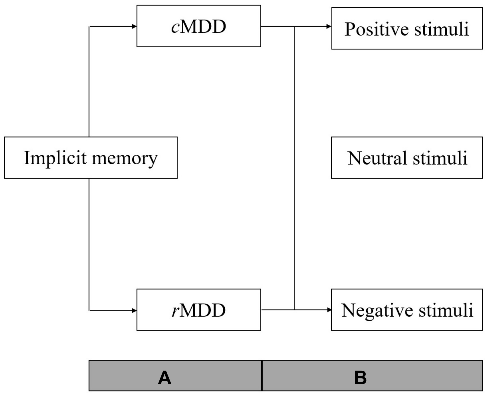


FIGURE 1
Over view of the analyses procedure. *c*MDD, patients with current major depressive disorder; *r*MDD, patients remitted from major depressive disorder; *A*, two initial meta-analyses were conducted to examine the general function of implicit memory in *c*MDD and *r*MDD patients; *B*, subgroup analyses of implicit memory to stimuli with different emotion types (positive, neutral, and negative) in *c*MDD and *r*MDD patients respectively.

Fields] OR "major depression" OR "depression" [All Fields] OR "depressed" [All Fields] OR "MDD" [All Fields]) AND ("implicit" [All Fields] OR "automatic" [All Fields]) AND ("memory" [All Fields] OR "learning" [All Fields]). Two authors screened studies and extracted data independently, and any disagreement was resolved by discussion until a consensus was reached or by consulting a third author.

To be included in the analysis, the selective criteria for studies were: (1) MDD patients (mean age ≥ 18 years) diagnosed with the Diagnostic and Statistical Manual of Mental Disorders (DSM) (27) or International Classification of Diseases (ICD) (28), which are free from psychotic features, bipolar disorder, comorbid ADHD, or substance abuse; (2) studies matched depressed patients with healthy controls; (3) studies using at least one psychological paradigm to measure the implicit memory; (4) sufficient data was reported to estimate effect sizes (e.g., mean and standard deviation or standard error data) for both groups; and (5) only case-control should be included.

2.2. Study selection and data extraction

All identified titles and abstracts were independently assessed for eligibility by two authors (XzL and YL) using a pilot form. Any disagreement in selected studies was resolved by discussion, and the arbitration of the third author (XW). One reviewer (XzL) conducted the full-text reviews of the reports and extracted the data into the structured forms. Then, another reviewer (XW) verified its completeness and accuracy. The included studies were attentively studied, the information collected is listed below: the location, author, and publication year of study; the age, gender, clinical data of participants, and outcomes of experiments (mean value and corresponding standard deviation of each experiment condition). For the outcome indicators, reaction time, numbers/percentage of recall, and fixed duration time of eye-tracking were collected.

2.3. Quality assessment

We applied the Newcastle-Ottawa Scale (29) to assess methodological quality of included studies in view of its' comparable comprehensive evaluate contents for case-control studies. The scores of 7-9, 4-6 and ≤ 3 in the Newcastle-Ottawa Scale are representative of high, moderate, and low quality in case-control studies accordingly. This part was performed by two investigators (XzL and XW). Any disagreements were resolved by consensus discussion with all authors.

2.4. Statistical analysis

As stated above, the analyses were conducted with the following steps: (1) two initial meta-analyses to data of *c*MDD and *r*MDD patients to examine the general function of implicit memory, (2) subgroup analyses of implicit memory to stimuli with different emotion types (i.e., positive, neutral, and negative) in *c*MDD and *r*MDD patients separately.

Stata version 12 was applied for data analysis. We calculated standardized mean differences (SMDs) and 95% confidence intervals (CIs) indicating the difference between patients and healthy controls. When experiment is conducted to same participants repeatedly (i.e., prior-treatment and post-treatment), only the performance in prior-treatment is included for the analysis. The magnitude of SMDs indicates: (0-0.2) = negligible effect, (0.2-0.5) = small effect, (0.5-0.8) = moderate effect, (0.8 +) = large effect (30). Heterogeneity is estimated with the I^2 statistic. I^2 statistic of 25, 50 and 75% were generally interpreted as small, moderate and high heterogeneity, respectively (31). In order to address heterogeneity, the random effect model is used. When the heterogeneity is high, we would conduct leave-one-out sensitivity analyses and random-effects meta-regression analyses to examine individual moderators, and if more than one moderator significantly predicted variance in effect size, we examined the moderators jointly as predictors.

3. Results

3.1. Included studies and quality assessment

3.1.1. Included studies

The procedure of literature searching is depicted in **Figure 2**. It was conducted through four steps (*Identification*, *Screening*, *Eligibility*, and *Inclusion* with two authors independently. In

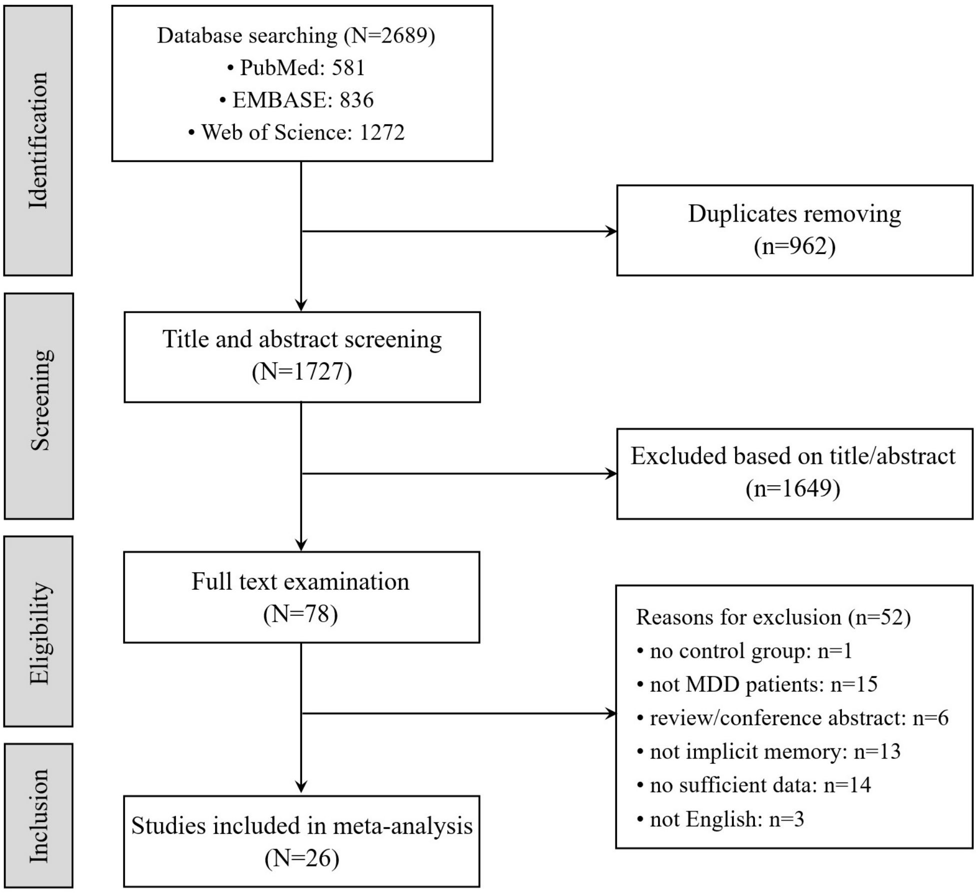

Identification
Screening
Eligibility
Inclusion
Database searching (N=2689)
• PubMed: 581
• EMBASE: 836
• Web of Science: 1272
Duplicates removing (n=962)
Title and abstract screening (N=1727)
Excluded based on title/abstract (n=1649)
Full text examination (N=78)
Reasons for exclusion (n=52)
• no control group: n=1
• not MDD patients: n=15
• review/conference abstract: n=6
• not implicit memory: n=13
• no sufficient data: n=14
• not English: n=3
Studies included in meta-analysis (N=26)

FIGURE 2
Flowchart of the trial selection process.

Identification, we searched three data bases (PubMed, Web of Science and EMBASE) with key words: ("major depressive disorder" [All Fields] OR "major depression" OR "depression" [All Fields] OR "depressed" [All Fields] OR "MDD" [All Fields]) AND ("implicit" [All Fields] OR "automatic" [All Fields]) AND ("memory" [All Fields] OR "learning" [All Fields]). There were 2,689 studies in total collected in the three data bases (581 studies in PubMed, 836 studies in EMBASE, 1,272 studies in Web of Science) and 1,727 studies after removed 962 duplicated studies from these data bases. In *Screening*, 1,649 of 1,727 studies are excluded because of their irrelevant title or abstract. In *Eligibility* and *Inclusion*, 26 studies are included in meta-analysis after removed 52 studies that did not meet standards.

There were total of 744 patients (143 *r*MDD and 601 *c*MDD) and 790 healthy controls included in 26 studies, and a group of healthy controls matched with patients in each included study. For *c*MDD patients, the sample size in these studies was ranged from 10 to 67, with mean age ranged from 23.5 ± 4.6 to 72.4 ± 9.0 years; for *r*MDD patients, the sample size ranged from 20 to 93, the mean age was ranged from 21.57 ± 1.43 to 36.2 ± 9.6 years; for healthy controls, the sample size and mean age range were 20 to 35 and 22.10 ± 1.95 to 35.1 ± 8.9 years. The detail characteristics of included studies were summarized in Table 1.

3.1.2. Quality assessment

As illustrated in Table 2, the average of the total score in NOS was 6.12. 17 studies showed moderate methodological quality, and nine of the rest are with high methodological quality.

In *Selection* part, the diagnostic criteria of all patients were DSM, ICD, or RDC. MDD patients and healthy controls in most of the included studies had corresponding representativeness; 19 studies recruited healthy controls from the community, and 23 studies defined healthy controls without any mental disorder history. In *Comparability* part, patients and controls in 25 studies matched age and/or other factors (e.g., gender, education, and IQ) to ensure the comparability of groups. There were only 2 studies that met the ascertainment of exposure criteria in *Exposure* part.

TABLE 1 Characteristic of included studies.

References	Patient					Control		Stimuli type
	N (F/M)	Age (years)	Diagnostic	Status	Depression severity	*N* (F/M)	Age (years)	
Elliott and Greene (32)	10 (NR)	31.5 ± NR	RDC	*cMDD*	HRSD: 27.3 (range = 20 ∼ 36)	10 (NR)	31.9 ± NR	Positive, Neutral, Negative
Bazin et al. (33)	23 (16/7)	43.22 ± 13.07	DSM-III-R	*cMDD*	BDI: 21.3 ± 6.30MADRS: 35.26 ± 4.92	37 (25/12)	45.24 ± 14.41	NC
Danion et al. (22)	30 (19/11)	41.2 ± 11.5	DSM-III-R	*cMDD*	HAMD-21: 29.5 (range = 18∼49)MADRS: 33.4 (range = 16∼48)	30 (19/11)	41.5 ± 11.6	Positive, Neutral, Negative
Ilsley et al. (34)	15 (6/9)	47.3 ± 16.2	DSM-III-R	*cMDD*	HRSD: 26.4 ± 5.9	15 (4/11)	43.6 ± 12.3	Positive, Negative
Bazin et al. (21)	23 (16/7)	43.2 ± 13	DSM-III-R	*cMDD*	BDI: 21.3 ± 6.30MADRS: 35.26 ± 4.92	37 (25/12)	44.2 ± 14	Positive, Negative
Watkins et al. (35)	67 (52/15)	NR	DSM-IV	*cMDD*	BDI-I: NR	67 (52/15)	NR	Positive, Neutral, Negative
Ellwart et al. (36)	36 (28/8)	42.06 ± 12.08	DSM-IV	*cMDD*	FDD: 39.9 ± 7.99	36 (26/10)	42.47 ± 12.76	Positive, Neutral, Negative
Tarsia et al. (23)	18 (8/10)	43.11 ± 8.93	DSM-IV ICD-10	*cMDD* *cMDD*	BDI-I: 25.67 ± 7.31	18 (10/8)	38.00 ± 9.91	Positive, Neutral, Negative
Aizenstein et al. (37)	11 (6/5)	68.70 ± 6.00	DSM-IV	cMDD	HAMD-17: 18.5 ± 4.8	11 (6/5)	71.3 ± 6.26	Neutral
Lim and Kim (38)	26 (NR)	36.53 ± 13.58	DSM-IV	cMDD	BDI-I: 24.20 ± 11.72	33 (16/17)	33.76 ± 7.96	Positive, Neutral, Negative
Rinck and Becker (39)	27 (NR)	23.5 ± 4.6	DSM-IV	cMDD	FDD: 25.3 ± 7.7	55 (NR)	21.4 ± 2.4	Positive, Neutral, Negative
Naismith et al. (40)	21 (NR)	53.9 ± 11.8	DSM-IV	cMDD	HAMD-17: 21.7 ± 4.4	21 (NR)	50.8 ± 11.7	Neutral
Lamy et al. (41)	18 (7/11)	38.8 ± 12.9	DSM-IV	cMDD	BDI-I: 24.8 ± 10.4	18 (7/11)	37.6 ± 11.8	Neutral
Vázquez et al. (7)	35 (26/9)	39.6 ± 12.2	DSM-IV	cMDD	BDI: 26.3 ± 1.1	36 (15/21)	30.4 ± 7.4	Positive, Neutral, Negative
Exner et al. (42)	26 (20/6)MEL 9 (4/5)Non-MEL	33.0 ± 10.0 35.0 ± 10.5	DSM-IV	cMDD	BDI MEL: 26.7 ± 10.3 BDI Non-MEL: 21.5 ± 8.3 HAMD-17 MEL: 22.1 ± 4.4 HAMD-17 Non-MEL: 16.3 ± 4.5	26 (18/8)	33.0 ± 8.9	Neutral
Ridout et al. (43)	16 (11/5)	43.7 ± 11.3	ICD-10	cMDD	BDI-I: 31.8 ± 1.8	18 (14/4)	39.3 ± 8.8	Positive, Neutral, Negative
Pedersen et al. (44)	20 (10/10)	36.2 ± 9.6	DSM-IV	rMDD	BDI: 10.9 ± 6.5 HDRS: 3.9 ± 2.8	20 (10/10)	35.1 ± 8.9	Neutral
Naismith et al. (45)	19 (14/5)	56.1 ± 9.8	DSM-IV	cMDD	HAMD-17: 21.6 ± 4.2	20 (14/6)	50.6 ± 11.9	Neutral
Elderkin-Thompson et al. (46)	32 (NR)	NR	DSM-IV	cMDD	BDI-I: 26.6 ± 8.2 HAMD-17: 18.3 ± 3.4	45 (NR)	NR	Neutral

(Continued)

TABLE 1 (Continued)

References	Patient					Control		Stimuli type
	N (F/M)	Age (years)	Diagnostic	Status	Depression severity	*N* (F/M)	Age (years)	
Romero et al. (47)	30 (24/6)	21.57 ± 1.43	DSM-IV	rMDD	BDI-II: 6.13 ± 3.39	30 (24/8)	22.10 ± 1.95	Positive, Neutral, Negative
Callahan et al. (5)	19 (15/4)	72.4 ± 9.0	DSM-IV	cMDD	GDS: 16.1 ± 3.8	28 (21/7)	72.1 ± 8.1	Positive, Neutral, Negative
Mörkl et al. (48)	44 (31/13)	39.27 ± 11.59	DSM-IV	cMDD	HRSD: 22.70 ± 4.90 BDI-I: 25.02 ± 9.23	44 (29/15)	40.5 ± 14.9	Neutral
Nemeth et al. (49)	28 (23/5)	49.22 ± 10.88	DSM-IV	cMDD	HRSD: 18.15 ± 5.21	28 (18/10)	46.83 ± 10.85	Positive, Neutral, Negative
Romero et al. (6)	38 (30/8)	26.79 ± 8.53	DSM-IV	cMDD	BDI-II: 26.74 ± 9.55	40 (31/9)	25.45 ± 7.93	Positive, Negative
Janacsek et al. (24)	10 (6/4)	47.90 ± 10.77	DSM-IV-TR	cMDD	BDI: 33.70 ± 8.68	10 (7/3)	44.58 ± 16.25	Neutral
Brian et al. (50)	93 (73/20)	23.17 ± 3.40	DSM-IV-TR	rMDD	HDRS: 3.63 ± 4.14	35 (20/5)	22.40 ± 3.08	Positive, Neutral, Negative

DSM-x, Diagnostic and Statistical Manual of Mental Disorder; ICD-x, International Classification of Diseases; cMDD, patients with current depression; rMDD, patients remitted from depression; HAMD-17/21/x, Hamilton Depression Rating Scale (17-/21-item/unstated version); BDI-I/13/II, Beck depression inventory (I-/II-version); MADRS, Montgomery-Asberg Depression Rating Scale; GDS, Geriatric Depression Scale; FDD, questionnaire for depression diagnosis (German version); QSD, severity of depression questionnaire; NR, not report; NC, not control.

3.2. Meta-analysis results

3.2.1. Initial meta-analysis

All included studies were divided into two data sets according to the depression status of patients (*c*MDD and *r*MDD). We then conducted two initial analyses to examine the general function of implicit memory in *c*MDD and *r*MDD patients (see **Figure 3**). For *c*MDD patients, the general function of implicit memory was significantly poorer than healthy controls (Effect size = −0.30; 95% CI: −0.53 to −0.08; $p < 0.05$), with the I^2 of 74.2%. Meanwhile, the Egger's test revealed no evidence for a publication bias (Egger's intercept = −1.21; 95% CI: −4.98 to 2.31, $p = 0.46$); for *r*MDD patients, the implicit memory was not significantly different between patients and controls for Effect size = −0.05, 95% CI: −0.32 to −0.22; $p = 0.71$, with I^2 of 0.0%. The Egger's test showed no publication bias to studies of *r*MDD patients (Egger's intercept = −1.36; 95% CI: −18.27 to 11.39, $p = 0.21$).

On account of the high heterogeneity ($I^2 = 74.2\%$) in studies of current depression, we then applied Galbraith graph to trace studies that possibly contributed to the heterogeneity (see **Supplementary material**). The graph indicated that there were five studies (32, 35, 41, 42, 48) primarily contributed the high heterogeneity. After removing these studies, the heterogeneity was decreased from 74.2 to 40.2%. Therefore, subsequent analyses would exclude the five studies.

3.2.2. Sub-group analyses

In this phase, we conducted three subgroup analyses according to the stimuli types (i.e., positive, neutral, and negative) in *c*MDD and *r*MDD separately. The data of *c*MDD patients was extracted from 18 studies, and the rest of 3 studies were *r*MDD patients. In studies of *c*MDD, one did not categorize the emotion types of stimuli (33), three only manipulated positive and negative stimuli (6, 21, 34), and 5 studies only adopted neutral stimuli (24, 37, 40, 45, 46). Two studies of rMDD patients only adopted positive and neutral stimuli severally [(50); Pedersen et al. (44)].

For *c*MDD patients, the implicit memory to neutral (19 studies) and positive stimuli (11 studies) were significantly poorer than controls for Effect size = −0.56, 95% CI = −0.86∼−0.25, $p < 0.001$, $I^2 = 84.3\%$ and Effect size = −0.60, 95% CI = −1.01∼−0.20, $p < 0.05$, $I^2 = 72.14\%$, respectively. However, the implicit memory performance to negative stimuli was similar between cMDD patients and controls (Effect size = 0.06, 95% CI = −0.30∼0.40, $p = 0.74$, $I^2 = 79.9\%$). For *r*MDD patients, the implicit memory to positive stimuli was still poorer than controls for Effect size: −0.80, 95% CI: −1.12 to −0.48; $p < 0.001$, $I^2 = 6.2\%$ while reversed to negative stimuli for *r*MDD patients could recall more negative stimuli than controls (Effect size = 0.82, 95% CI: 0.51 to 1.13; $p < 0.001$, $I^2 = 0.0\%$). The performance to neutral

TABLE 2 Quality assessment of included studies.

References	Selection				Comparability	Exposure		Total score[a]
	Case definition	Representa-tiveness of case	Selection of controls	Definition of controls	Comparability of cases and controls	Ascertain-ment of exposure	Same ascertain-ment for case/ Control	
Elliott and Greene (32)	☆	☆	☆		☆ ☆		☆	6
Bazin et al. (33)	☆	☆		☆	☆		☆	5
Danion et al. (22)	☆	☆	☆	☆	☆		☆	6
Ilsley et al. (34)	☆	☆	☆		☆		☆	5
Bazin et al. (21)	☆		☆	☆	☆		☆	5
Watkins et al. (35)	☆	☆	☆	☆	☆		☆	6
Ellwart et al. (36)	☆	☆	☆	☆	☆		☆	6
Tarsia et al. (23)	☆	☆		☆	☆		☆	5
Aizenstein et al. (37)	☆	☆	☆	☆	☆ ☆		☆	7
Lim and Kim (38)	☆	☆		☆	☆ ☆		☆	6
Rinck and Becker (39)	☆	☆	☆	☆			☆	5
Naismith et al. (40)	☆	☆	☆	☆	☆ ☆		☆	7
Lamy et al. (41)	☆	☆		☆	☆		☆	5
Vázquez et al. (7)	☆	☆		☆	☆		☆	5
Exner et al. (42)	☆	☆	☆	☆	☆ ☆		☆	7
Ridout et al. (43)	☆	☆		☆	☆ ☆		☆	6
Pedersen et al. (44)	☆		☆	☆	☆ ☆		☆	6
Naismith et al. (45)	☆	☆		☆	☆		☆	5
Elderkin-Thompson et al. (46)	☆	☆	☆	☆	☆		☆	6
Romero et al. (47)	☆	☆	☆		☆☆		☆	6
Callahan et al. (5)	☆	☆	☆	☆	☆ ☆	☆	☆	8
Mörkl et al. (48)	☆	☆	☆	☆	☆ ☆		☆	7
Nemeth et al. (49)	☆	☆	☆	☆	☆ ☆		☆	7
Romero et al. (6)	☆	☆	☆	☆	☆ ☆		☆	7
Janacsek et al. (24)	☆	☆	☆	☆	☆☆	☆	☆	8
Brian et al. (50)	☆	☆	☆	☆	☆ ☆		☆	7

A study can be awarded a maximum of 1 star for each item within the selection and exposure categories; a maximum of two stars can be given for comparability (☆ means yes, a total score of 7-9 indicates a high methodological quality, 4-6 indicates a moderate quality, and ≤3 indicates a low quality). ☆ ☆ Means studies met all two standards of the item.

stimuli was similar between patients and controls for Effect size = −0.24, 95% CI: −0.51 to 0.03; $p = 0.08$, $I^2 = 0.0\%$.

4. Discussion

4.1. Results summary

Present review mainly focused on examining the impairment of implicit memory in patients with current and remitted depression. Firstly, we conducted two initial meta-analyses to *c*MDD and *r*MDD patients separately to assess their general function of implicit memory. To further examine the implicit memory in detail, we categorized included studies into four groups based on the sub-function types of implicit memory (i.e., implicit learning group, positive, neutral, and negative groups of implicit memory bias) and conducted sub-group analysis to these groups.

The results of initial meta-analysis show that the general function of implicit memory in *c*MDD is impaired for Effect size = −0.30; 95% CI: −0.53 to −0.08; $p < 0.001$; $I^2 = 74.2\%$, but intact in *r*MDD for Effect size: −0.05, 95% CI: −0.32 to −0.22; $p = 0.7$; $I^2 = 0.0\%$. In subsequent sub-group analysis, *c*MDD patients are impaired to positive and neutral stimuli (Effect

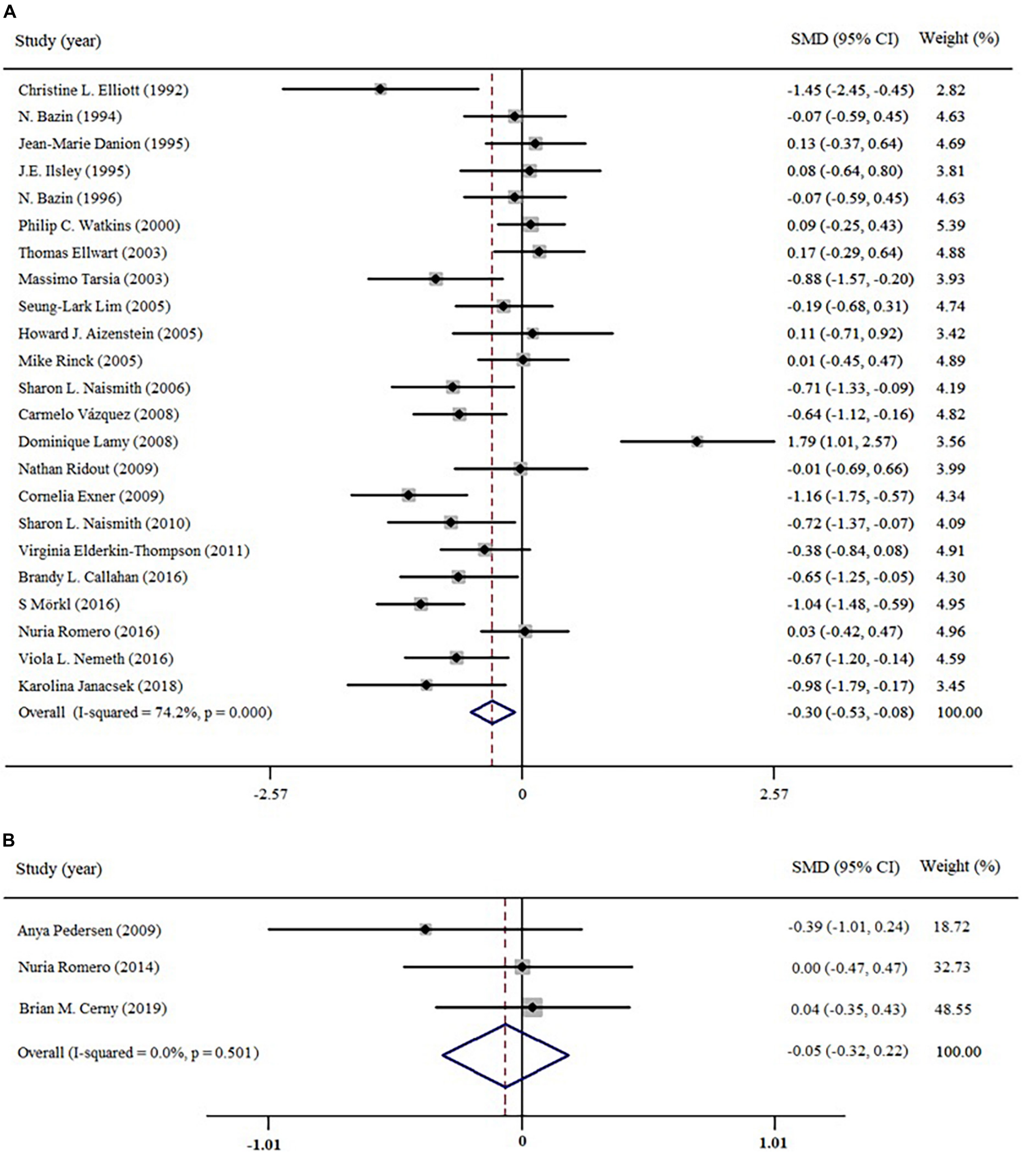


FIGURE 3
Forest plots of effect estimates of the general function of implicit memory in patients **(A)** current depression; **(B)** remitted depression compared to controls.

size = −0.66, 95% CI: −1.04 to −0.28; $p < 0.05$, $I^2 = 84.8\%$ and Effect size = −0.60, 95% CI = −1.01∼−0.20, $p < 0.05$, $I^2 = 72.14\%$, respectively), but similar with controls to negative stimuli (Effect size = −0.17, 95% CI: −0.61 to 0.28; $p = 0.46$, $I^2 = 89.0\%$). For *r*MDD patients, their implicit memory to neutral stimuli was similar to controls (Effect size = −0.24, 95% CI: −0.51 to 0.03; $p = 0.08$, $I^2 = 0.0\%$), whereas the implicit memory was still abnormal when processing positive and negative stimuli: *r*MDD patients exhibited poorer and better performance (Effect size = 0.82, 95% CI: 0.51 to 1.13; $p < 0.001$, $I^2 = 0.0\%$) to positive and negative stimuli (Effect size = −0.80, 95% CI: −1.12 to −0.48; $p < 0.001$, $I^2 = 6.2\%$) accordingly compared to controls. In brief, the implicit memory was generally impaired in *c*MDD patients, except for the negative stimuli. For *r*MDD patients, the implicit memory was recovered to neutral stimuli, but still abnormal to positive and negative stimuli.

4.2. Implicit memory to neutral stimuli

The implicit memory to neutral stimuli was assessed through two aspects. The first aspect was recall performance of neutral stimuli. In these studies, participants need to recall stimuli that they have processed but not intentionally memorized before (e.g., (22, 23, 32, 35, 36)). The second aspect regarded the latent regularity as memory content (e.g., (37, 40–42)). Relevant studies usually adopted visual search paradigms and shared similar experiment designs and logic. Generally, all the search displays could be divided into repeated and random conditions. Unlike random condition, the repeated condition contains a valid but latent regularity that could predict target location across trials. The search efficiency in both two conditions can improve over time by practice effect, but will be more significant in repeated condition if participants could memorize the regularities. The present review found that the recall performance in cMDD patients was poorer than controls, but recovered in rMDD patients. Therefore, it indicated that the abnormality of implicit memory to neutral stimuli was caused by depressive episode, and would recover with remission.

It should be mentioned that, one study of Lamy et al. (41) applied a unique and different paradigm to measure the implicit memory to neutral stimuli. They adopted the contextual cueing effect task that developed firstly by Chun and Jiang (51). Different from implicit sequence learning and weather prediction that are commonly used in MDD studies, the distractors in contextual cueing task share similar saliency (e.g., shapes, color, or topology) with the target so that participants need to pay attention (or top-down attention) for searching the target and making response. That is, the expression of the contextual cueing effect would be affected by multiple factors, such as the participations of selective attention (52, 53), working memory (54, 55), and successful attentional guidance and response selection (56, 57). Thus, future studies need to verify the mechanism of absent contextual cueing effect in MDD patients, and whether it depends on the development of depression.

4.3. Implicit memory to positive and negative stimuli

For studies applied the positive and negative stimuli, they shared same procedure and logic with neutral stimuli. The subgroup analyses showed that *c*MDD patients performed poorer than controls, but similar to controls when recalling negative stimuli. This indicated that the negative-biased memory tendency in *c*MDD patients might compensate the general impairment of implicit memory. In other words, the implicit memory in patients was biased to negative information, and the bias made them perform well as controls while deficit to neutral stimuli. For *r*MDD patients, recall performance to positive and negative stimuli was still abnormal; *r*MDD patients recalled more negative and less positive stimuli. These results might reveal a stable (or trait) cognitive characteristic in MDD patients for they possessed negative-biased and positive-avoided memory tendencies. However, whether the abnormal implicit memory bias was associated with the development of depression is still unclear. Future studies could longitudinally examine the abnormal implicit memory tendency to figure out latent cognitive indicators of depression development.

4.4. Implicit memory paradigms

The implicit memory is one categorization of memory that could process without participation of the conscious. It is generally assessed through the repetition priming effect in various paradigms with reaction time (RT), accuracy (ACC), and number/percentage of recall as indicators. We summarized the paradigms that commonly used in MDD patients' studies into two parts. The first part introduced the paradigms adopted emotional stimuli (i.e., positive, neutral, and negative); The second part focused on the paradigms that regarded latent regularities (e.g., spatial and semantic associations) as memory contents.

4.4.1. Emotional paradigms

The emotional paradigms usually manipulated emotional types of stimuli, and were primarily used to examine the implicit memory bias to stimuli with different emotional types. In these studies, the stimuli were regarded as memory contents and categorized into different emotion types (e.g., positive, neutral and negative). As listed in **Table 1**, this function could be assessed through various tasks, such as Self-referent incidental recall (6, 7, 38, 47), Mental imagery (36, 39), and Word-stem completion task (21, 22, 33–35). Most of them shared one similar procedure: in study phase, the stimuli that consisted of words or pictures with different emotion types would present to participants one at time, and they were asked to pronounce, imagine, or make decisions (e.g., matching faces; whether the word describes themself; evaluating the emotion types or valences of stimuli) to each of the presented words. The requirements were designed to ensure that participants could process the semantics of each stimulus without intentional memory. In the test phase, participants were informed to recall stimuli they processed before, or recognize them freely (or in a set mixed with new stimuli). The implicit memory bias would be considered happen if the recall performance (recall, recognition, or fixation duration with eye-tracking) to specific stimuli is better in patients/controls compared to another group.

4.4.2. Regularity paradigms

Each time we utilized the library to search for books of interests, the configuration of book categorizations and book placement within categorization we frequently browsed were stable. Such configurations helped us locate the target book more efficiently even though we have never intentionally memorized them. Psychologically, this improvement derived from repeated and stable configuration without conscious memory could be used to measure implicit memory.

Generally, most of these paradigms regarded regularities as memory contents and adopted various visual search tasks that participants will be asked to search specific targets in a series of search displays. Each display concluded one target and several distractors [e.g., single letter and geometric graphics, see (37, 41)]. All search displays could be divided into repeated and random conditions. In repeated condition, there was one or more regularities through trials, they were valid to predict the target location (or physical characteristics) so that the search performance would improve more efficiently than random condition, which has no such regularities. After completing the search task, participants usually needed to complete another recognition test to ensure that they were unconscious of the regularities in repeated condition. Tasks like implicit procedural learning (24, 37, 40, 42, 44, 45), and weather prediction (46, 48) tasks were commonly used. In an example of implicit procedural learning (40), participants need to judge which one of the four frames target would present in. All the search displays were categorized into repeated and random conditions. In repeated condition, the target location in each display was pseudo-random for the former target locations were associated with latter locations. Consequently, participants will perform better in repeated than random condition over time if they could implicitly memorize these location associations. Similarly, the weather prediction tasks (58) presented to participants with stimuli like words (*task 1*), geometric graphics (*task 2*), or artificial objects (*task 3*). Each display would be presented to participants with different combination patterns, and participants need to choose one of two possible outcomes while each combination pattern could predict the specific outcome with a different probability. Thus, the performance in each pattern should increase over time if participants could memorize the associations between combination patterns and outcomes. Lastly, Lamy et al. (41) adopted contextual cueing effect task that firstly adopted in Chun and Jiang (51) to assess patients' implicit memory. In this task, participants are informed to search for a target letter 'T' among distractor letters 'L's in each display. The displays could be categorized into repeated and novel conditions for the locations of target and distractors within each search display are stable in the former, but random in the latter. Over time, participants would gain an advantage for searching for targets in repeated condition over novel condition.

5. Limitations and new insights

There are several major limitations: For the first, the heterogeneity between included studies was moderate to high. Thus, we conducted a random effects model throughout to provide a conservative estimate. The second is that the sample size in some studies is insufficient. For example, the number of MDD patients in the study of Aizenstein et al. (37) and Janacsek et al. (24) were 11 and 10. That make it hard to decrease the affection of extraneous variables, and insufficient in the statistical validation. In addition, the included studies were search based on three databases (i.e., PubMed, Web of Science, and EMBASE). It was possible that some other relevant studies might not be collected so that limited the final sample size. Lastly, we did not examine the possible associations between implicit memory and clinical characteristics (e.g., severity and recurrence) in MDD patients. Hence, future studies could include more studies with sufficient sample sizes and conduct more detailed sub-group analyses for a more elaborate and accurate investigation of the implicit memory impairment in MDD patients.

6. Conclusion

The primary purpose of the present review was to examine the elaborate function of implicit memory in MDD patients with different statuses (i.e., current and remission). The results of the initial meta-analysis revealed a general impairment of implicit memory in both *c*MDD and *r*MDD patients. Then, the following sub-group analyses showed that the implicit memory of positive and neutral stimuli was abnormal in *c*MDD patients. Notably, the implicit memory of negative stimuli in cMDD patients was intact as healthy controls. This may suggest a compensatory effect of negative-biased memory tendency to a general memory deficit in *c*MDD patients. Further, the implicit memory to neutral stimuli was recovered in *r*MDD patients but remained abnormal to positive and negative stimuli. This result suggested that the abnormality of implicit memory to positive and negative stimuli was stable, indicating that implicit memory's positive-avoided and negative-biased tendencies might be a stable (or trait) dysfunction in MDD patients. Thus, we expect to provide a possible indicator (implicit memory to positive/negative stimuli) for the diagnosis and prediction of depression.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

XL: work design, data collecting, and manuscript writing. XW: technical and data collecting assistances. YL: data collecting assistance. FG, JX, and JF: make important revisions to this manuscript. XZ: guiding and co-ordinating the whole process. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2022.1043987/full#supplementary-material

References

1. Eaton WW, Alexandre P, Bienvenu OJ, Clarke D, Martins SS, Zablotsky B. The burden of mental disorders. In: Eaton WW editor. *Public Mental Health*. New York, NY: Oxford University Press (2012). p. 3–30.

2. Liu DY, Thompson RJ. Selection and implementation of emotion regulation strategies in major depressive disorder: an integrative review. *Clin Psychol Rev.* (2017) 57:183–94. doi: 10.1016/j.cpr.2017.07.004

3. Chen C, Jiang WH, Wang W, Ma XC, Li Y, Wu J, et al. Impaired visual, working, and verbal memory in first-episode, drug-naive patients with major depressive disorder in a Chinese population. *PLoS One.* (2018) 13:e0196023. doi: 10.1371/journal.pone.0196023

4. Withall A, Harris LM, Cumming SR. The relationship between cognitive function and clinical and functional outcomes in major depressive disorder. *Psychol Med.* (2009) 39:393.

5. Callahan BL, Simard M, Mouiha A, Rousseau F, Laforce R Jr., Hudon C. Impact of depressive symptoms on memory for emotional words in mild cognitive impairment and late-life depression. *J Alzheimers Dis.* (2016) 52:451–62. doi: 10.3233/JAD-150585

6. Romero N, Sanchez A, Vázquez C, Valiente C. Explicit self-esteem mediates the relationship between implicit self-esteem and memory biases in major depression. *Psychiatry Res.* (2016) 242:336–44. doi: 10.1016/j.psychres.2016.06.003

7. Vázquez C, Diez-Alegria C, Hernandez-Lloreda MJ, Moreno MN. Implicit and explicit self-schema in active deluded, remitted deluded, and depressed patients. *J Behav Ther Exp Psychiatry.* (2008) 39:587–99. doi: 10.1016/j.jbtep.2008.01.006

8. Beck AT, Rush AJ, Shaw BF, Emery G. *The Cognitive Therapy of Depression*. New York, NY: The Guilford Press (1979).

9. Abramson LY, Metalsky GI, Alloy LB. Hopelessness depression: a theory-based subtype of depression. *Psychol Rev.* (1989) 96:358–72.

10. Beck AT. Depression: clinical experimental, and theoretical aspects. *JAMA.* (1967) 203:1144–5. doi: 10.1001/jama.1968.03140130056023

11. Beck AT. The evolution of the cognitive model of depression and its neurobiological correlates. *Am J Psychiatry.* (2008) 165:969–77.

12. Bower GH. Mood and memory. *Am Psychol.* (1981) 36:129–48.

13. Ingram RE, Miranda J, Segal ZV. *Cognitive Vulnerability to Depression*. New York, NY: Guilford Press. (1998).

14. Teasdale JD. Cognitive vulnerability to persistent depression. *Cogn Emot.* (1988) 2:247–74.

15. Beevers CG. Cognitive vulnerability to depression: a dual process model. *Clin Psychol Rev.* (2005) 25:975–1002. doi: 10.1016/j.cpr.2005.03.003

16. Graf P, Schacter DL. Implicit and explicit memory for new associations in normal in amnesic subjects. *J Exp Psychol Learn Mem Cogn.* (1985) 11:501–18. doi: 10.1037//0278-7393.11.3.501

17. MacLeod C, Mathews AM. Cognitive–experimental approaches to the emotional disorders. In: Martin PR editor. *Handbook of Behavior Therapy and Psychological Science*. New York, NY: Pergamon Press (1991). p. 116–50.

18. Mulligan NW. The effect of generation on long-term repetition priming in auditory and visual perceptual identification. *Acta Psychologica.* (2011) 137:18–23.

19. Neill WT. Episodic rerieval in negative priming and repetition priming. *J Exp Psychol Learn Mem Cogn.* (1997) 23:1291–3105.

20. Schacter DL, Chiu CYP, Ochsner KN. Implicit Memory: a Selective Review. *Annu Rev Neurosci.* (1993) 16:159–82.

21. Bazin N, Perruchet P, Feline A. Mood congruence effect in explicit and implicit memory tasks: a comparison between depressed patients, schizophrenic patients and controls. *Eur Psychiatry.* (1996) 11:390–5. doi: 10.1016/s0924-9338(97)82575-8

22. Danion JM, Kauffmann-Muller F, Grangé D, Zimmermann MA, Greth P. Affective valence of words, explicit and implicit memory in clinical depression. *J Affect Disord.* (1995) 34:227–34. doi: 10.1016/0165-0327(95)00021-e

23. Tarsia M, Power MJ, Sanavio E. Implicit and explicit memory biases in mixed anxiety-depression. *J Affect Disord.* (2003) 77:213–25. doi: 10.1016/s0165-0327(02)00119-2

24. Janacsek K, Borbély-Ipkovich E, Nemeth D, Gonda X. How can the depressed mind extract and remember predictive relationships of the environment? Evidence from implicit probabilistic sequence learning. *Prog Neuropsychopharmacol Biol Psychiatry.* (2018) 81:17–24. doi: 10.1016/j.pnpbp.2017.09.021

25. O'Connor MG, Jerskey BA, Robertson EM, Brenninkmeyer C, Ozdemir E, Leone AP. The effects of repetitive transcranial magnetic stimulation (rTMS) on procedural memory and dysphoric mood in patients with major depressive disorder. *Cogn Behav Neurol.* (2005) 18:223–7. doi: 10.1097/01.wnn.0000187938.73918.33

26. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Ann Intern Med.* (2015) 162:777–84.

27. American Psychiatric Association [APA]. *Diagnostic and Statistical Manual of Mental Disorders: DSM-V*. 5th ed. Washington, DC: American Psychiatric Association (2013).

28. World Health Organization [WHO]. *The International Classification of Diseases (ICD), 2nded.* Geneva: World Health Organization (1993).

29. Wells G. *The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Non-Randomised Studies in Meta-Analyses. Paper presented at the Symposium on Systematic Reviews*. Milton Keynes, UK: Beyond the Basics (2014).

30. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. New York, NY: Lawrence Erlbaum Associates (1988).

31. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* (2003) 327:557–60.

32. Elliott CL, Greene RL. Clinical depression and implicit memory. *J Abnorm Psychol.* (1992) 101:572–4. doi: 10.1037//0021-843x.101.3.572

33. Bazin N, Perruchet P, De Bonis M, Féline A. The dissociation of explicit and implicit memory in depressed patients. *Psychol Med.* (1994) 24:239–45. doi: 10.1017/s0033291700027008

34. Ilsley JE, Moffoot APR, O'Carroll RE. An analysis of memory dysfunction in major depression. *J Affect Disord.* (1995) 35:1–9. doi: 10.1016/0165-0327(95)00032-I

35. Watkins PC, Martin CK, Stern LD. Unconscious memory bias in depression: perceptual and conceptual processes. *J Abnorm Psychol.* (2000) 109:282–9. doi: 10.1037//0021-843X.109.2.282

36. Ellwart T, Rinck M, Becker ES. Selective memory and memory deficits in depressed inpatients. *Depress Anxiety.* (2003) 17:197–206. doi: 10.1002/da.10102

37. Aizenstein HJ, Butters MA, Figurski JL, Stenger VA, Reynolds ICF, Carter CS. Prefrontal and striatal activation during sequence learning in geriatric depression. *Biol Psychiatry.* (2005) 58:290–6. doi: 10.1016/j.biopsych.2005.04.023

38. Lim SL, Kim JH. Cognitive processing of emotional information in depression, panic, and somatoform disorder. *J Abnorm Psychol.* (2005) 114:50–61. doi: 10.1037/0021-843x.114.1.50

39. Rinck M, Becker ES. A comparison of attentional biases and memory biases in women with social phobia and major depression. *J Abnorm Psychol.* (2005) 114:62–74. doi: 10.1037/0021-843X.114.1.62

40. Naismith SL, Hickie IB, Ward PB, Scott E, Little C. Impaired implicit sequence learning in depression: a probe for frontostriatal dysfunction? *Psychol Med.* (2006) 36:313–23. doi: 10.1017/s0033291705006835

41. Lamy D, Goshen-Kosover A, Aviani N, Harari H, Levkovitz H. Implicit memory for spatial context in depression and schizophrenia. *J Abnorm Psychol.* (2008) 117:954–61. doi: 10.1037/a0013867

42. Exner C, Lange C, Irle E. Impaired implicit learning and reduced pre-supplementary motor cortex size in early-onset major depression with melancholic features. *J Affect Disord.* (2009) 119:156–62. doi: 10.1016/j.jad.2009.03.015

43. Ridout N, Dritschel B, Matthews K, McVicar M, Reid IC, O'Carroll RE. Memory for emotional faces in major depression following judgement of physical facial characteristics at encoding. *Cogn Emot.* (2009) 23:739–52. doi: 10.1080/02699930802121137

44. Pedersen A, Kueppers K, Behnken A, Kroker K, Schoening S, Baune BT, et al. Implicit and explicit procedural learning in patients recently remitted from severe major depression. *Psychiatry Res.* (2009) 169:1–6. doi: 10.1016/j.psychres.2008.06.001

45. Naismith SL, Lagopoulos J, Ward PB, Davey CG, Little C, Hickie IB. Fronto-striatal correlates of impaired implicit sequence learning in major depression: an fMRI study. *J Affect Disord.* (2010) 125:256–61. doi: 10.1016/j.jad.2010.02.114

46. Elderkin-Thompson V, Moody T, Knowlton B, Hellemann G, Kumar A. Explicit and implicit memory in late-life depression. *Am J Geriatric Psychiatry.* (2011) 19:364–73. doi: 10.1097/JGP.0b013e3181e89a5b

47. Romero N, Sanchez A, Vázquez C. Memory biases in remitted depression: the role of negative cognitions at explicit and automatic processing levels. *J Behav Ther Exp Psychiatry.* (2014) 45:128–35. doi: 10.1016/j.jbtep.2013.09.008

48. Mörkl S, Blesl C, Jahanshahi M, Painold A, Holl AK. Impaired probabilistic classification learning with feedback in patients with major depression. *Neurobiol Learn Mem.* (2016) 127:48–55. doi: 10.1016/j.nlm.2015.12.001

49. Nemeth V, Csete G, Drotos G, Greminger N, Janka Z, Vecsei L, et al. The effect of emotion and reward contingencies on relational memory in major depression: an eye-movement study with follow-up. *Front Psychol.* (2016) 7:1849. doi: 10.3389/fpsyg.2016.01849

50. Brian MC, Jonathan PS, Leah RK, Elissa JH, Lisa A, O'Donnell C. Self-reported affective biases, but not all affective performance biases, are present in depression remission. *Br J Clin Psychol.* (2019) 58:274–88.

51. Chun MM, Jiang Y. Contextual cueing: implicit learning and memory of visual context guides spatial attention. *Cogn Psychol.* (1998) 36:28–71. doi: 10.1006/cogp.1998.0681

52. Jiang YH, Leung AW. Implicit learning of ignored visual context. *Psychon Bull Rev.* (2005) 12:100–6.

53. Jiménez L, Vázquez GA. Implicit sequence learning and contextual cueing do not compete for central cognitive resources. *J Exp Psychol Hum Percept Perform.* (2011) 37:222.

54. Manginelli AA, Baumgartner F, Pollmann S. Dorsal and ventral working memory-related brain areas support distinct processes in contextual cueing. *Neuroimage.* (2013) 67:363–74. doi: 10.1016/j.neuroimage.2012.11.025

55. Travis SL, Mattingley JB, Dux PE. On the role of working memory inspatial contextual cueing. *J Exp Psychol Learn Mem Cogn.* (2013) 39:208–19.

56. Wang C, Haponenko H, Liu X, Sun H, Zhao G. How attentional guidance and response selection boost contextual learning: evidence from eye movement. *Adv Cogn Psychol.* (2019) 15:265–75. doi: 10.5709/acp-0274-2

57. Zhao G, Liu Q, Jiao J, Zhou P, Li H, Sun H. Dual-state modulation of the contextual cueing effect: evidence from eye movement recordings. *J Vis.* (2012) 12:11. doi: 10.1167/12.6.11

58. Knowlton BJ, Squire LR, Gluck MA. Probabilistic classification learning in amnesia. *Learn Mem.* (1994) 1:106–20.

TYPE Original Research
PUBLISHED 16 February 2023
DOI 10.3389/fpsyg.2023.1125810

Check for updates

OPEN ACCESS

EDITED BY
Sergio Da Silva,
Federal University of Santa Catarina,
Brazil

REVIEWED BY
Raul Matsushita,
University of Brasilia,
Brazil
Travis Hodges,
Mount Holyoke College,
United States
Joao Plinio Juchem Neto,
Federal University of Rio Grande do Sul, Brazil

*CORRESPONDENCE
Gerda Ana Melnik-Leroy
✉ gerda.melnik@mif.vu.lt

SPECIALTY SECTION
This article was submitted to
Cognitive Science,
a section of the journal
Frontiers in Psychology

RECEIVED 16 December 2022
ACCEPTED 20 January 2023
PUBLISHED 16 February 2023

CITATION
Melnik-Leroy GA, Aidokas L, Dzemyda G, Dzemydaitė G, Marcinkevičius V, Tiešis V and Usovaitė A (2023) Is my visualization better than yours? Analyzing factors modulating exponential growth bias in graphs.
Front. Psychol. 14:1125810.
doi: 10.3389/fpsyg.2023.1125810

Is my visualization better than yours? Analyzing factors modulating exponential growth bias in graphs

Gerda Ana Melnik-Leroy*, Linas Aidokas, Gintautas Dzemyda, Giedrė Dzemydaitė, Virginijus Marcinkevičius, Vytautas Tiešis and Ana Usovaitė

Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania

Humans tend to systematically underestimate exponential growth and perceive it in linear terms, which can have severe consequences in a variety of fields. Recent studies attempted to examine the origins of this bias and to mitigate it by using the logarithmic vs. the linear scale in graphical representations. However, they yielded conflicting results as to which scale induces more perceptual errors. In the current study, in an experiment with a short educational intervention, we further examine the factors modulating the exponential bias in graphs and suggest a theoretical explanation for our findings. Specifically, we test the hypothesis that each of the scales can induce misperceptions in a particular context. In addition to this, we explore the effect of mathematical education by testing two groups of participants (with a background in humanities vs. formal sciences). The results of this study confirm that when used in an inadequate context, these scales can have a dramatic effect on the interpretation of visualizations representing exponential growth. In particular, while the log scale leads to more errors in graph description tasks, the linear scale misleads people when they have to make predictions on the future trajectory of exponential growth. The second part of the study revealed that the difficulties with both scales can be reduced by means of a short educational intervention. Importantly, while no difference between participants groups was observed prior to the intervention, participants with a better mathematical education showed a stronger learning effect at posttest. The findings of this study are discussed in light of a dual-process model.

KEYWORDS

cognitive bias, exponential growth, graph perception, logarithmic scaling, mathematical literacy, dual-process model

1. Introduction

Exponential growth is intrinsic to a large number of phenomena, ranging from the proliferation of microorganisms in biology, to compounding interests in economics or the nuclear chain reaction in physics (Lamarsh, 1983; Marr, 1991; Levy and Tasoff, 2017). Nevertheless, a growing body of literature confirms the difficulty of correctly perceiving this type of growth (Wagenaar and Sagaria, 1975; Wagenaar and Timmers, 1979). Specifically, people tend to systematically underestimate it and perceive it in terms of linear growth (Levy and Tasoff, 2017). This perceptual error has been termed 'the exponential growth bias'. Importantly, a biased perception of exponential growth has been shown to impact real-world behavior (Christandl and Fetchenhauer, 2009; Levy and Tasoff, 2016) and it recently attracted much attention due to its relevance in the context of the Covid-19-pandemic.

Namely, the infection rate of this virus follows an exponential trend, as according to estimations, the number of positive Covid-19 cases doubles every 3 days (Pellis et al., 2021). This growth has often been shown graphically in mainstream media (Engledowl and Weiland, 2021). Unfortunately, both the general public and government officials tended to misperceive it and to underestimate the risks and the severity of the disease (Gaissmaier, 2019; Lammers et al., 2020; Podkul et al., 2020). In particular, when asked to intuitively predict the number of COVID-19 cases in the future, many people underestimated how fast this value will increase (Banerjee and Majumdar, 2020; Jäckle and Ettensperger, 2021). They tended to think that the infections increase by a constant amount over each time interval (as is the case in linear growth), whereas in reality, exponential growth accelerates over time. Thus, as the quantity increases, so does that rate at which it grows: the more infections occur at the beginning of a disease outbreak, the more people will get infected. Nevertheless, people not only fail to perceive this growth, but they are also unaware of their errors (Cordes et al., 2019) and are even overconfident in their ability to deal with exponential growth (Levy and Tasoff, 2017). Growing evidence points to the fact that this has directly impacted the compliance with safety measures and therefore the spread of the virus (Muñiz-Rodríguez et al., 2020; Banerjee et al., 2021).

Several attempts have been made to find pedagogical ways to mitigate this bias. The most straightforward method, i.e., explaining about the bias and the potential perceptual mistakes it can induce, seems to work in certain cases (Lammers et al., 2020), but fails in others (Schonger and Sele, 2020). Other rather simple interventions, such as instructing participants to make estimates through intermediate steps (Lammers et al., 2020) or framing the scenario in terms of doubling times rather than growth rates (Schonger and Sele, 2020) have been shown to significantly reduce the bias. Despite these rather positive findings, other studies did not succeed in attenuating the bias *via* short graphical (Levy and Tasoff, 2016) or other types (using tables or direct non-numerical ways; Wagenaar and Sagaria, 1975; Wagenaar and Timmers, 1979) of interventions.

These mixed results point to the need of understanding better the mechanisms that induce the bias in order to mitigate it more effectively. In the relatively few studies assessing this question, such factors as the level of expertise of the participants (Christandl and Fetchenhauer, 2009) or the manipulation of the relevance of the topic (Romano et al., 2020) seem to have little or no effect on the occurrence of the exponential growth bias. Recently growing attention has been paid to the choice of the scale used in graphical representations of exponential growth (usually, line charts or scatterplots). In particular, as the logarithmic scale makes the exponential curve look linear, it can eliminate the underestimation bias and thus render the graphs more comprehensible (Ciccione et al., 2022). For instance, Hutzler et al. (2021) showed that participants looking at epidemiological data with logarithmically scaled growth curves have made significantly more accurate estimates than those who looked at linearly scaled graphs. In addition to this, with logarithmic scaling, their predictions were not susceptible to range changes on the y-axis as was the case in the linear scale condition, suggesting that participants could compare countries in different phases of infection growth more accurately. Similarly, Ciccione et al. (2022) also identified scaling as one of the factors that can attenuate the misperception of exponential growth when making predictions, alongside the noisiness of data, the task to be performed by the user (pointing vs. guessing a number) and his/her level of mathematical knowledge. However, other studies show that the logarithmic scale induces even stronger exponential growth bias. For instance, Romano et al. (2020) found that when participants are shown exponential growth on a logarithmic scale, they have much more difficulty in describing the graph and making predictions compared to a graph with a linear scale. In a similar vein, Menge et al. (2018) demonstrates that even professional scientists in ecology interpret graphs more accurately when they have linear rather than log-scaled axes.

In the current study, in an experiment with a short educational intervention, we further examine the factors modulating the exponential bias in order to shed more light on the somewhat conflicting results described above and suggest a theoretical explanation for these findings. First, we investigate in more detail the effect of using the linear versus the logarithmic scale in graphs when dealing with exponential growth. Studies on the visualization of other phenomena point out that differences between visualizations of the same data can drastically change the viewer‘s interpretation of information (Padilla et al., 2022). We hypothesize that the contradictory results found in the studies arise from the fact that they test the use of the two scales for different tasks: describing the data in the graph (or simply graph-reading) vs. making predictions on the trajectory of the growth. Specifically, we suggest that when a viewer has to read or describe a graph by attending to the values on the axes and extrapolating them, the linear scale is easier to use, as it can be interpreted straightforwardly, using the habitual tendency to reason linearly (Van Dooren et al., 2007). Indeed, adults with formal Western education tend to map numbers onto space in a linear manner (Dehaene et al., 2008). In this context, the log scale can be difficult to grasp and seem counterintuitive, as steps on a logarithmic scale are not additive but multiplicative (Menge et al., 2018). Several studies have shown that when reading a log-scaled graph, participants with different educational backgrounds confuse the values of the tick marks (Heckler et al., 2013) or tend to make numerical overestimations (Romano et al., 2020; Ciccione et al., 2022). On the other hand, when a person has to make predictions from a graph on the future trajectory of a growth, the log scale seems preferable, as it can help him/her notice the exponentially increasing growth rate even at its beginning, when it can look misleadingly slow on a linear scale (Hutzler et al., 2021). This is especially relevant, when data with differing growth trajectories and/or different orders of magnitude is plotted in the same graph (Perneger et al., 2020). In other words, the overreliance on linearity characteristic to many viewers (Van Dooren et al., 2007) can cause difficulties when using each of the scales in an unsuitable context: on one hand, if the log scale is perceived as linear, there is a risk of misinterpreting the values of the axis in graph description tasks. On the other hand, when the linear scale is used in prediction tasks, viewers might fail to perceive the slope of the growing curve and its exponential trends, leading to less accurate predictions. We investigate this issue by presenting two groups of participants with the same data plotted either on the log, or the linear scale. In both scale conditions, we ask participants the same questions that involve describing the graphs (questions 1–3) and making predictions based on it (questions 4–5). If the exponential bias in graphs is modulated by the presentation of a particular scale in the suitable context, participants in different scale conditions should respond differently to the same questions.

A second factor we examine in this study is the role that mathematical education can have on the perception of exponential growth in graphs. A body of literature demonstrates that mathematical skills and higher levels of numeracy can act as a protective mechanism against cognitive biases and oversimplifications through heuristics (Munoz-Rubke et al., 2022). For instance, higher numeracy was found to be associated with less confirmation bias (Hutmacher et al., 2022), while short educational interventions of mathematical nature were shown to reduce whole-number bias (Thompson et al., 2021). In the context of the exponential growth bias, only two studies directly looked at the effect of mathematical education. While Wagenaar and Sagaria (1975) found that mathematical sophistication of the subjects nor experience with growth processes modulated the bias, Ciccione et al. (2022) showed that higher mathematical knowledge led to smaller underestimation of exponential growth. Nevertheless, these papers assessed only indirectly the level of mathematical education through subjective questionnaires. Other studies on the exponential growth bias just looked at the general education level of their participants (Christandl and Fetchenhauer, 2009; Levy and Tasoff, 2016; Menge et al., 2018). In the current study we tested two groups of undergraduate students who had differing levels of math knowledge due to the nature of their respective curricula. Specifically, one group studied foreign languages and had few basic courses in math at secondary school and no math at university, while the other group studied computer science and had a substantial number of math courses both at secondary school and university. In this way, we ensured that alongside subjective self-evaluations of their math level, we had objective evidence about the education in math that both groups underwent.

For the second part of the experiment, we designed a short educational intervention in order to test if the difficulties of graph interpretation leading to exponential growth bias could be reduced in both scale conditions and across participant groups. Recent papers have called for designing interventions that could increase statistical literacy in general (Gal, 2002; Gould, 2017; Weiland, 2017; Engel, 2021), and the understanding of the exponential bias (Sieroń, 2020; Munoz-Rubke et al., 2022) alongside with the scales used (Menge et al., 2018; Watson and Callingham, 2020; Ciccione et al., 2022) in particular. For each scale condition we came up with short explanations accompanied by graphs that take into account the propositions expressed in several recent studies, including instructions on the organization of the log scale (Ciccione et al., 2022); the presentation and labelling in the graphs (Heckler et al., 2013; Menge et al., 2018); driving the participants' attention to certain elements of the graph (da Silva et al., 2021) etc.

Finally, following calls to investigate decision making with visualizations in terms of human perception and cognitive theory (Alhadad and Alhadad, 2018), we propose to interpret the results of this study in light of a dual-process model. According to this model, two types of decision-making processes exist: System 1 is used for fast automatic decisions and can be identified with intuitions; while System 2, or reasoning, is used for more rational analytical decisions (Stanovich, 1999; Kahneman and Frederick, 2002; Kahneman and Klein, 2009). This model helps to explain how the human mind deals with the limitations of its processing capacity and, in our view, can shed more light on the causes of the misperceptions arising in graph reading with different scales. We will address this issue in the Discussion section of this paper.

2. Part I: Pretest

2.1. Methods

2.1.1. Participants

99 participants took part in this online experiment. They were all recruited at Vilnius University and Vilnius Gediminas Technical University. 49 participants were enrolled in a BA degree in foreign languages, while the remaining 50 participants studied computer science. Note that students in computer science were chosen instead of students in mathematics intentionally, as we aimed at testing participants with an intermediate to high level of math, who have had math courses at university and who could represent a more general population with a background in natural/formal science, not just professionals in math. For simplicity, the first group will be labelled in this paper "humanities" and the second "science" group. Participants in each group were randomly assigned to one of the two experimental scale conditions. A between-subject design was chosen in order to avoid possible bias when dealing with both scales at a time. All participants were free to quit the experiment whenever they wanted, thus making sure that only interested and fully engaged participants were completing the experiment. After an initial screening of the data, 4 participants were excluded based on short completion time, resulting in a total of 47 participants in the humanities group, and 48 in the science group.

2.1.2. Stimuli

Two line charts for time series representing hypothetical Covid-19 daily case data from 3 countries were designed for the experiment. The Czech Republic, Finland and Spain were chosen for the examples, as they are well known for the Lithuanian participants, but are not associated with specific Covid-19 surges or containment policies, as Italy or Sweden would be. The hypothetical data for the three countries was distributed in such a way as to allow comparisons of large, small and asynchronous outbreaks (Perneger et al., 2020). Namely, Finland represented a smaller outbreak, while the other two showed a larger outbreak that progressed earlier in Spain, compared to the Czech Republic (see Figure 1). The line charts differed only in the scale used for the y axis (representing the daily new cases). Specifically, a linear scale was used for one condition, and a logarithmic scale for the other. For both conditions, only the major labels were shown (0–200–400-600-800-1,000 in the linear scale condition; 1–10–100-1,000 in the log scale condition), as it is common practice in online platform and media coverage across countries (Clement et al., 2020; Idogawa et al., 2020; Wissel et al., 2020). In both conditions the labels went up till 1,000 and were accompanied by grey major gridlines in order to facilitate the readability. In addition to this, minor tick marks without labels were also included in the graphs. This was especially important for the log scale, as there is evidence that in the absence of minor tick marks, people tend to interpret the logarithmic scale as linear (Heckler et al., 2013).

2.1.3. Procedure

Participants in each condition were presented with the corresponding line chart with either the linear or the log scale. In both scale conditions they were then asked the same 5 questions. We came up with three questions testing different aspects of graph description, and two questions assessing prediction-making from graphs. The questions and suggestions on the location (which condition) and the nature of the perceptual errors participants might make are presented in Table 1.

These questions were followed by two additional questions assessing subjective autoevaluation. Participants were asked to assess their confidence in their answers on a scale from 1 to 5 and to evaluate how difficult the tasks were (scale 1 to 5). At the very end of the experiment (following parts 1 and 2) participants had to indicate their level of math on a scale from 1 to 10.

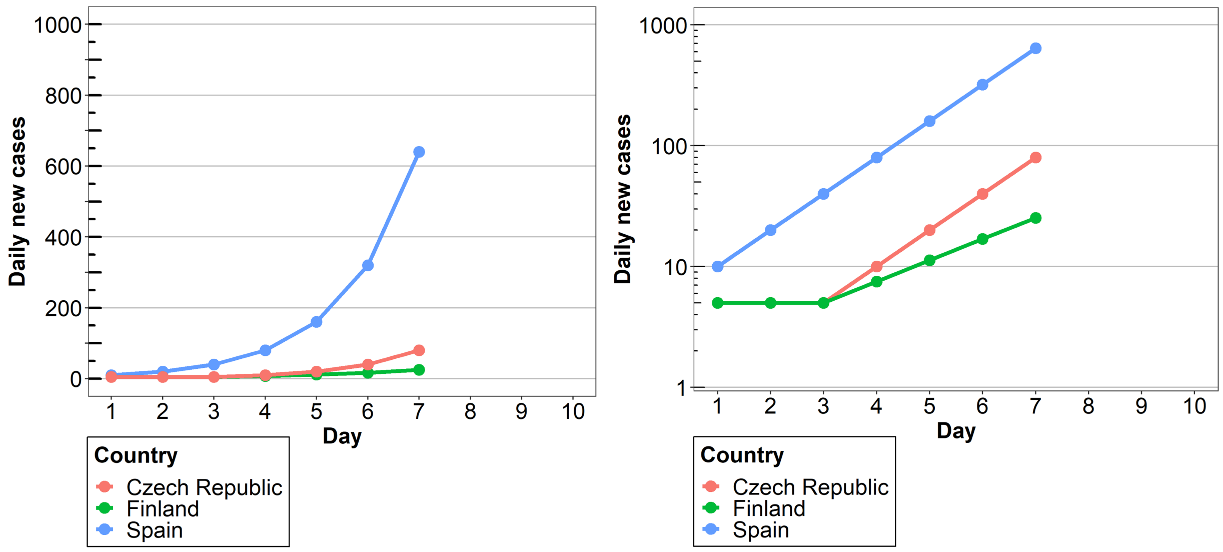


FIGURE 1
The two graphs presented to participants used in the linear and the log conditions, respectively.

TABLE 1 Questions asked during the experiment and their characteristics.

Question	Question type	Answer type (and proposed answer options)	Expected location and nature of misperceptions
1. Evaluate how many new cases occurred on day 6 in Spain.	Graph description	Ordinal (150; 300; 500; 800)	Log condition: while the correct answer is located halfway between the major tick marks 100 and 1,000, a linear interpretation of the log scale would lead to answering 500 instead of 300.
2. When did the number of daily new cases in Spain increase more?	Graph description	Ordinal (between day 4 and 5; between day 6 and 7; likewise)	Log condition: if participants interpret steps on this scale as linear, they would tend to think that the increase in both periods was identical and would fail to perceive the increasing outburst of cases over time.
3. Look at the difference in daily cases between Spain and the Czech Republic. How did the difference in cases from day 3 to day 7 change?	Graph description	Ordinal (decreased; remained stable; increased)	Log condition: An incorrect interpretation of the log scale would lead to a faulty perception of the growth dynamics of the two lines and the daily increasing difference in cases between them
4. Are cases in the Czech Republic more likely to grow like in Spain or in Finland?	Prediction making	Binary (like in Spain; like in Finland)	Linear condition: at first glance the growth of daily new cases in Finland and the Czech Republic looks more similar than that in Spain due to the differing growth rates between the first two and different growth progressions between the last two. This can lead to choosing the wrong growth trajectory (Finland instead of Spain).
5. What will approximately be the number of new cases in the Czech Republic on day 10?	Prediction making	Continuous (manual entry)	Linear condition: at first glance the growth of daily new cases in Finland and the Czech Republic looks more similar than that in Spain due to the differing growth rates between the first two and different growth progressions between the last two. This might lead the participants to providing a much lower estimate of future growth for cases in the Czech Republic than they actually are.

2.2. Results

We use an ordered logistic regression model to analyze the data from the first three questions, as they had ordered responses, which, however, cannot be considered continuous (Long, 1997). These analyses were performed using the polr command from the MASS package in R (Venables and Ripley, 2002). A logistic regression was used to analyze responses from question 4 (binary dependent variable), and a simple linear regression for question 5 (continuous dependent variable). For each of the five questions we constructed a model with Response to question as the dependent variable. Scale condition (linear vs. log) and Group (humanities vs. science), as well as their interaction were included as contrast-coded fixed factors. In questions 1–4, p-values were obtained by Wilks' likelihood ratio tests of the full model against the model without the effect or interaction in question. For question 5, an *Anova* was used for model comparison.

The analysis revealed that for all the questions there was a significant effect of Scale condition, but no effect of Group, nor an interaction between them. Specifically, in question 1, there was a significant difference between the linear and the log conditions [$\beta = 2.00$, SE = 0.52, $\chi^2(1) = 18.51$, $p < 0.0001$], with 86% of the participants (across groups) answering accurately in the linear condition compared to only 42% in the log condition (see Figure 2). This indicates that in the log condition many participants wrongly interpreted the values of the intermediate tick marks. As the target point was located halfway between the tick marks 100 and 1,000, they estimated that the value on the y axis was 500, instead of 300. Interestingly, although the difference between groups was not significant, we can see from the graph that many more participants from the science group made this mistake compared to the humanities group (62% vs. 31%). Turning to question 2, there were more than twice as many correct answers in the linear scale condition (95%) compared to the log scale condition (42%) [$\beta = 2.79$, SE = 0.67, $\chi^2(1) = 27.22$, $p < 0.0001$; see Figure 3]. Importantly, in the log condition a striking 52% of the participants clearly misunderstood the log scale in terms of a linear scale. When describing changes on this single curve, they misunderstood the pattern of change between two consecutive points: they thought that distances between points on the y axis are the same, independently on their location. A similar tendency can be observed in

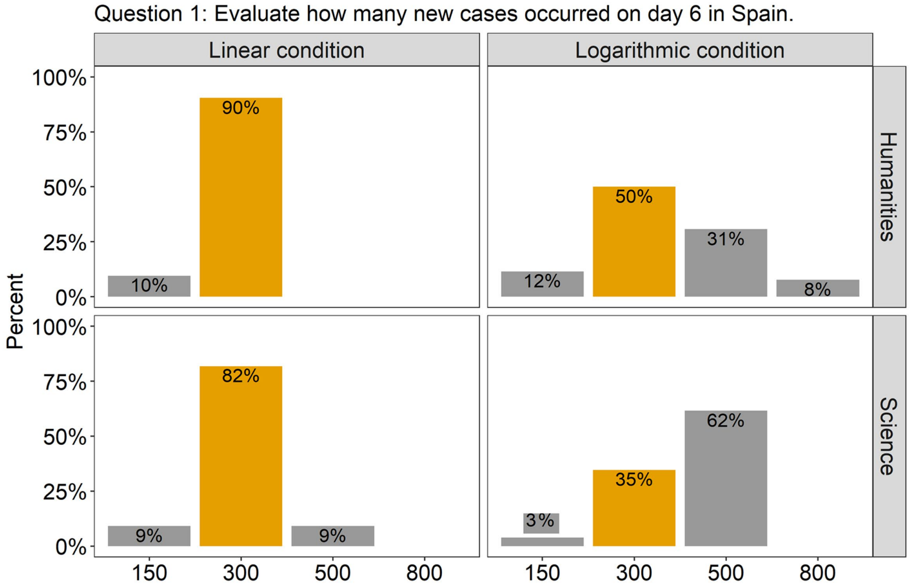


FIGURE 2
The barplots present the distribution of the participants' answers to Question 1 (graph description question). The correct answer is indicated with orange-colored bars.

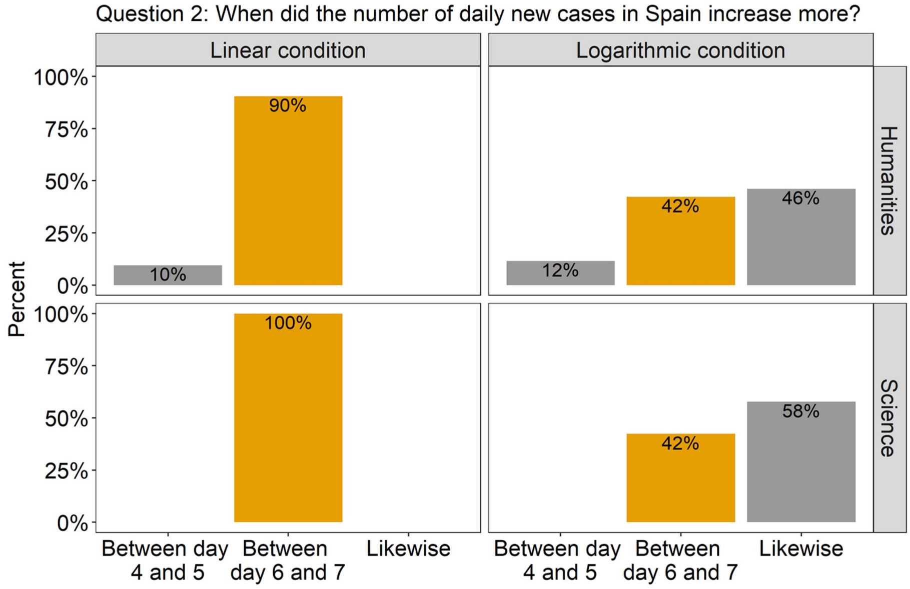


FIGURE 3
The barplots present the distribution of the participants' answers to Question 2 (graph description question). The correct answer is indicated with orange-colored bars.

question 3 (see Figure 4), where the difference between scale conditions [$\beta = -2.45$, $SE = 0.69$, $\chi^2(1) = 19.38$, $p < 0.0001$] was due to the great majority of participants (93%) answering correctly in the linear condition as opposed to 54% in the log one. This suggests that in the log condition participants misperceived the increasing distance between the two curves.

Turning to the questions involving predictions, in question 4 (Figure 5), the effect of Scale condition was again significant [$\beta = 2.80$, $SE = 0.79$, $\chi^2(1) = 20.39$, $p < 0.0001$], but this time the participants were much more accurate in the log scale condition (96% answered correctly) than in the linear scale condition (only 60% answered correctly). Finally, in question 5, where participants had to estimate the approximate number of new cases in the Czech Republic on day 10, a difference between the scale conditions was also observed [$\beta = 756.23$, $SE = 163.71$, $F(1, 91) = 21.34$, $p < 0.0001$]. As can be seen from Figure 6, participants in both

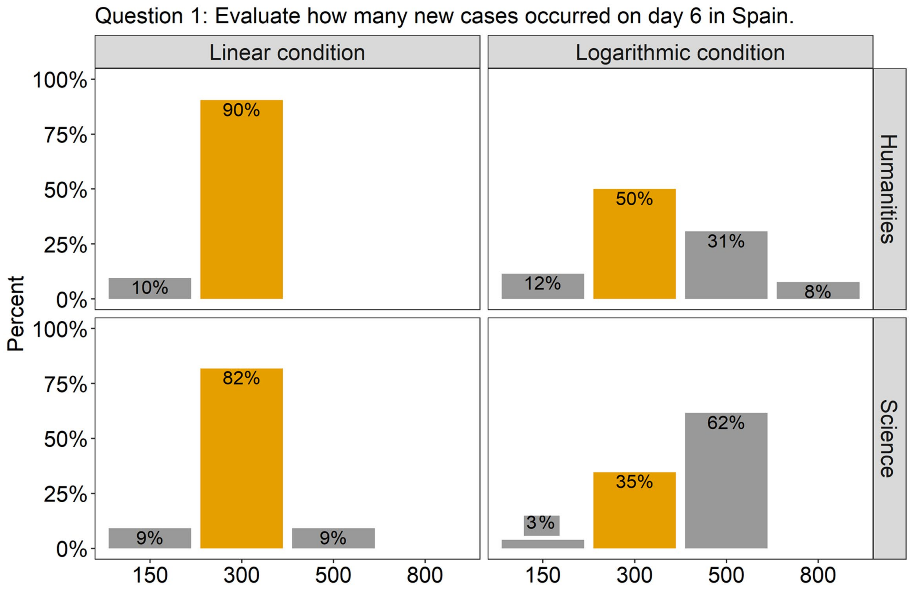


FIGURE 4
The barplots present the distribution of the participants' answers to Question 3 (graph description question). The correct answer is indicated with orange-colored bars.

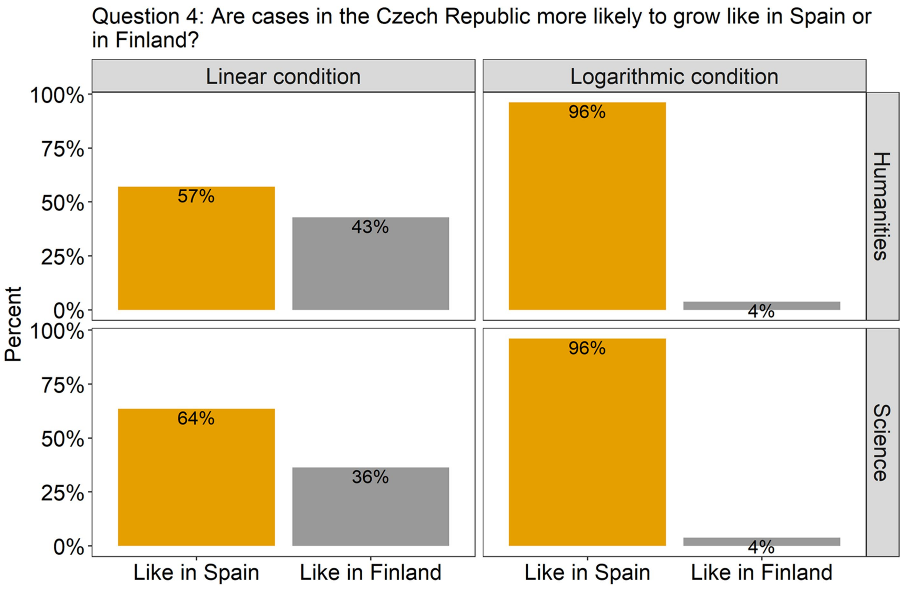


FIGURE 5
The plots present the distribution of the participants' answers to question 4 (prediction). The correct answer is indicated with orange-colored bars.

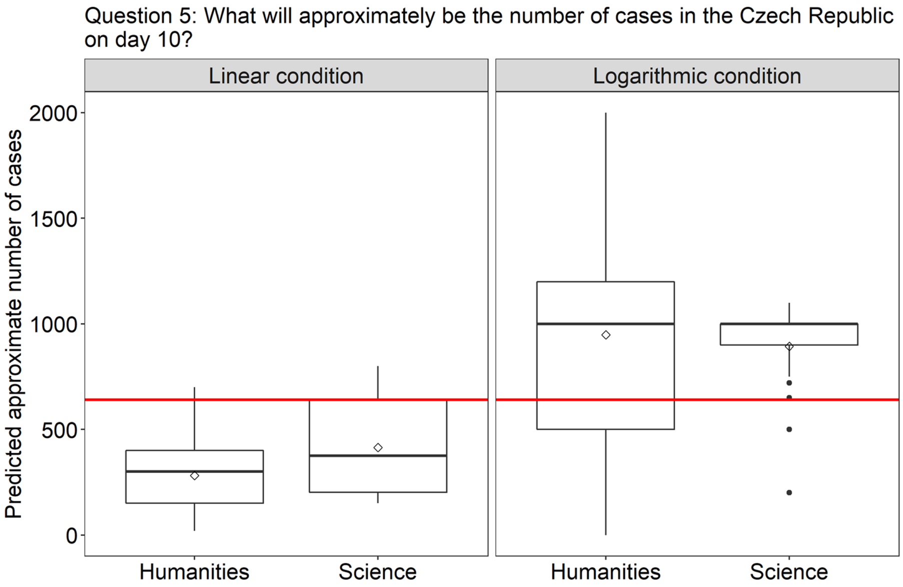


FIGURE 6
The plots present the distribution of the participants' answers to question 5 (prediction). The correct answer is indicated with a red line.

groups underestimated the growth in the linear scale condition, but overestimated it in the log condition. There was homogeneity of variances, as assessed by the Levene's test for equality of variances, for both Group ($p = 0.06$) and Scale condition ($p = 0.2$).

Finally, we carried out t-tests to examine whether the participants' answers on the additional questions differed depending on their educational background. First, we found that participants from the science group reported significantly higher scores on math autoevaluation than participants from the humanities group [on a scale from 1 to 10: science: mean = 6.94; humanities: mean = 5.06; $t(91.79) = -4.79$, $p < 0.0001$]. Next, we looked at the difference between groups in their level of confidence. Here too the difference was significant, namely, participants with a background in science felt more confident in their answers [on a scale from 1 to 5: science: mean = 3.56; humanities: mean = 3.02; $t(90.75) = -2.98$, $p < 0.01$]. Finally, the analyses revealed that participants in science found the tasks to be easier than their peers in humanities did [on a scale from 1 to 5: science: mean = 1.42; humanities: mean = 1.85; $t(91.31) = -2.29$, $p = 0.02$].

2.3. Discussion

The choice of the scale impacts indeed the responses of the participants. However, it is not the case that one scale is overall better than the other. Rather, each of the scales can induce errors in a particular context. Similarly to previous studies (Heckler et al., 2013), our experiment has shown once more that people misunderstand the minor tick marks on the log scale, and instead process them in terms of a linear scale. For this reason, the complexity of the log scale resulted in an inability to use it effectively to describe data on a graph. On the other hand, it proved to be very helpful in making predictions about future growth. Conversely, the experiment showed that the linear scale is much easier to use when describing a graph. Note, that participants in the linear scale condition reached very high accuracy (around 90% correct) on the first three questions. Nevertheless, this tendency was reversed in question 4, where participants had to compare the curves and predict their future growth. In this case, around 40% of the participants chose the wrong answer.

Interesting results were obtained on question 5. Here, participants in both conditions gave slightly inaccurate responses, but the nature of their mistakes was diametrically opposed. In particular, they underestimated the growth in the linear, but overestimated it in the log scale condition. Concerning the linear condition, this tendency reflects the typical exponential growth bias found in a variety of studies (Wagenaar and Sagaria, 1975; Hutzler et al., 2021). The tendency found in the log scale condition can seem more surprising, but it has already been observed by two other studies (Romano et al., 2020; Ciccione et al., 2022). The latter study found that the overestimation effect occurred when participants were presented with a noiseless exponential function rather than noisy data, which was also the case in our theoretical data scenarios. The overestimation effect found in the log scale condition could be overall considered preferable to the underestimation bias in many contexts. For instance, in cases of epidemics, the mere detection of exponential growth *per se* matters, while the exact estimate of the final numbers is not indispensable (Hutzler et al., 2021). On the other hand, these findings suggest that the choice of the scale could be motivated by the message one would wish to convey. Specifically, the log scale could be used in order to stress the importance of the growth of a phenomenon, while the linear scale could help to downplay its gravity. Note, however, that these tactics could also be employed to manipulate the viewer, and therefore a better understanding of these perceptual effects in the general public would be preferable.

Turning to the second factor we examined in this study, mathematical education does not seem to play a major role in the perception of exponential bias. Specifically, independently on their background in humanities or in science, both groups of participants were susceptible to the exponential growth bias when interpreting graphs plotted with an inappropriate scale. Yet, we found that participants in science group reported significantly higher autoevaluation in math levels, they felt more confident in their answers and had lower scores on perceived difficulty of the tasks. This points to one of the causes for the persistence of this bias: people are simply not aware of their lack of understanding of exponential growth. Christandl and Fetchenhauer (2009) also found that people are overconfident in their capacity to solve problems that involve exponential growth, which results in a low demand for corrective tools.

3. Part 2: Intervention and posttest

In order to test whether the difficulties experienced when using the log scale in describing a graph, and the linear scale when making predictions can be overcome, we designed a short intervention, which will be described next.

3.1. Methods

3.1.1. Participants

The same participants were tested as in part 1.

3.1.2. Stimuli

For each of the scale conditions we designed a short intervention consisting of two-slides-long instructions with graphs. We presented the same data as in part 1, but with additional information on the graphs and/or in the description, based on cumulated recommendations from several previous studies. For the log condition, we encouraged the participants to examine the y axis, and explained briefly the principles behind the log scale. We first explicitly drew their attention to the major tick marks and explained that each label is 10 times as large as the previous one (i.e.,$1 > 10 > 100 > 1{,}000$). We then provided information about the uneven distribution of the minor tick marks. In order to facilitate the understanding of this concept, we added a higher density of numerical labels in between major ticks (Heckler et al., 2013; Ciccione et al., 2022). In addition to this, we included additional gridlines at each minor tickmark.

In the linear condition we also added intermediate tick labels and gridlines. Moreover, we emphasized the fact that outbreaks can differ in their size and/or their timing. We encouraged participants to compare the three lines in terms of this information. Similar instructions that encourage the noticing of particular elements in the graph have been shown to help the viewers (Chang et al., 2016; Boone et al., 2018). The materials used for the educational intervention can be found in Supplementary materials.

3.1.3. Procedure

Participants were first presented with the educational intervention consisting of two slides with graphs and instructions. Following these slides, they saw the modified graphs with the same data and were asked to answer again the same questions. Participants were told that they either could answer as in the pretest, or modify their answers if needed. They were then asked to evaluate how useful the intervention was. Finally, participants were asked basic demographic questions (age, studies, gender) and to evaluate their level in math.

3.2. Results

As we have already shown in part 1 each of the scales can cause difficulties in a specific context. Therefore, results from the intervention will be presented by scale condition for those specific difficult questions, namely, the graph description questions for the log scale, and the prediction questions for the linear scale.

For the log condition, we looked at the first three questions (i.e., description of the graph) and we used an ordered logistic regression model to analyze the data. For each of them, we constructed a model with Response to question as the dependent variable. Session (pretest vs. posttest) and Group (humanities vs. science), as well as their interaction were included as contrast-coded fixed factors. p-values were obtained by likelihood ratio tests of the full model against the model without the effect or interaction in question. A summary of main effects and interactions that turned out to be significant in both parts of the experiments is presented in Table 2. The figures presenting the results for all five questions of the posttest can be found in Supplementary materials. For question 1 we found a significant effect of Session [$\beta = 1.04$, SE $= 0.41$, $\chi^2(1) = 6.63$, $p < 0.01$] and an interaction between Session and Group [$\beta = -2.11$, SE $= 0.82$, $\chi^2(1) = 6.87$, $p < 0.01$]. *Post-hoc* analyses revealed that the interaction was due to the fact that the difference between sessions was significant in the science group [$\beta = -3.02$, SE $= 0.84$, $\chi^2(1) = 51.68$, $p < 0.001$], but not in humanities ($p > 0.05$). That is, while in science group the accuracy improved from 35% to 89%, it only raised from 50% to 58% in the humanities group. Thus, there was a learning effect following the intervention in the former group, but not in the later. Turning to question 2, there was a significant effect of Session [$\beta = -1.57$, SE $= 0.45$, $\chi^2(1) = 13.24$, $p < 0.001$]. Specifically, in both study groups the correct answer was chosen only 42% of the times at pretest. At posttest, however, the accuracy improved in both groups, raising to 73% and 96% of correct responses in humanities and in science groups, respectively. Although the effect of Group was not significant, nor was the interaction, we still can note that the participants in science group benefited more from the intervention, almost reaching a ceiling effect at posttest. In question 3, we found significant effects of Session [$\beta = 0.94$, SE $= 0.45$, $\chi^2(1) = 4.62$, $p < 0.05$] and Studies [$\beta = 1.08$, SE $= 0.45$, $\chi^2(1) = 6.17$, $p < 0.05$]. Although the interaction only marginally approached significance ($p = 0.06$), we can observe a much stronger improvement in the science group following the intervention (the correct response rate raised from 58% to 89%, compared to 50% vs. 54% in the humanities group).

Turning to the linear scale condition and the questions involving predictions, a logistic regression was used to analyze responses from question 4, and a simple linear regression for question 5. Here too, Session (pretest vs. posttest) and Group (humanities vs. science), as well as their interaction were included as contrast-coded fixed factors. A significant effect of Session was found for question 4 [$\beta = -2.22$,

TABLE 2 Summary of the main effects and interactions that turned out to be significant in both parts of the experiment.

Question	1	2	3	4	5
Part 1: Pretest	Scale condition	Scale condition	Scale condition	Scale condition	Scale condition
Part 2: Posttest	Session Session × Group	Session	Session Group	Session	Session Group

SE = 0.71, $\chi^2(1) = 14.02$, $p < 0.001$]. Both groups showed similar levels of improvement, on average from 61% at pretest to 93% at posttest.

Finally, for question 5, both the factors Session [$\beta = 162.36$, SE = 47.39, $F(1, 82) = 11.74$, $p < 0.001$] and Studies [$\beta = 170.74$, SE = 47.39, $F(1,81) = 12.98$, $p < 0.001$] turned out to be significant. While both groups showed improvement after the intervention by reducing the underestimation tendency, this effect was much stronger in the science group (the correct answer to question 5 was 640; the mean predicted value in humanities at pretest was 282 and 406 at posttest; while in science it was 414 at pretest and 615 at posttest).

3.3. Discussion

The results of part 2 of the experiment show that even a short educational intervention can improve the reading and interpretation of graphs involving exponential growth bias. Specifically, it proved to be helpful in dealing with graphically presented data, under conditions when the use of the log and the linear scales causes the most mistakes (i.e., the log scale for the description of a graph and the linear scale for predictions). This result is important as in present times far-reaching measures related to crucial issues such as economic crises and hyperinflation or outbreaks of infectious deceases, such as Covid-19, are often explained with the help of data visualizations, while the general population has low levels of statistical literacy (Bakker and Wagner, 2020). As providing full-scale courses in statistics would be hardly possible for obvious reasons, the effectiveness of such short interventions is encouraging.

Nevertheless, we found a difference between groups in the majority of questions, that did not occur at pretest. Namely, participants in the science group seemed to benefit more from the intervention and showed a greater learning effect. This suggests that the intervention could potentially be adapted to different groups in order to maximize its effectiveness. The possible causes of the difference observed between groups at posttest are discussed in the following section.

4. General discussion

The current study demonstrated that the choice of the scale used to represent exponential growth in graphs can have a dramatic effect on the interpretation of these visualizations. The results confirmed our hypothesis that one scale is not overall better than the other. Rather, each of them can cause difficulties in a specific context. In particular, while the log scale leads to more errors when describing a graph, the linear scale can mislead people when they have to make predictions on the future trajectory of exponential growth. This at least partly explains why different studies obtained conflicting results as to which scale is more difficult to use. The second part of the study revealed that these difficulties with both scales can be reduced by means of a short educational intervention. Interestingly, while in the first part (pretest) there was no difference between participants with a background in science and those with a background in humanities, this difference was observed in the posttest. In particular, although both groups benefited to a certain extent from the intervention, the learning effect was much stronger in the science group.

We propose that our findings can be interpreted in light of a dual-process model. In particular, according to this model, reasonably accurate and effective decisions provided by System 1 are sufficient for the hundreds of decisions one has to make on a daily basis, although they might be prone to some errors (Stanovich, 1999). However, in situations where the mental shortcuts are not available and/or high levels of accuracy are required, the effortful System 2 comes at hand. In the field of visual processing, research has demonstrated that a limited set of visual features are detected preattentively (Healey and Enns, 2012). According to Padilla et al. (2018), who proposed an integrated model of decision making with visualizations, decisions based on graphs can be made by using either System 1 or System 2 processing. In the first scenario, viewers unconsciously focus on the aforementioned salient features and use minimal working-memory capacity, while in the second one they employ top-down attentional search of the visual array, which is taxing working memory, but might be more accurate.

In light of this theory, our results could be interpreted in the following manner: when describing graphs (question 1–3) in the linear condition, the viewers could rely on the salient graphical features, such as slopes, and automatically extract the necessary visual information. In other words, they could answer the questions in one or two steps, without having to analytically examine the different elements of the visualization, thus engaging little working memory. In this case, the use of System 1 was sufficient to provide accurate answers to the questions. On the contrary, more complex reasoning and more steps had to be involved in the log condition for the same questions. In particular, the reading and interpretation of the log scale *per se* required more attentional resources (i.e., driving one's attention to the scale of the y-axis, extrapolating the values of the major ticks, then the minor ticks, etc.). The resulting difficulty of participants to interpret the graph in this condition points to a persistent use of System 1 instead of the required System 2. In particular, it is likely that the viewers used heuristics usually employed to view graphs on the linear scale, which turned out to be misleading in this context.

Conversely, in the prediction question 4, participants could provide effortless accurate answers in the log condition, as it required only minimal reading and interpretation effort (they only had to look at the slopes of the curves and mentally prolongate them, as here the reading of the scale was not necessary to provide the correct answer). On the contrary, in order to be able to answer these questions in the linear condition, one would have to resort to graph analysis and inference making (compare the growth rate of all lines, evaluate their level of progression and synchronicity etc.). In question 5, where more analysis and the use of system 2 was necessary in both scale conditions (in both cases the viewers had to identify the 10th day on the x axis, then decide on where the line must continue, mentally draw it, after that extrapolate a number from the scale on the y axis etc.), many participants still applied linear thinking which turned out to be inadequate for the task. This resulted in underestimation in the linear condition and in overestimation in the log condition.

Thus, the results of the present study suggest that when dealing with graphs representing the exponential growth, viewers rely on salient features without examining the graph analytically and tend to use heuristics characteristic to System 1 processing. These involuntary shifts in focus to salient features bias the perception of graphs and can be detrimental to decision making (Padilla et al., 2018). Unfortunately, we did not record reaction times, which could provide further support for this interpretation of the results. The inclusion of reaction times along with accuracy scores would allow future studies to examine in more detail the possibility of using a computationally high System 2 for more difficult graph reading and prediction-making tasks.

Our second finding that the intervention overall improved the performance of both participant groups might also be explained in light of dual-processing theories. Specifically, it could be the case that both groups made mistakes at pretest as they tried to answer the questions intuitively by using System 1 in tasks which required the application of System 2. However, the instructions and information provided in the educational intervention pushed participants to deliberately pay more attention to certain elements of the graphs and to examine them analytically, thus employing System 2. This resulted in an overall better performance at posttest across groups. This raises the question whether the intervention was effective due to its pedagogical content, or rather it acted as a trigger to switch to a more analytical mode of processing. A future study could address this question by comparing the effect of two interventions, one containing pedagogical content, another – simple instructions to pay more attention to the different parts of the graph.

Turning to the difference found between groups at posttest, several explanations could account for a higher learning effect in participants with a background in science. For example, it is likely that the knowledge and skills they ought to have acquired during their studies got activated following the intervention. Alternatively, it could be the case that these participants were more receptive to the educational contents of the intervention. Previously acquired mathematical knowledge has been found to improve the overall capacities in conditional reasoning (Lehman and Nisbett, 1990; Gillard et al., 2009; Toplak et al., 2012). This entails that students who took a relatively large number of courses in math were more likely to employ strenuous System 2 processing (Borodin, 2016).

Note, however, that even if this was the case, the comparable performance of both groups at pretest on difficult questions point to a persistent use of System 1 processing. This suggests that when viewing graphs which use the inappropriate scale to represent exponential growth in the data, even the relatively "trained" viewers do encounter problems. For this reason, it is crucial for graph design to choose the scale that would direct participants' attention to the most important information. This would allow them to accurately and effortlessly extract the necessary information without having to resort to System 2. As pointed by Card et al. (1999) visualizations should capitalize on those visual biases which are consistent with the correct interpretation of the data. In our case, this would mean using the linear scale for the description of graphs representing exponential growth, and using the log scale to emphasize the growth when predictions have to be made. This is especially relevant when graphs are used to convey important information to the general public (Bakker and Wagner, 2020).

Finally, it is likely that the performance of the viewers, irrespective of their background, could be improved by teaching them general principles of graph reading. In particular, these skills are not necessarily directly trained in traditional math courses, thus even viewers with a background in science might benefit from such training (Bakker and Wagner, 2020). At the same time, it would not require specific mathematical knowledge, and thus would be easily applicable in curricula in various fields. For instance, da Silva et al. (2021) propose that the viewer should be encouraged to engage in four levels of graph reading and interpretation (i.e., reading, interpretation, prediction making and critical assessment) in order to develop a habit to examine graphs analytically and thus improve their accuracy. Overall, it is equally important to both avoid common pitfalls when designing graphs, as well as to improve the skills of the viewers by means of educational interventions. This is in line with the numerous calls to improve the didactics of many mathematical and statistical topics, as well as statistical literacy and graph-reading skills in the general population, which became crucial in our modern data-driven society (Gal, 2002; Sharma, 2017; Bakker and Wagner, 2020; Watson and Callingham, 2020).

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://midas.lt:443/action/resources/836f4caf-d7a0-480b-aa7a-5d5e0b1c6b90.

Ethics statement

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

GM-L, VT, and GinD contributed to grant application. All authors contributed to conception and design of the study. LA and AU conducted the study and collected the data. GM-L performed the data analysis. All authors contributed to the article and approved the submitted version.

Funding

This project has received funding from European Social Fund (project No 13.1.1-LMT-K-718-05-0034) under grant agreement with the Research Council of Lithuania (LMTLT).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1125810/full#supplementary-material

References

Alhadad, S., and Alhadad, S. S. J. (2018). Visualizing data to support judgement, inference, and decision making in learning analytics: insights from cognitive psychology and visualization science. *Learning Analytics* 5, 60–85. doi: 10.18608/jla.2018.52.5

Bakker, A., and Wagner, D. (2020). Pandemic: lessons for today and tomorrow? *Educ. Stud. Math.* 104, 1–4. doi: 10.1007/s10649-020-09946-3

Banerjee, R., Bhattacharya, J., and Majumdar, P. (2021). Exponential-growth prediction bias and compliance with safety measures related to COVID-19. *Soc. Sci. Med.* 268:113473. doi: 10.1016/j.socscimed.2020.113473

Banerjee, R., and Majumdar, P. (2020). *Exponential growth bias in the prediction of COVID-19 spread and economic expectation* IZA DP. Bonn, Germany.

Boone, A. P., Gunalp, P., and Hegarty, M. (2018). Explicit versus actionable knowledge: the influence of explaining graphical conventions on interpretation of hurricane forecast visualizations. *J. Exp. Psychol. Appl.* 24, 275–295. doi: 10.1037/xap0000166

Borodin, A. (2016). The need for an application of dual-process theory to mathematics education. *Cambridge Open-Review Educ. Res. e-Journal* 3, 1–31. doi: 10.17863/CAM.41156

Card, S. K., Mackinlay, J. D., and Shneiderman, B. (1999). *Readings in information visualization: Using vision to think*, Morgan Kaufmann Publishers, San Francisco, CA.

Chang, R., Yang, F., and Procopio, M. (2016). From vision science to data science: applying perception to problems in big data. *Soc. Imaging Sci. Technol.* 16, 1–7. doi: 10.2352/ISSN.2470-1173.2016.16HVEI-131

Christandl, F., and Fetchenhauer, D. (2009). How laypeople and experts misperceive the effect of economic growth. *J. Econ. Psychol.* 30, 381–392. doi: 10.1016/j.joep.2009.01.002

Ciccione, L., Sablé-Meyer, M., and Dehaene, S. (2022). Analyzing the misperception of exponential growth in graphs. *Cognition* 225:105112. doi: 10.1016/j.cognition.2022.105112

Clement, F., Kaur, A., Sedghi, M., Krishnaswamy, D., and Punithakumar, K. (2020). Interactive data driven visualization for COVID-19 with trends, analytics and forecasting. *Proc. Int. Conf. Inf. Vis.* 4, 593–598. doi: 10.1109/IV51561.2020.00101

Cordes, H., Foltice, B., and Langer, T. (2019). Misperception of exponential growth: are people aware of their errors? *Decis. Anal.* 16, 261–280. doi: 10.1287/deca.2019.0395

da Silva, A. S., Barbosa, M. T. S., de Souza Velasque, L., da Silveira Barroso Alves, D., and Magalhães, M. N. (2021). The COVID-19 epidemic in Brazil: how statistics education may contribute to unravel the reality behind the charts. *Educ. Stud. Math.* 108, 269–289. doi: 10.1007/s10649-021-10112-6

Dehaene, S., Izard, V., Spelke, E., and Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in western and Amazonian indigene cultures. *Science* 320, 1217–1220. doi: 10.1126/science.1156540

Engel, J. (2021). Statistical literacy for active citizenship: a call for data science education. *Stat. Educ. Res. J.* 16, 44–49. doi: 10.52041/serj.v16i1.213

Engledowl, C., and Weiland, T. (2021). Data (Mis)representation and COVID-19: leveraging misleading data visualizations for developing statistical literacy across grades 6–16. *J. Stat. Data Sci. Educ.* 29, 160–164. doi: 10.1080/26939169.2021.1915215

Gaissmaier, W. (2019). A cognitive-ecological perspective on risk perception and medical decision making. *Med. Decis. Mak.* 39, 723–726. doi: 10.1177/0272989X19876267

Gal, I. (2002). Adults' statistical literacy: meanings, components, responsibilities. *Int. Stat. Rev.* 70, 1–25. doi: 10.1111/j.1751-5823.2002.tb00336.x

Gillard, E., Van Dooren, W., Schaeken, W., and Verschaffel, L. (2009). Dual processes in the psychology of mathematics education and cognitive psychology. *Hum. Dev.* 52, 95–108. doi: 10.1159/000202728

Gould, R. (2017). Data literacy is statistical literacy. *Stat. Educ. Res. J.* 16, 22–25. doi: 10.52041/serj.v16i1.209

Healey, C. G., and Enns, J. T. (2012). Attention and visual perception in visualization and computer graphics. *IEEE Trans. Vis. Comput. Graph.* 18, 1170–1188. doi: 10.1109/TVCG.2011.127

Heckler, A. F., Mikula, B., and Rosenblatt, R. (2013). Student accuracy in reading logarithmic plots: the problem and how to fix it. *Proc. - Front. Educ. Conf. FIE*. 1066–1071. doi: 10.1109/FIE.2013.6684990

Hutmacher, F., Reichardt, R., and Appel, M. (2022). The role of motivated science reception and numeracy in the context of the COVID-19 pandemic. *Public Underst. Sci.* 31, 19–34. doi: 10.1177/09636625211047974

Hutzler, F., Richlan, F., Leitner, M. C., Schuster, S., Braun, M., and Hawelka, S. (2021). Anticipating trajectories of exponential growth. *R. Soc. Open Sci.* 8. doi: 10.1098/rsos.201574

Idogawa, M., Tange, S., Nakase, H., and Tokino, T. (2020). Interactive web-based graphs of coronavirus disease 2019 cases and deaths per population by country. *Clin. Infect. Dis.* 71, 902–903. doi: 10.1093/cid/ciaa500

Jäckle, S., and Ettensperger, F. (2021). Boosting the understanding and approval of anti-Corona measures–reducing exponential growth bias and its effects through educational nudges. *Schweizerische Zeitschrift für politische Wissenschaft* 27, 809–821. doi: 10.1111/spsr.12479

Kahneman, D., and Frederick, S. (2002). "*Representativeness revisited: Attribute substitution in intuitive judgment, Heuristics and Biases*." Cambridge: Cambridge University Press, pp. 49–81.

Kahneman, D., and Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *Am. Psychol.* 64, 515–526. doi: 10.1037/a0016755

Lamarsh, J. R. (1983). *Introduction to nuclear reactor theory*, Addison-Wesley, Reading, Addison-We, United States.

Lammers, J., Crusius, J., and Gast, A. (2020). Correcting misperceptions of exponential coronavirus growth increases support for social distancing. *Proc. Natl. Acad. Sci. U. S. A.* 117, 16264–16266. doi: 10.1073/pnas.2006048117

Lehman, D. R., and Nisbett, R. E. (1990). A longitudinal study of the ef ects of undergrad-uate training on reasoning. *Dev. Psychol.* 26, 952–960. doi: 10.1037/0012-1649.26.6.952

Levy, M., and Tasoff, J. (2016). Exponential-growth bias and lifecycle consumption. *J. Eur. Econ. Assoc.* 14, 545–583. doi: 10.1111/jeea.12149

Levy, M. R., and Tasoff, J. (2017). Exponential-growth bias and overconfidence. *J. Econ. Psychol.* 58, 1–14. doi: 10.1016/j.joep.2016.11.001

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*, Thousand Oaks, CA: Sage.

Marr, A. G. (1991). Growth rate of Escherichia coli. *Microbiol. Rev.* 55, 316–333. doi: 10.1128/mmbr.55.2.316-333.1991

Menge, D. N. L., MacPherson, A. C., Bytnerowicz, T. A., Quebbeman, A. W., Schwartz, N. B., Taylor, B. N., et al. (2018). Logarithmic scales in ecological data presentation may cause misinterpretation. *Nat. Ecol. Evol.* 2, 1393–1402. doi: 10.1038/s41559-018-0610-7

Muñiz-Rodríguez, L., Rodríguez-Muñiz, L. J., and Alsina, Á. (2020). Deficits in the statistical and probabilistic literacy of citizens: effects in a world in crisis. *Mathematics* 8, 1–20. doi: 10.3390/math8111872

Munoz-Rubke, F., Almuna, F., Duemler, J., and Velásquez, E. (2022). Mathematical tools for making sense of a global pandemic. *Int J Sci Educ B Commun Public Engagem* 12, 1–10. doi: 10.1080/21548455.2022.2100941

Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., and Stefanucci, J. K. (2018). Correction to: decision making with visualizations: a cognitive framework across disciplines. *Cogn. Res. Princ. Implic.* 3:34. doi: 10.1186/s41235-018-0126-3

Padilla, L., Hosseinpour, H., Fygenson, R., Howell, J., Chunara, R., and Bertini, E. (2022). Impact of COVID-19 forecast visualizations on pandemic risk perceptions. *Sci. Rep.* 12:2014. doi: 10.1038/s41598-022-05353-1

Pellis, L., Scarabel, F., Stage, H. B., Overton, C. E., Chappell, L. H. K., Fearon, E., et al. (2021). Challenges in control of COVID-19: short doubling time and long delay to effect of interventions. *Philos. Trans. R. Soc. Lond.* 376:20200264. doi: 10.1098/rstb.2020.0264

Perneger, T., Kevorkian, A., Grenet, T., Gallée, H., and Gayet-Ageron, A. (2020). Correction to: alternative graphical displays for the monitoring of epidemic outbreaks, with application to COVID-19 mortality. *BMC Med. Res. Methodol.* 20, 1–9. doi: 10.1186/s12874-020-01147-z

Podkul, A., Vittert, L., Tranter, S., and Alduncin, A. (2020). The coronavirus exponential: a preliminary investigation into the Public's understanding. *Harvard Data Science Rev.* 1. doi: 10.1162/99608f92.fec69745

Romano, A., Sotis, C., Dominioni, G., and Guidi, S. (2020). The scale of COVID-19 graphs affects understanding, attitudes, and policy preferences. *Health Econ.* 29, 1482–1494. doi: 10.1002/hec.4143

Schonger, M., and Sele, D. (2020). How to better communicate the exponential growth of infectious diseases. *PLoS One* 15, e0242839–e0242813. doi: 10.1371/journal.pone.0242839

Sharma, S. (2017). Definitions and models of statistical literacy: a literature review. *Open Rev. Educ. Res.* 4, 118–133. doi: 10.1080/23265507.2017.1354313

Sieroń, A. (2020). Does the COVID-19 pandemic refute probability neglect? *J. Risk Res.* 23, 855–861. doi: 10.1080/13669877.2020.1772346

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*, NJ: Lawrence Erlbaum Associates Publishers. *20*

Thompson, C. A., Taber, J. M., Sidney, P. G., Fitzsimmons, C. J., Mielicki, M. K., Matthews, P. G., et al. (2021). Math matters: a novel, brief educational intervention decreases whole number bias when reasoning about COVID-19. *J. Exp. Psychol. Appl.* 27, 632–656. doi: 10.1037/xap0000403

Toplak, M. E., West, R. F., and Stanovich, K. E. (2012). "Education for rational thought" in *Enhancing the quality of learning: Dispositions, instruction, and learning processes*. ed. J. R. Kirb (Cambridge: Cambridge University Press), 51–92.

Van Dooren, W., De Bock, D., Janssens, D., and Verschaffel, L. (2007). Pupils' over-reliance on linearity: a scholastic effect? *Br. J. Educ. Psychol.* 77, 307–321. doi: 10.1348/000709906X115967

Venables, W. N., and Ripley, B. D. (2002). *Modern applied statistics with S*, Springer, New York, Fourth edi.

Wagenaar, W. A., and Sagaria, S. D. (1975). Misperception of exponential growth. *Percept. Psychophys.* 18, 416–422. doi: 10.3758/BF03204114

Wagenaar, W. A., and Timmers, H. (1979). The pond-and-duckweed problem; three experiments on the misperception of exponential growth. *Acta Psychol.* 43, 239–251. doi: 10.1016/0001-6918(79)90028-3

Watson, J., and Callingham, R. (2020). COVID-19 statistical literacy. *Aust. Math. Educ. J.* 2, 16–22.

Weiland, T. (2017). Problematizing statistical literacy: an intersection of critical and statistical literacies. *Educ. Stud. Math.* 96, 33–47. doi: 10.1007/s10649-017-9764-5

Wissel, B. D., Van Camp, P. J., Kouril, M., Weis, C., Glauser, T. A., White, P. S., et al. (2020). An interactive online dashboard for tracking COVID-19 in U.S. counties, cities, and states in al time. *J. Am. Med. Inform. Assoc.* 27, 1121–1125. doi: 10.1093/jamia/ocaa071

TYPE Brief Research Report
PUBLISHED 23 February 2023
DOI 10.3389/fpsyg.2023.1128200

Check for updates

OPEN ACCESS

EDITED BY
Dario Monzani,
University of Palermo, Italy

REVIEWED BY
Steven Phillips,
National Institute of Advanced Industrial Science and Technology (AIST), Japan
Athanasios Drigas,
National Centre of Scientific Research Demokritos, Greece

*CORRESPONDENCE
Leona Polyanskaya
✉ leona.polyanskaya@uni-goettingen.de

SPECIALTY SECTION
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

RECEIVED 20 December 2022
ACCEPTED 30 January 2023
PUBLISHED 23 February 2023

CITATION
Polyanskaya L (2023) I know that I know. But do I know that I do not know? *Front. Psychol.* 14:1128200. doi: 10.3389/fpsyg.2023.1128200

I know that I know. But do I know that I do not know?

Leona Polyanskaya[1,2,3]*

[1]Georg-August-Universität Göttingen, Allgemeine Sprachwissenschaft, Göttingen, Germany, [2]CIBIT-Coimbra Institute for Biomedical Imaging and Translational Research, Coimbra, Portugal, [3]Faculty of Psychology and Educational Sciences, University of Coimbra, Coimbra, Portugal

Metacognition–the ability of individuals to monitor one's own cognitive performance and decisions–is often studied empirically based on the retrospective confidence ratings. In experimental research, participants are asked to report how sure they are in their response, or to report how well their performance in high-level cognitive or low-level perceptual tasks is. These retrospective confidence ratings are used as a measure of monitoring effectiveness: larger difference in confidence ratings assigned to correct and incorrect responses reflects better ability to estimate the likelihood of making an error by an experiment participant, or better metacognitive monitoring ability. We discuss this underlying assumption and provide some methodological consideration that might interfere with interpretation of results, depending on what is being asked to evaluate, how the confidence response is elicited, and the overall proportion of different trial types within one experimental session. We conclude that mixing trials on which decision confidence is assigned when positive evidence needs to be evaluated and the trials on which absence of positive evidence needs to be evaluated should be avoided. These considerations might be important when designing experimental work to explore metacognitive efficiency using retrospective confidence ratings.

KEYWORDS

metacognition, decision confidence, retrospective confidence judgments, error detection, conscious awareness, metacognitive bias

Introduction

Living beings, including humans, constantly monitor environment and their own cognitive states in order to evaluate their past decisions (Kepecs et al., 2008; Smith, 2009). This is known as metacognitive monitoring – ability to evaluate one's own cognition – and it is based on error-detection mechanisms. Meta-monitoring is an important component of metacognition because it lays the foundation for meta-control, or adjusting future behavior in accordance with goodness of past decisions and the ratio of resolved/retained uncertainty about the state of the world and mind (Nelson and Narens, 1990; Drigas and Mitsea, 2020). Meta-monitoring relies on estimating the probability of an error on each decision (Ordin et al., 2020). If the estimated probability of an error is high, then people tend to assign lower confidence to their decisions than in the cases when estimated error probability is low. Hence, confidence ratings assigned tend to better discriminate between correct and incorrect decisions of those individuals who are better at estimating the probability of committing an error, i.e., better metacognitive monitoring skills. Efficient metacognition is reflected in larger difference in average confidence ratings assigned to correct and incorrect responses.

Metacognition is not necessarily correlated with cognitive performance (Flavell, 1979; Nelson, 1996), leading to under- or overconfidence bias. Some individuals may be very good at a particular cognitive task without realizing that their performance is high and thus assigning low confidence to their answers. Other individuals, on the contrary, may perform in the same task poorly without realizing it. Regardless of how well these individuals perform in cognitive tasks, their metacognitive skills are poor. On the other hand, individuals with good metacognitive abilities do not have to perform a cognitive task at a high level. Good metacognition is reflected in being able to realize how well the task is performed and adjust confidence ratings accordingly (Smith et al., 2003; Persaud et al., 2007). Metacognitive efficiency is studied most frequently by explicitly asking people in experimental setting to report how sure they are in their response on each trial. These confidence ratings are then used to measure metacognition. Three important concepts need to be distinguished in the study of metacognition: metacognitive sensitivity, metacognitive efficiency, and metacognitive bias, which will be defined below.

Metacognitive sensitivity is the accuracy with which participants discriminate between potentially correct and incorrect decisions. The percentage of correct responses on trials to which higher confidence is assigned, tends to be higher than the percentage of correct decisions to which lower confidence is assigned. The percentage of correct decisions on trials assigned the lowest confidence ratings should be at the chance level, given that overall performance (average accuracy of decisions) is above chance.

Metacognitive efficiency reflects how well confidence ratings discriminate between correct and incorrect responses. Efficient metacognition manifests as bigger differences in confidence between correct and incorrect decisions. In laboratory settings, this is often limited by the confidence rating scale. If participants are asked to report whether they are sure or not sure about a given answer using a binary scale, estimating metacognitive efficiency as the difference in confidence assigned to correct and incorrect responses becomes methodologically more challenging.

Metacognitive bias is the general tendency of an individual to assign higher or lower confidence ratings to his decisions. Metacognitive bias can expand or contract the scale for shifting confidence ratings up or down to reflect fluctuations in the degree of decision confidence. In extreme cases, over- or under-confidence can limit the discriminative aspect of confidence: when an under-confident individual correctly estimates that the likelihood of an error in a particular case is high, he may not be able to assign a lower confidence rating to another response because the base reference for his confidence is already at the lowest level. The opposite logic might also be true for over-confident individuals, who tend to assign ratings at ceiling, and are not able to push the ratings higher on trials where they estimate the likelihood of an error to be very low.

Since task performance and metacognitive bias can influence metacognitive sensitivity and efficiency (Galvin et al., 2003; Fleming and Lau, 2014; Rouault et al., 2018), Maniscalco and Lau (2012) proposed using a signal detection analytic approach. The basic idea behind this approach is that cognitive hits and correct rejections, to which high confidence rating is attached, are considered to be metacognitive hits, and cognitive hits and correct rejections, to which low confidence is attached, are considered metacognitive misses. Cognitive false alarms and misses with high confidence are metacognitive false alarms, and cognitive false alarms and misses with low confidence are metacognitive correct rejections. Confidence ratings do not have to be binary, leading to more precise modeling, as described below.

Metacognitive sensitivity is estimated as task performance (D') that would lead to the observed ROC curve for confidence ratings, given the absence of imprecision in assigned confidence ratings (modeling an ideal observer for confidence estimates). This fitted D' is referred to as meta-D' and may be higher or lower than D', correspondingly signaling better or worse metacognitive sensitivity. If metacognitive judgments and cognitive decisions are based on partially parallel processing streams (Fleming and Daw, 2017), participants can perform at chance in a cognitive or perceptual task, yet exhibit high metacognitive sensitivity, meaning that their confidence ratings will discriminate correct and incorrect decisions. Metacognitive efficiency within this framework is defined as metacognitive sensitivity relative to individual task performance (e.g., M-ratio, measured as meta-D'/D' or M-difference, measured as meta-D'-D'). Meta-D' shows how accurately correct and incorrect decisions are discriminated, while M-ratio shows how well confidence tracks performance on a particular task given an individual level of performance on this task. This then allows comparing meta-efficiency across tasks of different difficulty, in different domains and modalities (important is that the task structure remains the same across modalities and domains, Ruby et al., 2017).

While this approach has clear advantages (e.g., Maniscalco and Lau, 2012; Fleming, 2017), it is important to also be aware of its limitations. Metacognitive hits include task cognitive hits and correct rejections with high confidence, placing, for example, equal weight on cognitive correct responses, regardless of whether they are given based on positive evidence (detection of signals, i.e., hits) or absence of evidence (signals not present and not detected, i.e., correct rejections). However, Meuwese et al. (2014) showed that metacognition is superior on trials that require estimating positive evidence compared to trials that require estimating absence of evidence. Kanai et al. (2010) showed that in some cases cognitive misses and correct rejections are not discriminated by confidence ratings, while hits and false alarms are discriminated. That said, the structure of the task (Ruby et al., 2017) and individual decision making strategies (explore vs. exploit; reject vs. accept, Kanai et al., 2010; Meuwese et al., 2014) might lead to multiple individual differences in metacognitive sensitivity and efficiency, as measured by meta-D' and M-ratio.

In this report, we will look at discriminability of confidence ratings between correct and incorrect trials given based on presence and absence of evidence in an artificial language learning task with a yes/no recognition test. The task involves familiarizing people with a continuous sensory input with embedded recurrent discrete constituents. People detect and memorize these constituents during familiarization, and then they are subject to a recognition test, when they hear or see a token and need to respond whether this token is a constituent from the familiarization input or not. To measure metacognition, people are asked to assign confidence rating upon responding "yes" or "no" on each trial. This is a tricky test because on presenting the actual tokens from the familiarization input, participants need to estimate how sure they are in what they know. By contrast, on trials when foils are presented, participants' confidence ratings reflect how sure they are in what they do not know. In SDT approach, however, both types of responses are used within one framework. But we can calculate individual difference between correct and wrong responses separately for foils and actual constituents, hence tapping on whether people "know what they do not know" (on trials with foils), and whether metacognitive processing on trials with foils and actual tokens differs. This might have important methodological considerations for future experimental designs.

Method

The material for analysis was the same as described in details in Ordin et al. (2021). No experimental data was collected specifically for this study, an existing dataset (completely anonymized) was used, the ethical approval was obtained prior to collecting the primary dataset for the original study. For the readers' convenience, the material and the procedure is outline below, without details, which are presented in the original article. I used the data collected on 48 Spanish-Basque bilinguals from students' population at the university of the Basque country in Donostia-San Sebastian, Spain.

The data was obtained by running an artificial language learning experiments to investigate efficiency of statistical learning in the visual and auditory modalities on linguistic and non-linguistic material (semi-linguistic stimuli in the original dataset were not used in this analysis). The study was designed so that each participant performed all experiments, in a counter-balanced order.

For linguistic material, recurrent triples of syllables (further referred to as words) were embedded into a syllabic stream and presented *via* headphones in the auditory modality. In the visual modality, a different set of syllables was used to make another set of tri-syllabic words. Syllables were presented one by one in the middle of the screen. People listened/watched the familiarization sequence, and their task was to detect and memorize the words of this artificial language (explicit instructions were given as to what they will be tested on following the familiarization phase). Upon familiarization, we played *via* headphones or presented visually a tri-syllabic sequence. Participants had to report whether the sequence is a word from the artificial language or not, and how sure they were in their response (confidence rating was collected on a 4-point scale).

For non-linguistic material, we used fractals in the visual modality and environmental sounds in the auditory modality. The sounds/fractals were arranged into recurrent triplets embedded into familiarization input, and participants were explicitly instructed to detect and memorize these sequences. A yes/no recognition test followed.

For the recognition test in the auditory modality, eight words/sequences were created. The tokens for the test represented either recurrent sequences from the familiarization input (aka words) or foils. On foils, the transitional probabilities between separate elements (syllables/fractals/sounds) were 0% (i.e., the consecutive elements in the foils never occurred consecutively in the familiarization input). Eight foils preserved the ordinal position of elements, and eight foils violated the ordinal position of the elements in the words/sequences (i.e., if a particular element was used in the unit-initial position, it could only be used in the foil-medial or foil-final position).

In the visual modality, the number of words and foils was reduced by two. The order of sessions (modalities*domains) were counterbalanced across participants. During the tests, each token was used twice, yielding 48 trials in the auditory modality and 24 trials in the visual modality on each type of material.

Results

Data from one participant was discarded because he always gave the same confidence rating across all trials. The remaining data was screened for outliers (defined as data values exceeding 3SE deviations from the mean in z-transformed scores) and for deviations from normality (using Kolmogorov–Smirnov tests). Neither significant deviations from normality nor extreme outliers capable of distorting the test results were detected.

Analysis of metacognitive sensitivity

To analyze metacognitive sensitivity, we calculated the percentage of correct trials for all trials on which participants assigned high vs. low confidence ratings (as a number of participants did not use extreme confidence ratings at all, we lumped together all responses with confidence ratings 3-"sure" and 4-"absolutely sure" as high-confidence trails, and responses with confidence ratings 2-"not very sure" and 1-"unsure" as low-confidence trails). Here, we calculated the number of responses to which high or low confidence was assigned, and the number of correct responses among these responses, and calculated the ratio multiplied by 100. If a participant gave only 5 responses with high confidence, but all 5 responses were correct (100%), his metacognitive sensitivity was considered to be higher than that of a participant who gave 20 responses with high confidence, but only 10 were correct (50%).

The difference in the percentage of correct responses with high vs. low confidence ratings was significant for all token types (both linguistic and non-linguistic, both in visual and auditory modalities). On triplets, as predicted, responses to which high confidence is assigned are more likely to be correct than responses to which low confidence is assigned. On foils, although, the trend is the opposite: responses with low confidence are more likely to be correct than those with high confidence. This pattern is evident in **Figure 1**. All paired 2-tailed t-tests comparing the number of correct responses per confidence level in each modality and domain were significant ($p < 0.0005$ after Bonferroni correction), except t-tests for both types of non-linguistic foils in the visual modality ($p > 0.5$ before correcting for multiple comparison), also confirming that people are not sensitive to the likelihood of an error when they need to estimate how likely it is that they do not know something (evaluate absence of knowledge).

A more insightful result section below is related to metacognitive efficiency.

Analysis of metacognitive efficiency

People who exhibit equally high metacognitive sensitivity may nevertheless differ in metacognitive efficiency, i.e., in the magnitude of the difference in confidence ratings assigned to correct and

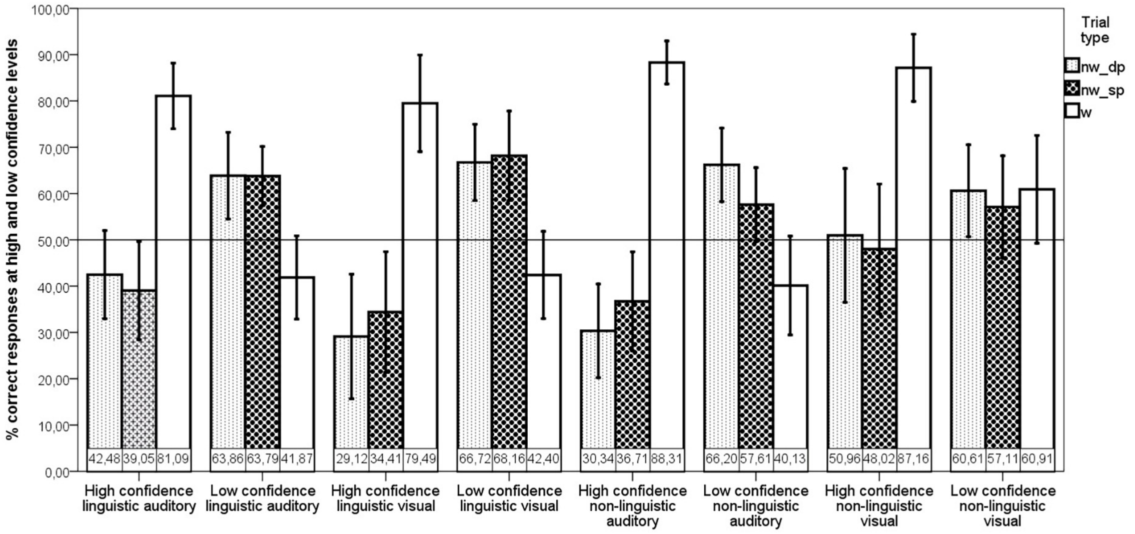


FIGURE 1
Meta-sensitivity measured as the percentage of correct responses for high and low confidence levels. If a participant gave only 5 responses with high confidence but all 5 responses were correct (100%), his metacognitive sensitivity was considered to be higher than that of a participant who gave 20 responses with high confidence, with only 10 correct (50%). Meta-sensitivity was calculated separately in the visual and auditory modalities, linguistic and non-linguistic domains, and on three different token types: random foils (nw_dp), ordered foils (nw_sp) and triplets (w). The horizontal line represents performance at the chance level.

incorrect responses. Figure 2 shows that correct responses are assigned higher confidence than incorrect responses only on trials with triplets, while on trials with foils, higher confidence is more often assigned to incorrect than correct responses. A series of two-tailed t-tests showed that the differences in mean confidence assigned to correct and incorrect responses were significant for all token types in both modalities and for both stimulus types – linguistic and non-linguistic - with all p-values, corrected, <0.0005. The only exceptions where this difference was not observed were for foils in the non-linguistic domain in the visual modality (corrected, p = 0.72 for ordered foils and p = 0.12 for random foils). As metacognition is evidenced by assigning a higher confidence rating to correct than to incorrect responses (Galvin et al., 2003; Persaud et al., 2007; Maniscalco and Lau, 2012), the data suggests that metacognitive processes did not operate on the trials in which foils were presented. This conclusion agrees with the analysis that revealed metacognitive sensitivity only on trials in which triplets were presented.

Overall, the data is in line with Kanai et al. (2010) and Meuwese et al. (2014), showing that metacognitive sensitivity is higher when people need to estimate how confident they are in what they know. Our results are even stronger suggesting that metacognition fails when people need to estimate their confidence in absence of evidence.

Earlier studies showed that on trials, in which the test tokens were endorsed, participants tend to assign higher confidence than on trials, in which the test tokens were rejected (Kanai et al., 2010; Maniscalco and Lau, 2014; Meuwese et al., 2014). To verify whether this pattern is observed in our sample, the data was re-analyzed conditional on the response type (*yes* vs. *no*), with *response type* and *correctness* as within-subject factors. In audio modality, the analysis on linguistic material revealed a significant effect of *correctness*, $F(1{,}44)$ = 15.06, $p < 0.001$, $\eta^2_p = 0.255$; and of *response type* $F(1{,}44)$ = 49.01, $p < 0.001$, η^2_p = 0.527, with insignificant interaction between the factors, $F(1{,}44)$ = 0.92, p = 0.34, η^2_p = 0.02. For each response type, correct responses were assigned higher confidence than wrong responses (confidence on *hits* was higher than on *false alarms*, and confidence on *correct rejections* was higher than on *misses*). On non-linguistic material in audio modality, the pattern was the same: a significant effect of *correctness*, $F(1{,}46)$ = 14.299, $p < 0.001$, η^2_p = 0.355; and of *response type* $F(1{,}46)$ = 67.64, $p < 0.001$, η^2_p = 0.59; yet the interaction between the factors was also significant, $F(1{,}46)$ = 15.18, $p < 0.001$, η^2_p = 0.25. The interaction is revealed in significant difference in confidence ratings between hits and false alarms, and lack of significant difference in confidence ratings between misses and correct rejections. These patterns are displayed on Figures 3A, B, for linguistic and non-linguistic material correspondingly.

In the visual modality on linguistic material, the analysis showed a significant effect of *response type* $F(1{,}43)$ = 62.99, $p < 0.001$, η^2_p = 0.59, while neither effect of *correctness*, $F(1{,}43)$ = 2.41, p = 0.128, η^2_p = 0.05, nor interaction between the factors, *correctness*, $F(1{,}43)$ = 1.04, p = 0.314, η^2_p = 0.02 turned out significant. On non-linguistic material in the visual modality, the pattern is identical to what we observed in the visual modality, with significant effect of *correctness*, $F(1{,}43)$ = 9.38, p = 0.004, η^2_p = 0.202; and of *response type* $F(1{,}43)$ = 13.34, $p < 0.001$, η^2_p = 0.265, and with insignificant interaction between the factors, $F(1{,}43)$ = 1.82, p = 0.185, η^2_p = 0.05. These patterns are displayed on Figures 3C, D, for linguistic and non-linguistic material correspondingly.

Higher confidence on correct responses than on incorrect responses on the constituents extracted from the familiarization sensory input and the reverse pattern on foils is possible if the participants exhibit a lenient response criterion on the cognitive task (i.e., if the tendency to endorse the presented test token is stronger than the tendency to reject the tokens, irrespective of their correctness). Given that each test token is presented twice during the recognition test, participants might develop a lenient criterion *via* familiarization with the test tokens after the first presentation. As “yes” responses tend to attract higher confidence compared to “no” responses, the lenient criterion may lead to higher confidence on

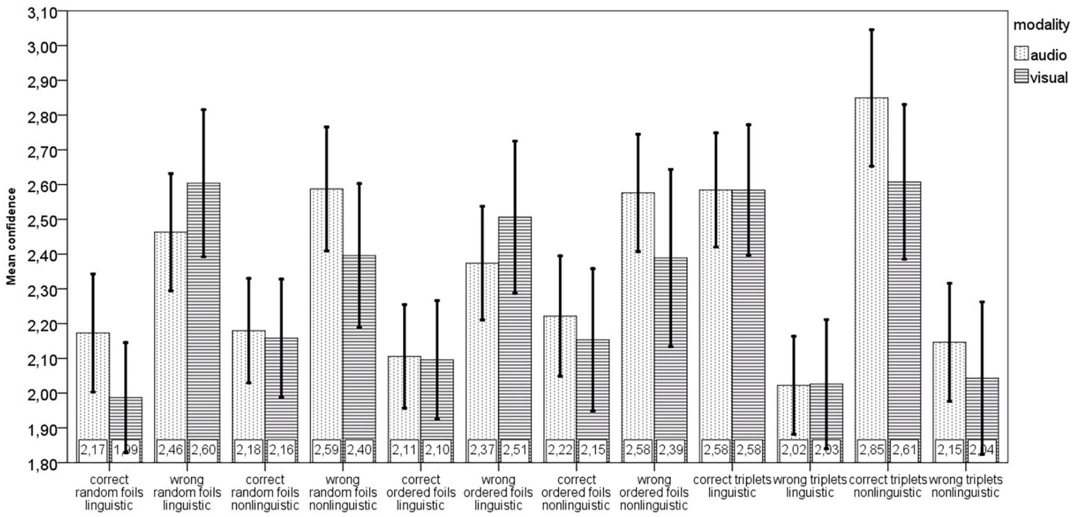


FIGURE 2

Confidence ratings assigned to correct and incorrect responses for different token types (random and ordered foils and triplets), stimulus types (linguistic and non-linguistic), and modalities (visual and auditory).

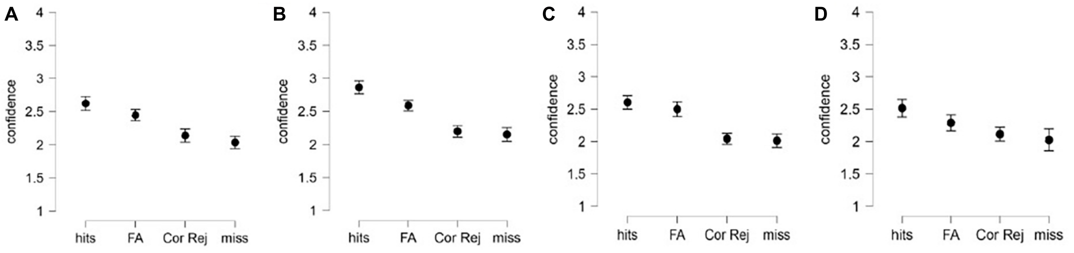


FIGURE 3

Confidence ratings assigned to correctly endorsed constituents (hits, "yes" responses), incorrectly endorsed foils (false alarms, "yes" responses), correctly rejected foils (correct rejections, "no" responses), and incorrectly rejected constituents (misses, "no" responses). Confidence ratings are represented separately for auditory **(A)** linguistic material, **(B)** non-linguistic material and visual, **(C)** linguistic material, and **(D)** non-linguistic material modalities. Error bars stand for 95%CI.

hits (endorsed constituents from the sensory input) than on *misses* (rejected constituents from the sensory input) and lower confidence on *correct rejections* (rejected foils) than *false alarms* (endorsed foils). To consider this possibility, we calculated the response bias using the classical SDT approach. Positive bias signals an overall tendency to endorse items and negative bias signals an overall tendency to reject items. A score of 0 indicates no bias, hence significant deviations from 0 (using four one-sample t-tests, separate for each material type and perceptual modality) reveal the overall tendency to accept or reject the test tokens (the normality assumption was tested by the Shapiro–Wilk test).

For linguistic material, the difference from zero was not significant, $t(47) = 0.98$, $p = 0.331$, $d = 0.14$ in the auditory modality and significant, albeit with low effect size, $t(47) = 2.305$, $p = 0.026$, $d = 0.33$ in the visual modality. For non-linguistic material, the difference from zero was significant and important, with a moderate effect size, both for the auditory, $t(47) = 3.57$, $p < 0.001$, $d = 0.52$, and for the visual, $t(47) = 2.944$, $p = 0.005$, $d = 0.42$, modalities. The result pattern is displayed in **Figure 4** (adapted from Ordin et al., 2021).

Whether lenient criterion fully accounts for the difference in confidence ratings on foils and constituents remains an open question because the bias to endorse the test tokens (i.e., to respond "yes") is not different between modalities and material types (Ordin et al., 2021), and that significant deviations from zero were not observed in the auditory condition on linguistic material, yet the confidence

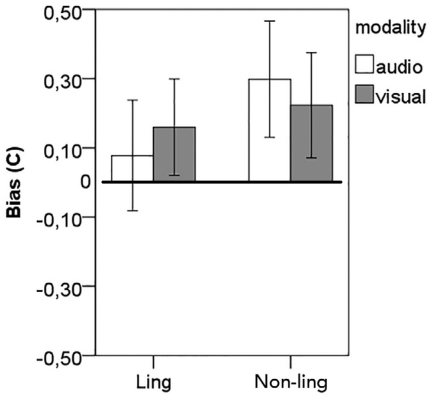


FIGURE 4

Response bias. A score of 0 indicates no bias, positive bias reflects a tendency to accept (endorse) test tokens (a tendency to respond "yes"). Error bars stand for 95%CI.

pattern conditioned to the stimuli type (*foil* vs. *actual constituent*) was the same across all modalities and material types. Besides, the difference in confidence ratings assigned to correct and wrong responses on trials, in which constituents were presented is larger than the difference in confidence on correct and wrong responses on trials, in which foils were presented: Δ_foils < Δ_constituents, $p < 0.001$ in paired t-tests for each modality on both linguistic and non-linguistic material.

Taken together, the data suggest that metacognitive monitoring is differentially affected on constituents and foils, with metacognitive monitoring on trials when constituents are presented being stronger than on trials when foils are presented (weak version of the hypothesis), or metacognitive failure when foils are presented (strong version of the hypothesis, which required further empirical testing).

Discussion

We found that the confidence ratings are discriminative of correct and incorrect responses in the expected manner (i.e., higher on correct than on incorrect responses) only on trials when people had to recognize words or recurrent sequences from the familiarization input. On trials in which foils were presented the confidence ratings revealed the reverse pattern. This was confirmed across four experimental sessions: in visual and in auditory modalities both on linguistic and non-linguistic material. This suggests that metacognition is efficient in those cases when people need to evaluate how well they have learnt something. When people need to report how sure they are in what they have not been learning, metacognition fails (or we fail to capture metacognitive efficiency based on retrospective confidence ratings). This should be considered in experimental design, in terms of wording for the tasks and structure of the trials.

Intriguingly, there is no observable difference in confidence assigned to foils of different types, although random foils should be easier to reject because people need to detect the novel element at the triplet-initial position in order to be able to reject the foil, the confidence in decision should increase once the second element, also violating the expectations, is processed, leading to higher confidence on rejected foils. On ordered foils, besides positional information, relational information (which element is expected given preceding one(s). This can be calculated once the first element has happened on a test token, giving less time for confidence accumulation toward the end of the triplet. According to the Relational Complexity Theory (Halford et al., 1998), in the process of conceptual segmentation (i.e., during the learning stage of the artificial language learning experiments), new representations of segmented units are formed by reducing complexity *via* collapsing dimensions (sources of variation). A new holistic representation is easier to process, but different sources of variation within the segmented and consolidated unit can no longer be unpacked. Hence we did not observe the effect of difference in complexity on confidence. However, to further explore the potential relation between complexity and confidence, in the future studies we will need to focus on the foils that violate relational information (ABC–target vs. ABD–foil), introducing multiple dimensions (sources of variability) of complexity.

Why we observed a reverse result pattern in how confidence is assigned on trials with foils remains unclear. We expected the confidence being not discriminative between correct and incorrect responses, which would indicate a poor metacognitive efficiency. Neither did we expect any difference in the number of correct responses per confidence level on the trails with foils. However, on foils trials participants consistently assigned significantly higher rating to wrong responses. We propose several explanations why our expectations were violated on foils trials.

(1) Correct response on foils is rejection, while correct response on words is acceptance. Rejection and acceptance might rely on differential neuro-cognitive mechanisms, and monitoring of these mechanisms might also differ, leading to differential result patterns on trials with foils and words.
(2) We have twice as many foils as words in each session; hence people should reject items twice as frequently as accept items. However, given the dual choice, participant might have expected an equal distribution of trials when they need to accept and reject presented test tokens. Thus, with each new token that had to be correctly rejected, participants' confidence in their response might decrease.
(3) Accepting items elicits higher confidence overall, leading to higher confidence on correct responses on trials with words and on incorrect responses on trials with foils. In other words, we can evaluate the changes in mental states based on evidence, but not absence of evidence. This is an important confounding factor that also undermines the experimental design that incorporates the analysis of "yes" versus "no" responses with the analysis of responses on items (or rules for constructing novel items) that have been learnt and those that have not been learnt. A potentially promising approach might be based on the differences in searching or decision-making time on the trials in which the uncertainty that the "award" is expected is high, versus trials in which the uncertainty is low).

Another important consideration is the degree of conscious awareness into metacognitive judgments. Metacognition relies both on conscious and unconscious processing (Nelson, 1996; Kentridge and Heywood, 2000; Jachs et al., 2015). The nature of this task diminishes the contribution of the latter because people, when explicitly asked to rate their confidence, are more likely to consciously contemplate on their decisions (Ordin and Polyanskaya, 2021). This might highlight awareness of what is learnt and known (Drigas et al., 2023), but hinders awareness of what has not been learnt, yielding different result patterns in terms of retrospective confidence on trials with foils and words. Alternative procedures are also necessary to study the contribution of unconscious processing into metacognitive efficiency because the ability to discriminate on the basis of confidence is often assumed to rely on conscious awareness of stimuli (Smith et al., 2003; Persaud et al., 2007). Explicit instructions to evaluate one's performance with confidence ratings skew the balance between conscious and unconscious processes in metacognition in favor of the former.

These methodological considerations do not undermine the usefulness of the signal detection theoretic approach to modeling metacognition using confidence ratings. Hits attract higher confidence rating than false alarms, but correct rejections attract lower confidence rating than misses. However, the difference in confidence between hits and false alarms is greater than between correct rejections and misses, thus the modeling approach nevertheless provides useful information is we need to compare metacognition between groups, between tasks, or between modalities/domains. The signal detection modeling approach offers a clear advantage when comparing metacognition across tasks, domains, and modalities that vary in terms of task performance and metacognitive bias, which affect alternative measures of metacognition based on retrospective confidence (Masson and Rotello, 2009; Maniscalco and Lau, 2012; Barrett et al., 2013). Also, it provides a clearer link to conscious awareness because M-ratio effectively shows the extent to which a metacognitively ideal observer is aware of his task performance. Also, as meta-D' and D' are measured in the same units, sensitivity in the task and metacognitive sensitivity can be explicitly compared, and given larger difference in confidence between hits and false alarms than between misses and correct rejections, the SDT will nevertheless yield valid results. However, care should be taken in regard how questions are asked and whether people are indeed asked to evaluate what they know rather than what they do not know (in the latter case, differences between conditions might be diminished due to reverse confidence patterns on hits and false alarms versus misses and correct rejections. In statistical learning experiments, it might be useful, for example, to implement alternative forced-choice methods, when people need to select between a foil and a word, which of the two tokens is embedded into familiarization stream (e.g., Ordin et al., 2020; Ordin and Polyanskaya, 2021). Such trials always include evaluation of what people (supposedly) know. Avoiding mixing trials that require estimating positive evidence and trails that require estimating absence of evidence will increase the strength of the SDT analytic approach.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://doi.org/10.6084/m9.figshare.20995345.

Ethics statement

A secondary dataset that was originally collected for other research purposes was used for this study. Ethical review and approval, and written informed consent, were not required for re-analysis of the existing dataset in accordance with the national legislation and the institutional requirements.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

This study has received funding from the European Research Council Executive Agency (ERCEA), grant agreement 101040850. Open access was supported by the Alexander von Humboldt Stiftung and the Open Access Publication Funds of the Göttingen University. Views and opinions expressed are, however, those of the author only and do not necessarily reflect those of the funders.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Barrett, A. B., Dienes, Z., and Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychol. Methods* 18, 535–552. doi: 10.1037/a0033268

Drigas, A., and Mitsea, E. (2020). The 8 pillars of metacognition. *Int. J. Emer. Technol. Learn.* 15, 162–178. doi: 10.3991/ijet.v15i21.14907

Drigas, A., Mitsea, E., and Skianis, C. (2023). Meta-learning: A Nine-layer model based on metacognition and smart technologies. *Sustainability* 15:1668. doi: 10.3390/su15021668

Flavell, J. H. (1979). Metacognition and cognitive monitoring, A new area of cognitive-development inquiry. *Am. Psychol.* 34, 906–911. doi: 10.1037/0003-066X.34.10.906

Fleming, S. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neurosci. Conscious.* 1:mix007. doi: 10.1093/nc/nix007

Fleming, S. M., and Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychol. Rev.* 124, 91–114. doi: 10.1037/rev0000045

Fleming, S., and Lau, H. C. (2014). How to measure metacognition. *Front. Hum. Neurosci.* 8:443. doi: 10.3389/fnhum.2014.00443

Galvin, S. J., Podd, J. V., Drga, V., and Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychon. Bull. Rev.* 10, 843–876. doi: 10.3758/BF03196546

Halford, G. S., Wilson, W. H., and Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behav. Brain Sci.* 21, 803–831. doi: 10.1017/S0140525X98001769

Jachs, B., Blanco, M., Grantham-Hill, S., and Soto, D. (2015). On the independence of visual awareness and metacognition: A signal detection theoretic analysis. *J. Exp. Psychol.* 41, 269–276. doi: 10.1037/xhp0000026

Kanai, R., Walsh, V., and Tseng, C. H. (2010). Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. *Conscious. Cogn.* 19, 1045–1057. doi: 10.1016/j.concog.2010.06.003

Kentridge, R. W., and Heywood, C. A. (2000). Metacognition and awareness. *Conscious. Cogn.* 9, 308–312. doi: 10.1006/ccog.2000.0448

Kepecs, A., Uchida, N., Zariwala, H., and Mainen, Z. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature* 455, 227–231. doi: 10.1038/nature07200

Maniscalco, B., and Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.* 21, 422–430. doi: 10.1016/j.concog.2011.09.021

Maniscalco, B., and Lau, H. (2014). "Signal detection theory analysis of type 1 and type 2 data: Meta-d, response-specific meta-d, and the unequal variance SDT mode", in *The cognitive neuroscience of metacognition*, eds S. M. Fleming and C. D. Frith (New York, NY: Springer), 25–66.

Masson, M. E., and Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *J. Exp. Psychol.* 35, 509. doi: 10.1037/a0014876

Meuwese, J. D., van Loon, A. M., Lamme, V. A., and Fahrenfort, J. J. (2014). The subjective experience of object recognition: Comparing metacognition for object detection and object categorization. *Attent. Percept. Psychophys.* 76, 1057–1068. doi: 10.3758/s13414-014-0643-1

Nelson, T. (1996). Consciousness and Metacognition. *Am. Psychol.* 51, 102–116. doi: 10.1037/0003-066X.51.2.102

Nelson, T., and Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychol. Learn. Motiv.* 26, 125–173. doi: 10.1016/S0079-7421(08)60053-5

Ordin, M., and Polyanskaya, L. (2021). The role of metacognition in recognition of the content of statistical learning. *Psychon. Bull. Rev.* 28, 333–340. doi: 10.3758/s13423-020-01800-0

Ordin, M., Polyanskaya, L., and Samuel, A. (2021). An evolutionary account of intermodality differences in statistical learning. *Ann. N. Y. Acad. Sci.* 1486, 76–89. doi: 10.1111/nyas.14502

Ordin, M., Polyanskaya, L., and Soto, D. (2020). Metacognitive processing in language learning tasks is affected by bilingualism. *J. Exp. Psychol.* 46, 529–538. doi: 10.1037/xlm0000739

Persaud, N., McLeod, P., and Cowey, A. (2007). Postdecision wagering objectively measures awareness. *Nat. Neurosci.* 10, 257–261. doi: 10.1038/nn1840

Rouault, M., Seow, T., Gillan, C. M., and Fleming, S. M. (2018). Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biol. Psychiatry* 84, 443–451. doi: 10.1016/j.biopsych.2017.12.017

Ruby, E., Giles, N., and Lau, H. (2017). Finding domain general metacognitive mechanisms requires using appropriate tasks. *bioRxiv* [Preprint] doi: 10.1101/211805

Smith, J. D. (2009). The study of animal metacognition. *Trends Cogn. Sci.* 13, 389–396. doi: 10.1016/j.tics.2009.06.009

Smith, J. D., Shields, W. E., and Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behav. Brain Sci.* 26, 317–339. doi: 10.1017/S0140525X03000086

TYPE Perspective
PUBLISHED 28 February 2023
DOI 10.3389/fpsyg.2023.1129835

OPEN ACCESS

EDITED BY
Sergio Da Silva,
Federal University of Santa Catarina,
Brazil

REVIEWED BY
Bin Liu,
Hainan University,
China
Newton Da Costa,
Federal University of Santa Catarina,
Brazil

*CORRESPONDENCE
Johan. E. (Hans) Korteling
✉ hans.korteling@tno.nl

SPECIALTY SECTION
This article was submitted to
Cognitive Science,
a section of the journal
Frontiers in Psychology

RECEIVED 22 December 2022
ACCEPTED 06 February 2023
PUBLISHED 28 February 2023

CITATION
Korteling JE, Paradies GL and Sassen-van Meer JP (2023) Cognitive bias and how to improve sustainable decision making. *Front. Psychol.* 14:1129835.
doi: 10.3389/fpsyg.2023.1129835

Cognitive bias and how to improve sustainable decision making

Johan. E. (Hans) Korteling*, Geerte L. Paradies and Josephine P. Sassen-van Meer

TNO Netherlands Organization for Applied Scientific Research, The Hague, Netherlands

The rapid advances of science and technology have provided a large part of the world with all conceivable needs and comfort. However, this welfare comes with serious threats to the planet and many of its inhabitants. An enormous amount of scientific evidence points at global warming, mass destruction of bio-diversity, scarce resources, health risks, and pollution all over the world. These facts are generally acknowledged nowadays, not only by scientists, but also by the majority of politicians and citizens. Nevertheless, this understanding has caused insufficient changes in our decision making and behavior to preserve our natural resources and to prevent upcoming (natural) disasters. In the present study, we try to explain how systematic tendencies or distortions in human judgment and decision-making, known as "cognitive biases," contribute to this situation. A large body of literature shows how cognitive biases affect the outcome of our deliberations. In natural and primordial situations, they may lead to quick, practical, and satisfying decisions, but these decisions may be poor and risky in a broad range of modern, complex, and long-term challenges, like climate change or pandemic prevention. We first briefly present the social-psychological characteristics that are inherent to (or typical for) most sustainability issues. These are: experiential vagueness, long-term effects, complexity and uncertainty, threat of the status quo, threat of social status, personal vs. community interest, and group pressure. For each of these characteristics, we describe how this relates to cognitive biases, from a neuro-evolutionary point of view, and how these evolved biases may affect sustainable choices or behaviors of people. Finally, based on this knowledge, we describe influence techniques (interventions, nudges, incentives) to mitigate or capitalize on these biases in order to foster more sustainable choices and behaviors.

KEYWORDS
cognitive bias, nudging, decision making, behavioral influence, sustainability, sustainable behavior

1. Introduction: The challenges of human welfare

Supported by science and technology, the world has undergone an explosively rapid change in only a few centuries which offers humanity enormous practical advantages in a large number of areas. Misery and misfortune as a result of food shortages, diseases, and conflicts that were previously considered unsolvable have been adequately tackled (Pinker, 2018). A large part of the world has achieved unprecedented economic growth, and on the waves of globalization, it is assumed that the less developed countries can in principle also benefit from this development (Harari, 2017). However, the technologies we use to increase our welfare today have effects, not only across the whole planet, but also stretching far into the future. In the wake of our pursuit of prosperity, humanity has created a number of new, and possibly even greater, problems. The economic growth, that has provided us with an abundance of food, energy, medicines, and living comfort, simultaneously destabilizes the ecological balance. To date, scientists have gathered broad and convincing evidence that under the influence of fossil energy consumption, there is a rapid global warming that may have devastating consequences for the health, wellbeing, and flourish of future generations. This includes sea level rise, droughts, floods, water shortage, and refugee flows (e.g., Meadows et al., 1972; Meadows, 1997; Kates and Parris, 2003; Millenium Ecosystem Assessment, 2005; Biermann et al., 2012; IPCC, 2013, 2014, 2021, 2022; Steffen et al., 2015). Other examples of ecological destabilization are: environmental pollution, pandemics, and massive extinction of plant and animal species. All these ecological imbalances pose a serious threat to the continued existence of the world and the survival of our civilization. In the Stone Age, the average person had around 4,000 cal. of energy per day at their disposal. Today, the average American uses around 230,000 cal., sixty times as much (Harari, 2017). To offer everyone in this world the same standard of living as persons living in the USA, we would need at least four planets, but we only have one (OECD, 2012). At the same time, the world seems hesitating to take decisive preventative action.[1] So, despite that most scientists and an increasing number of politicians and citizens acknowledge these facts, this common understanding has not caused much change in our collective behavior. Humanity thus seems to lack the kind of rationality or wisdom that is needed to make substantial financial, social, or material changes in order to stop possible disasters that threaten long-term wellbeing, i.e., to create a world in which people can flourish and be happy.

1.1. Cognitive bias in sustainability issues

How can this be? Human decision making can be quite questionable at times. For example, it often seems to underestimate the long-term dangers of things like global warming and species extinction. This can make even major future threats seem insufficient motivation for determined action (Berger, 2009). In general, we see these types of typical, and often flawed, decision making patterns in many different contexts of our society (Eigenauer, 2018). For instance, Flyvbjerg (2009) showed that 9 out of 10 transportation infrastructure projects end up in large cost overrun, which did not improve over time, even over a period of 70 years. Other examples of persisting problems that for a major part follow from poor decision making are: improper and incorrect diagnoses as well as harmful patient decisions in medicine and health care (Croskerry, 2003; Groopman, 2007); overly optimistic growth assessments and ill-advised lending policies in global finance (Shiller, 2015); optimistic decision making in personal finance, like susceptibility to scams (Modic and Lea, 2013); against all knowledge continue a chosen course or investment with negative outcomes rather than alter it (Arkes and Blumer, 1985; Garland and Newport, 1991); perpetuating injustice through personal prejudice and unjust sentencing (Benforado, 2015); and accepting superstitions or conspiracy theories while rejecting scientific findings that contradict these beliefs (Yasynska, 2019).

In this article, we will focus on how the human brain and its evolved psychological characteristics affect people's decision making. Effects of the workings of our brain and of our evolutionary heritage on decision making manifest most prominently in cognitive biases (Kahneman et al., 1982; Hastie and Dawes, 2001; Shafir and LeBoeuf, 2002; Haselton et al., 2005; van Vugt et al., 2014; Korteling et al., 2018). Cognitive biases can be generally described as systematic, universally occurring, tendencies, inclinations, or dispositions in human decision making that may make it vulnerable for inaccurate, suboptimal, or wrong outcomes (e.g., Tversky and Kahneman, 1974; Kahneman, 2011; Korteling and Toet, 2022). Well-known examples of biases are hindsight bias (once we know the outcome, we tend to think we knew that all along), tunnel vision (when we are under pressure, we tend to overfocus on our goal and ignore all other things that are happening), and confirmation bias (we tend to only see information that confirms our existing ideas and expectations). People typically tend to pursue self-interest at the expense of the community (Tragedy of the commons). We tend to over-value items we possess (Endowment effect) and we have a strong urge to persist in courses of action, with negative outcomes (Sunk-cost fallacy). What is more, biased decision making feels quite natural and self-evident, such that we are quite blind to our own biases (Pronin et al., 2002). This means we often do not recognize it, and therefore do not realize how our biases influence our decision making.

Cognitive biases are robust and universal psychological phenomena, extensively demonstrated, described, and analyzed in the scientific literature. In a wide range of different conditions, people show the same, typical tendencies in the way they pick up and process information to judge and decide. In line with their systematic and universal character, cognitive biases are also prominent in societal issues and policymaking (e.g., Levy, 2003; McDermott, 2004; Mercer, 2005; Baron, 2009; Flyvbjerg, 2009; Vis, 2011; Arceneaux, 2012; Shiller, 2015; Bellé et al., 2018). For example, Arceneaux (2012) has shown that in discussing political arguments, individuals are more likely to be persuaded by arguments that evoke loss aversion, even in the face of a strong counterargument. And it has been demonstrated in many instances that policy makers tend to make risk-aversive decisions when they expect gains, whereas when facing losses they accept taking more risk (e.g., McDermott, 2004; Vis, 2011).

There are already many publications on cognitive biases showing how human psychological tendencies underly the choices and behaviors of people (e.g., Kahneman et al., 1982; Shafir and LeBoeuf, 2002; Kahneman, 2011). There is also some literature on which biases and human mechanisms play a role in our difficulties with preventing climate change (e.g., Gifford, 2011; van Vugt et al., 2014; Marshall, 2015; Stoknes, 2015). However, there is still lack of insight into how biases play a role in the process of environmental policymaking and how this knowledge may be used to deal with the major systemic challenges that the modern world is confronted with. Despite their possible substantial effects on society and human wellbeing, cognitive biases have never been a serious matter of concern in the social and political domain (Eigenauer, 2018). In this paper, we will therefore analyze the constellation of psychological biases that may hinder behavioral and policy practices addressing sustainability challenges. We will also look for ways to mitigate the potential negative effects of biases through influence techniques, like nudging (e.g., Thaler and Sunstein, 2008).

[1] The problem of climate change was put on the agenda by the Club of Rome, with their report *Limits to Growth* (Meadows et al., 1972). Since then numerous countries have agreed that action is needed. Climate goals were set numerous times, of which the last two were the Paris climate goals (Paris Climate Conference, COP21, 2015, and COP26 in Glasgow, 2021).

1.2. The rationale and drawback of biases

Given the inherent constraints of our information processing system (i.e., the limited cognitive capacities of the human brain) our intuitive inclinations, or heuristics, may be considered effective, efficient, and pragmatic. And indeed, intuitive or heuristic decision making may typically be effective in; natural (primal) conditions with time-constraints, lack (or overload) of relevant information, when no optimal solution is evident, or when we have built up sufficient expertise and experience with the problem (Simon, 1955; Kahneman and Klein, 2009; Gigerenzer and Gaissmaier, 2011). In these cases, the outcomes of heuristic decision making may be quite acceptable given the invested time, effort, and resources (e.g., Gigerenzer et al., 1999).

The fact that heuristic thinking deals with information processing limitations and/or data limitations (Simon, 1955) does not alter the fact that many of our judgments and decisions may systematically deviate from what may be considered optimal, advisable, or utile given the available information and potential gain or risk (Shafir and LeBoeuf, 2002). This has been demonstrated by a large body of literature, showing how cognitive heuristics or biases may lead to poor decisions in a broad range of situations, even including those without complexity, uncertainty, or time constraints (Korteling et al., 2018). Imagine, for instance, a board of directors that has to decide about the continuation of a big project. Typically, the more they have invested so far, the less likely they are to pull the plug. This is not rational (and is therefore called the sunk cost fallacy), because what should matter is what the costs and benefits will be from this point forward, not what has already been spent. The Sunk-cost fallacy, like various other psychological biases affecting decision making, may continuously pop up in the world we live in. Examples are the Anchoring bias (Tversky and Kahneman, 1974; Furnham and Boo, 2011), Authority bias (Milgram, 1963), Availability bias (Tversky and Kahneman, 1973, 1974), and Conformity bias (Cialdini and Goldstein, 2004).

A large number of different biases have been identified so far and specific biases are also likely to occur in the domain of public decision making. By public decision making, we mean not only collective and democratic decision making, but also individual decision making. For different kinds and domains of decision making, different biases may occur. It may be expected that in decision making within the sustainability domain, certain (categories of) biases may more often occur than others. In this paper, we try to present the most relevant biases and the associated nudges, focusing on public decision making with regard to sustainability challenges.

2. Methods

Decision making in our modern society may be done on an individual basis, but may also involve many participants or stakeholders with their own perspectives and background, i.e., citizens, policy makers, company representatives, and interest groups (e.g., Steg and Vlek, 2009). To come to a comprehensive understanding of which psychological biases are likely to pop up in this context, we selected those biases that would likely be most prominent, given the typical (psychological) characteristics of sustainability issues. Next, we described interventions or influence techniques (incentives, nudges) to overcome, mitigate, or capitalize on these biases. This was done in three steps.

Step 1: Defining psychological characteristics of sustainability problems

Sustainability issues have characteristics that may evoke certain biases. Here, we define "sustainability" as: a balanced development in which the exploitation of resources, the direction of investments, the orientation of technological development, and institutional change are all in harmony and enhance both current and future potential to meet long-term wellbeing. First, on the basis of the literature (e.g., Schultz, 2002; Steg and Vlek, 2009; van Vugt, 2009; van Vugt et al., 2014; Engler et al., 2018; Toomey, 2023) and a workshop with experts we defined a set of general social-psychologically relevant characteristics or factors, like "experiential vagueness" or "long-term effects" or "threat of the status quo" that are associated with most sustainability issues.

Step 2: Biases per sustainability characteristic

Each characteristic of sustainability issues may relate to a few specific biases that may hamper sustainable choices and behaviors of people. For example, the long-term character of sustainability implies may be in conflict with our tendency to short-term thinking (Hyperbolic time discounting) or the tendency to underestimate both the likelihood of a disaster and its possible consequences, and to believe that things will always function the way they normally function (Normalcy bias). The subsequent identification of thinking tendencies and biases related to these characteristics was based on the literature entailing overviews of multiple biases (e.g., Korteling et al., 2020a), a Neuro-Evolutionary Bias Framework (Korteling et al., 2020a,b; Korteling and Toet, 2022), and on the literature on cognitive biases and sustainability challenges (e.g., Gardner and Stern, 2002; Penn, 2003; Fiske, 2004; Wilson, 2006; Steg and Vlek, 2009; van Vugt, 2009; van Vugt et al., 2014; Marshall, 2015; Engler et al., 2018).

Step 3: Influence techniques per sustainability characteristic

Also, for each group of biases, some relevant intervention techniques that can be used, by for example government or policy makers, were briefly described. These interventions, incentives, or nudges, may be applied to mitigate the relevant biases or to capitalize on them for the purpose of stimulating decision making that is more in line with sustainability goals in the context of the current world. On the basis of a previous literature review (Korteling et al., 2021), we have chosen not to advocate specific educational approaches, aiming at bias mitigation training in order to foster sustainable decision making. Instead, our approach aims at interventions with regard to the context or environment in which people live order to promote more sustainable choices.

Example of the approach

Finally, we will illustrate our approach with the help of an example: A conflict between personal versus community interest is a typical characteristic that is associated with sustainability issues. Natural selection has favored individuals who prioritize personal benefits over those of unrelated others (Hardin, 1968; van Vugt et al., 2014). This means that making choices in the public interest is often hindered by our personal interests (Step 1). Sustainability also often involves a trade-off between personal interests, such as driving a car or flying, against collective interests, such as fresh air and a peaceful environment. This conflict relates to the bias called the *Tragedy of the commons*, i.e., the tendency to prioritize one's own interests over the common good of the community (Step 2). Because we share our genes with our relatives, this tendency may be countered by invoking kinship as a nudge. Pro-environmental actions or appeals may thus be more effective if they emphasize the interests of our ingroup, children, siblings, and grand-children (Step 3).

3. Most relevant psychological characteristics of sustainability challenges

Below, we list a set of prominent psychological characteristics that we consider relevant for sustainability issues. Although biases are inherent to the thinking and decision making in all people, it may be supposed that biases may differ depending on peoples' places, functions, and roles in decision situations. On the other hand, there are many mutual influences and dependencies in the policymaking arena. Therefore, we have decided not to make clear distinctions between the specific roles people play in this arena. So, we do not discern biases for citizens, politicians or policy makers.

- *Experiential vagueness:* Sustainability problems are slowly and gradually evolving. Therefore, the impact of the issue is difficult, if not impossible, to perceive or experience directly with our body and senses. Our knowledge of the issue is largely built on indirect and abstract cognitive information, i.e., on conceptual reasoning, abstract figures, written papers, and quantitative models.
- *Long-term effects and future risk*: The negative consequences of green practices follow directly, whereas the positive aspects of green practices may emerge only after many years in the (far) future. The same counts for the positive consequences of not taking green action. In addition, sustainability concerns an unknown future with an abundance of possibilities that easily go beyond our imagination.
- *Complexity and uncertainty:* The sustainability issue is very complicated (socially, technically, logistically, economically) and even "wicked." Being able to judge and reason over most topics within the field requires multi- and transdisciplinary knowledge. Sustainability challenges are (therefore) accompanied by a high degree of uncertainty about their future progression and how it should be tackled and addressed.
- *Threat to the status quo:* Many sustainability measures more or less have impact on (sometimes even threaten) our established way of living and basic societal infrastructure. When new measures have an impact on our "normal," established way of living and basic societal infrastructure, this may be experienced as a threat that will result in losing our freedom and/or comfort ("fear of falling").
- *Threat of social status*: Many environmental problems result from a desire to possess or consume as much as possible, instead of consuming "enough" for a good life. Consumptive behavior and high energy consumption are intrinsically related to high social status, which is something most people do not want to lose.
- *Social dilemma's:* The sacrifices that have to be made in order to foster sustainability are mainly beneficial for the collective, whereas direct individual gains are often limited. In this "social dilemma," humans tend to prioritize direct personal interests relative to more sustainable ones that benefit the planet.
- *Group pressure*: Norms, values, and standards for what is considered as 'normal' or what is considered "desirable" are determined and reinforced by group pressure. Also with regard to green choices, we are often more strongly influenced by the behaviors and opinions of our peers than by our personal views and attitudes toward conservation.

4. Biases and interventions per psychological sustainability characteristic

For each of the above-mentioned general psychological characteristics of sustainability issues, the next subsections will provide an analysis and inventory of the (kinds of) cognitive biases that are probably most relevant and critically involved in the associated public and political decision making processes. Finally, for each general characteristic, influence techniques (interventions) to mitigate or capitalize on the relevant/critical biases will be briefly described. These interventions are based on the literature concerning "psychological influence" (e.g., Jowett and O'Donnell, 1992; Cialdini, 2006; Adams et al., 2007; Cialdini, 2009; Hansen, 2013; Heuer, 2013; Korteling and Duistermaat, 2018; Toomey, 2023). The influence techniques have an informational nature. They can be utilized in public communication, education, and policy making, especially in communication to the public, in different forms of media. Because the biases mentioned show a great deal of overlap and similarity—it was more about groups or types of similar biases—we chose not to make explicit links between specific biases and the associated nudge.

4.1. Experiential vagueness

Social scientists have long been puzzled as to why people are so poor at recognizing environmental risks and ignore global environmental hazards (Slovic, 1987; Hardin, 1995). Such apathy is probably a product of our evolutionary heritage that produced a brain that is optimized to perform biological and perceptual-motor functions (Haselton and Nettle, 2006; Korteling et al., 2018; Korteling and Toet, 2022). For example, the vertebrate eye evolved some 500 billion years ago, compared to 50,000 years ago for human speech; while the first cave drawings are dated at 30,000 years, compared to the earliest writing system approximately 5,000 years ago (Parker, 2003; see also Grabe and Bucy, 2009). This comparatively more ancient visual perceptual and communicative apparatus enables us to quickly extract meaning from eye-catching images (Powel, 2017). In addition, there was always a tangible link between behavior and the environment. That is: if you do not eat, you will become hungry and search for food. If it starts raining, you may look for shelter in order to prevent becoming wet. A critical difference between the modern world and our ancestral environment is that we rarely see, feel, touch, hear, or smell how our behaviors gradually impact the environment (Uzzell, 2000; Gifford, 2011). Because our ancestors were not confronted with the relatively remote, slowly evolving, or abstract problems (Toomey, 2023), we probably are not well-evolved to be alarmed when confronted with potential or novel dangers that we cannot directly see, hear, or feel with our perceptual systems (van Vugt et al., 2014).

The human senses and nervous system show a gradual decrease in responsiveness to constant situations. In general, we are more sensitive to, and more easily triggered by, sudden changes and differences in the stimulus (contrasts). Because of this neural adaptation, we often may have difficulty with perceiving and appreciating slow and gradual processes of change. Therefore, the gradual changes that are implied in our environment, like global warming, are not very easily noticed. So, most people are generally not really alarmed by the gradual evolving and remote environmental challenges that the world is facing. This may contribute to the relatively low public interest in the issue of environmental threats such as global climate change, pollution of the oceans, extinction of species, the negative health effects of particulate matter, and decreasing biodiversity (Swim et al., 2011).

4.1.1. Most relevant biases with regard to experiential vagueness

- *Experience effect*: the tendency to believe and remember things easier when they are experienced directly with our physical body and senses instead of abstract representations, like graphs and statistics, or text about scientific data (van Vugt et al., 2014).
- *Contrast effect*: having difficulty with perceiving and appreciating gradual changes or differences (instead of contrasting ones), such as gradually decreasing biodiversity and climate change (Plous, 1993).
- *Story bias*: the tendency to accept and remember more easily than simple or basic facts (Alexander and Brown, 2010).

4.1.2. Interventions to mitigate these biases

Key: Make the consequences of possible ecological breakdown tangible

- To increase awareness of environmental threats people should experience by their senses (e.g., vision, sound, proprioception, and smell) how future situations will look and feel, e.g., by gaming, simulation or "experience tanks." In raising and education, positive "nature experiences" can be used in order to promote a pro-environmental perspective of the world.
- People have difficulty with correctly perceiving and judging abstract figures. Quantitative data, tables, and numbers do not really make an impression and are thus easily ignored or forgotten.[2] Make people therefore aware of environmental challenges using concrete examples and narratives that are related to real individuals with whom they can empathize and reinforce messages with vivid and appealing images, frames, and metaphors.
- Use pictures, animations, artist impressions, podcasts, and video's instead of (or to support) written information.
- Focus on the concrete *consequences* of severe threats.
- Humans are evolved to love nature. So, increase the availability and number of opportunities (especially for city dwellers) to appreciate, experience and protect the healing value of the real nature, i.e., the fields, the woods, the waters, and the mountains (Schultz, 2002).
- Sustainability interventions that imply the loss of assets or privileges should proceed slowly, gradual, and in small steps. The more positive and rewarding aspects of transitions can be presented as more contrasting, sudden and discrete events.
- Narratives and stories consisting of coherent events and elements—real or imaginary—are more easily accepted and remembered than plain facts, which may be useful to create or enhance feelings of connectedness and commitment to pro-environmental initiatives.
- From a psycho-social perspective face-to-face communication is probably the richest (and most natural) form of communication and interaction. Use therefore face-to-face communication to promote pro-environmental behavior.

4.2. Long-term effects and future risk

Sustainable choices are often only rewarded in the long-term future, while the costs and sacrifices have to take place in the present. Given two similar rewards, humans show a preference for one that arrives sooner rather than later. So, humans (and other animals) are said to *discount* the value of the later reward and/or delayed feedback (Alexander and Brown, 2010). In addition, this effect increases with the length of the delay. According to van Vugt et al. (2014), our tendency to discount future outcomes may have had substantial benefits in primitive ancestral environments, suggesting it is an evolved psychological trait (Wilson and Daly, 2005). If our ancestors had put too much effort into meeting future needs rather than their immediate needs, they would have been less likely to survive and pass on their genes in the harsh and unpredictable natural environment in which they lived (Boehm, 2012). Human psychology is thus naturally formed to maximize outcomes in the here and now, rather than in the uncertain future (van Vugt et al., 2014). Thus people in modern societies still may weigh immediate outcomes much more heavily than distant ones (Green and Myerson, 2004). This preference for today's desires over tomorrow's needs—and the conflict between people's desire for immediate rather than delayed rewards—may be the cause of the persistence of many environmental problems.

Our brain tends to build general conclusions and predictions on the basis of a (small) number of consistent, previous observations (inductive thinking). A typical and flawed inductive statement is: "Of course humanity will survive. Up to now, we have always survived our major threats and disasters."[3] Even in highly educated and experienced people, inductive reasoning may lead to poor intuitive predictions concerning the risks in the (long-term) future (Taleb, 2007). We tend to focus on risks that we clearly see, but whose consequences are often relatively small, while ignoring the less obvious, but perhaps more serious ones. Next to such poor statistical intuitions, we have a

[2] Although exact "numbers" may sometimes provide information with an aura of objectivity and certainty.

[3] However, most human-like races, such as the Neanderthals, are now extinct and real major threats of humanity are those of a globalized world (which only exists for less than a couple of centuries) such as nuclear or biochemical weapons, global warming, or pandemics.

preference for optimistic perspectives. This leads us to ignore unwelcome information and to underestimate the severity and probability of future (environmental) challenges and hazards (Ornstein and Ehrlich, 1989). This may be especially devasting when considering rare and unpredictable outlier events with high impact ("black swans"). Examples of black swans from the past were the discovery of America (for the native population), World War I, the demise of the Titanic, the rise of the Internet, the personal computer, the dissolution of the Soviet Union, and the 9/11 attacks. Many people ignore possible rare events at the edges of a statistical distribution that may carry the greatest consequences. According to Taleb (2007), black swans (or "unknown-unknowns") rarely factor into our planning, our economics, our politics, our business models, and in our lives. Although these black swans have never happened before and cannot be precisely predicted, they nevertheless need much more attention than we give them. Also global warming may trigger currently unknown climate *tipping points* when change in a part of the climate system becomes self-perpetuating beyond a warming threshold, which will lead to unstoppable earth system impact (IPCC, 2021, 2022).

4.2.1. Most relevant biases related to long-term effects

- *Hyperbolic time discounting:* the tendency to prefer a smaller reward that arrives sooner over a larger reward that arrives later. We therefore have a preference for immediate remuneration or payment compared to later, which makes it hard to withhold the temptation of direct reward (Alexander and Brown, 2010).
- *Normalcy bias*: the tendency to underestimate both the likelihood of a disaster and its possible consequences, and to believe that things will always function the way they normally function (Drabek, 2012). By inductive reasoning, we fail to imagine or recognize possible rare events at the edges of a statistical distribution that often carry the greatest consequences, i.e., black swans (Taleb, 2007).
- *Optimism bias:* (Positive outcome bias, Wishful thinking): the tendency to overestimate the probability of positive (favorable, pleasing) outcomes and to underestimate the probability of negative events (O'Sullivan, 2015).

4.2.2. Interventions to deal with these biases

Key: Bring the rewards of more sustainable choices to the present

- In general, immediate reinforcements are usually better recognized or appreciated and have more effect. Provide thus immediate rewards for green choices, e.g., through subsidy and tax policy, so that it pays more directly to make them.
- Bring long-term benefits in line with short-term ones. For example: investing in solar panels with a quick payback period, subsidizing the purchase of pro-environmental goods, or taxing the use of fossil fuels.
- Make people aware that we live in a world that inherently involves unpredictable and (system-) risks with high impact, e.g., like the corona pandemic. These risks may have severe negative consequences, maybe not yet for themselves in the short term, but much more for their beloved children and grandchildren.
- Present required changes as much as possible in terms of positive challenges, that is in terms of potential benefits rather than negative terms: a more "relaxed and natural way of life" instead of "costs of energy transition." Green policy will deliver a stable and predictable future within the foreseeable future that makes prosperity and well-being possible.

4.3. Complexity and uncertainty

The modern global world we live in is very complex with many intricate causal relationships. Everything is connected to everything, making it very difficult to see what exactly is going on in this dense network and how the interplay of societal, technological, economic, environmental, and (geo)political forces develops. Our wealth and comfort are made possible by many "hidden" enablers, such as child labor in third world sweatshops and animal suffering out of sight in the bio industry. The complexity of interrelated and hidden causes, consequences, or remedies is also very prominent in sustainability issues. Sustainability issues are about by a fine-grained logistic infrastructure and sophisticated technological inventions and their massive application. For example, the energy transition involves complex socio-technical systems that usually involve a high degree of uncertainty about how this will ultimately work out. Our cognitive capacities to pick up and understand all this technical, statistical, and scientific information are inherently limited (e.g., Engler et al., 2018; Korteling et al., 2018). How can we intuitively calculate how much CO_2 emission reduction is required and how much (or little) certain technical or economical interventions contribute to the reduction of greenhouse gases? Many people have also poor capacities for calculation and logic reasoning and a poor intuitive sense for coincidence, randomness, statistics, and probability reasoning (e.g., Monat et al., 1972; Sunstein, 2002; Engler et al., 2018). For instance, concepts like "exponential growth"—i.e., when the instantaneous rate of change of a quantity in time is proportional to the quantity itself—are generally poorly understood.

The inherent constraints of our cognitive system to collect and weight of all this information in a proper and balanced way may result in various biases preventing good judgment and decision making on the basis of the most relevant evidence. Our brain tends to selectively focus on specific pieces of information that 'resonate' with what we already know or expect and/or what associatively most easily pops up in the forming of judgments, ideas, and decisions (Tversky and Kahneman, 1974; Korteling et al., 2018; Toomey, 2023). The fact that other (possible relevant or disconfirming) information may exist beyond what comes up in our mind may be insufficiently recognized or ignored (Kahneman, 2011). This often may lead to a rather simplistic view of the world (e.g., populism). We trust and focus on what is clearly visible or (emotionally) charged, what we (accidentally) know, what we happened to see or hear, what we understand, what intuitively feels true, or what associatively comes to mind (the known-knowns). In contrast, we are rather insensitive to the fact that much information does not easily come to us, is not easily comprehensible, or simply is unknown to us. So we easily may ignore the fact that there usually is a lot that we do not know (The unknowns). This characteristic of neural information processing has been termed: *the Focus principle* (Korteling et al., 2018) or "What You See Is All There

Is" (WYSIATI, Kahneman, 2011). An important consequence of this principle is that we tend to overestimate our knowledge with regard to complex issues about which we lack experience or expertise (Kruger and Dunning, 1999). A situation may also be deemed as too uncertain or complicated and a decision is never made due to the fear that a new approach may be wrong or even worse. An abundance of possible options may aggravate this situation rendering one unable to come to a conclusion. In sustainability challenges, people may thus be very motivated to improve the situation, but still can be hampered by uncertainty and lack of understanding to take action.

4.3.1. Most relevant biases related to complexity and uncertainty

- *Confirmation bias*: the tendency to select, interpret, focus on and remember information in a way that confirms one's preconceptions, views, and expectations (Nickerson, 1998).
- *Neglect of probability*: the tendency to completely disregard probability when making a decision under uncertainty (Sunstein, 2002).
- *Zero-risk bias*: The tendency to overvalue choice options that promise zero risk compared to options with non-zero risk (Viscusi et al., 1987; Baron et al., 1993).
- *Anchoring bias*: Biasing decisions toward previously acquired information. In this way, the early arrival of irrelevant information can seriously affect the outcome (Tversky and Kahneman, 1974; Furnham and Boo, 2011).
- *Availability bias*: the tendency to judge the frequency, importance, or likelihood of an event (or information) by the ease with which relevant instances just happen to pop up in our minds (Tversky and Kahneman, 1973; Tversky and Kahneman, 1974).
- *Focusing illusion*: the tendency to place too much emphasis on one or a limited number of aspects of an event or situation when estimating the utility of a future outcome (Kahneman et al., 2006).
- *Affect heuristic*: basing decisions on what intuitively or emotionally feels right (Kahneman, 2011).
- *Framing bias*: the tendency to base decisions on the way the information is presented (with positive or negative connotations), as opposed to just on the facts themselves (Tversky and Kahneman, 1981; Plous, 1993).
- *Knowledge illusion* (Dunning-Kruger Effect): the tendency in laymen to over-estimate their own competence (Kruger and Dunning, 1999).
- *Surrogation* (means-goal): the tendency to concentrate on an intervening process instead of on the final objective or result, e.g., concentrating on means *vs* goals or on measures *vs* intended objectives (Choi et al., 2012).
- Ambiguity effect: the tendency to avoid options or actions for which the probability of a favorable outcome is unknown (Baron, 1994).

4.3.2. Interventions to deal with these biases

Key: Provide more information and education especially to better understand the environmental consequences of human decisions and actions

- Consistency is more convincing than quantity. We believe that our judgments are accurate, especially when available information is consistent and representative for a known situation. Therefore, conclusions based on a very small body of consistent information are more convincing for most people than much larger bodies of (less consistent) data (i.e., "The law of small numbers").
- Repetition of a pro-environmental message has more impact than just *one* attempt. This exposure effect can be enhanced by using all possible communication channels and media.
- Start with providing information the positive way you want it to taken by the target audience. Later the message may be extended by the less favorable nuances and details.
- Provide better statistical education and training and improve the communication on uncertainty and risk. When it comes to numbers, quantities, and changes therein, focus on total amounts rather than on proportions.
- Make pro-environmental information (e.g., about actions, initiatives, techniques etc….) salient and conspicuous. Focus (in a simple visual way) on the severe *consequences* of global warming and biodiversity loss (desertification, crop failure, and famine, millions of homeless and displaced people, risk of wars) instead of on the complex underlying mechanisms and processes.
- Influence is unlikely to fail due to information that is not provided. Therefore, in setting up an information campaign, it is generally not needed to invest all efforts in providing maximum possible "evidence" that is intended to confirm the deception. Consistency is dominant. In general, clear, recognizable, and simple information will be most easily picked up and accepted.
- Influence and persuasion is not only determined by *what* is, or is not, communicated (i.e., the content) but also by *how* it is communicated or presented (i.e., the frame or form). These latter superficial aspects are more easily, intuitively, and quickly processed than the deeper content of the message. This "framing" can thus be very well exploited for influencing peoples' choices. Each message can be framed in numerous ways. So it may be very effective to analyze how to wrap up a message in the way you want it to be taken.
- Different people value, and pick-up, different information at different levels. Therefore, communicate messages at different levels of understanding, from the direct immediate consequences for the individual (micro) to the overarching long-term consequences for the world of the future and for future generations (macro).
- Present and facilitate as much as possible "total solutions." Which are tailor-made to the target audiences.

4.4. Threat of the status quo

A basic premise of evolution is that all organisms strive for the continuation of their existence. This not only concerns the existence *per se*, but also the maintenance of stable living conditions (that are instrumental to this ultimate goal). For this reason (under normal circumstances and to prevent unexpected risk), we tend to strive at maintaining the present situation and to remain consistent with previous patterns (default effect). So, we easily accept, or prefer, to continue on the path taken and to maintain the status quo (default options) and we are afraid of choosing alternative, options that may turn out suboptimal (Kahneman and Tversky, 1979; Johnson and Goldstein, 2003; Chorus, 2010). Energy transition, as a possible solution of a future problem, is by many people experienced as threatening, not only to our established comfortable way of living, but to our individual and social basic needs as well. A transition to more sustainable practices may thus cause bad feelings of losing security and possessions, sometimes termed "fear of falling."

In line with this, people have an overall tendency to experience the disutility of giving up an object as greater than the utility associated with acquiring it (i.e., Loss aversion). Thaler (1980) recognized this pattern, and articulated it as such: people often demand much more to give up an object than they would be willing to pay to acquire it. This is called the Endowment effect. In contrast to what most authors on cognitive biases suppose, we here speculate that the emotions that we feel when we anticipate possible loss of our assets are not the *cause* of our bias to avoid loss. Instead, they are the *result* of our pervasive bias for self-preservation and for maintenance our (neurobiological) integrity (Korteling et al., 2018). So in brief: we often prefer to hold on to the current situation and to continue on previous (al) choices. As such, we default to the current situation or status quo.

4.4.1. Most relevant biases related to threat of the status quo

- *Status Quo bias*: the tendency to maintain the current state of affairs (Samuelson and Zeckhauser, 1988).
- *Default effect*: the tendency to favor the option that would be obtained if the actor does nothing when given a choice between several options (Johnson and Goldstein, 2003).
- *Sunk cost fallacy* (also known as Irrational escalation or Concorde effect): the tendency to consistently continue a chosen course with negative outcomes rather than alter it. The effort previously invested is the main motive to continue (Arkes and Ayton, 1999).
- *System justification*: the tendency to believe that the current or prevailing systems are fair and just, justifying the existing inaccuracies or inequalities within them (social, political, legal, organizational, and economical) (Jost and Banaji, 1994; Jost et al., 2004).
- *Cognitive dissonance*: the tendency to search for and select consistent information in order to try to reduce discomfort when confronted with facts that contradict own choices, beliefs, and values (Festinger, 1957).
- *Fear of regret*: feeling extra regret for a wrong decision if it deviates from the default (Dobelli, 2011; Kahneman, 2011).
- *Loss aversion*: the tendency to prefer avoiding losses to acquiring equivalent gains. Loss takes an (emotionally) heavier toll than a profit of the same size does (Kahneman and Tversky, 1984).
- *Endowment effect*: the tendency to value or prefer objects that you already own over those that you do not (Thaler, 1980).

4.4.2. Interventions to deal with these biases

Key: Make sustainable options the default or easiest choice and present them as a gains rather than losses

- Make desired pro-environmental choices and behavior the default (the normal standard) or easiest choice. For example, providing only reusable unless specifically request a single-use plastic shopping bag, or designing buildings and cities to make walking and biking more convenient.
- Encourage active participation can be a major tool for triggering cognitive consistency pressures to build more sustainable habits. In general: active participation signals commitment to subjects, increasing their likely identification with the message or goal of the persuasion. Subsequently, they will tend to make choices that are consistent with their previous—in this case pro-environmental—actions.
- Based on cognitive dissonance theory (Festinger, 1957), the expression of self-criticism in peer (discussion) groups is a major influence technique. Making people vocalize promises (or sins) in public drives subjects to remain consistent with their and words.
- We believe that our judgements are accurate, especially when available information is consistent and representative for a known situation. It is therefore always important to provide consistent information.
- People tend to focus on, interpret, and remember information in ways that *confirm* their existing ideas, expectations or preconceptions. Therefore, in order to create an open mind, it is better to start with undeniable, true evidence and take care to not to start with highly disputable information evidence. The more complicated and contradictory aspects can be tackled later.
- The first goal in any effort to change another person's mind must be to ensure that the subject is at least seriously considering the desired alternative. This requires to start with strong and obvious evidence which fits into the target's existing conceptions of the world. In contrast, starting with less dramatic evidence tends to be unsuccessful since the information will be ignored, unnoticed, forgotten, or misperceived.
- Present changes in terms of gains instead of losses and circumvent the loss felt by people when they are asked to invest funds and provide support to acquire the necessary funds for the transition.
- Create a story different from loss: what are we gaining? For example: more rest, less rat race. Do not address people as consumers, but as citizens, changemakers, parents, etc.

4.5. Threat of social status

People are more focused on relative status than absolute status. This is, for example, demonstrated by the fact that people find an increase in wealth relative to their peers more important than their absolute wealth (Diener and Suh, 2000). In an experimental setting, researchers found that when presented with financial options, most people chose to earn less in absolute terms, as long as they relatively earned more than their peers (Frank, 1985). Not unrelated to our status-seeking tendency, humans tend to consume more than they need. In many historical civilizations, we find a penchant toward (excessive) consumption and showing of materials and riches (Bird and Smith, 2005; Godoy et al., 2007). From an evolutionary point of view, such displays of status may be rooted in a social advantage (Penn, 2003; Saad, 2007; Miller, 2009). Ancestors who strived for improvement of their situation and who tried to do better than their peers, probably have passed their genes better than those who had a more comfortable attitude. The wry side effects, however, are that the tendency to seek status through material goods—nowadays more than ever—may contribute substantially to the production of waste and the depletion of nonrenewable resources. Because we seek relative wealth, as opposed to seeking an absolute point of satisfaction, we are not easily satisfied and we tend to persistently strive for ever more status and wealth. Whether it be our smartphone, our sense of fashion, or our household appliances, they all rapidly become outdated as soon as newer or more fashionable versions enter the horizon. As economists say: we compare ourselves continuously with our neighbors; we want to "keep up with the Joneses." Finally, items that are scarce or hard to obtain have typically more perceived quality and status than those that are easy to acquire. So many environmental problems can therefore be the result of a conflict between status-enhancing overconsumption versus having enough for a good life. This 'Hedonic treadmill' is encouraged by commercials offering us a never ending stream of new products that should make us, in one way or the other, happy and thus hungry to buy more.

4.5.1. Most relevant biases related to threat of social status

- Affective forecasting (Hedonic forecasting, Impact bias): the tendency to overestimate the duration and intensity of our future emotions and feelings regarding events, encouraging putting effort into favorable results (greed) and into avoiding threats (Wilson and Gilbert 2005).
- Hedonic adaptation (Hedonic treadmill): the tendency to quickly return to a relatively stable level of happiness despite major positive or negative life events (Brickman and Campbell, 1971).
- Social comparison bias: The tendency, when making decisions, to favor individuals who do not compete with one's own particular strengths (Garcia et al., 2010).
- Scarcity bias: the tendency to attribute greater subjective value to items that are more difficult to acquire or in greater demand (Mittone and Savadori, 2009).

4.5.2. Interventions to deal with these biases[4]

Key: Connect sustainable options and choices with concepts, persons or goods that emanate a high social status

- Frame pro-environmental choices or options (like solar panels, bikes, or electric cars) as status symbols that show good beliefs and an exemplary way of life.
- In contrast, frame counter-environmental options (mopeds, flying, and meat consumption) as unattractive or associate them with low-status.
- Use high-status and admired or popular influencers and celebrities to promote pro-environmental options, e.g., in social media campaigns.
- Educate people to assess their quality of life in absolute terms of health, freedom, and comfort instead of in relative terms towards 'the Jonesses'.
- Present the benefits of environmental as scarce. This can be done, for example, by pointing out others (competitors) who want the same goods or by drawing attention to possible future supply problems.

4.6. Personal versus community interest

Individual self-interest is often in conflict with the interest of the whole group. This is generally conceptualized as a social dilemma. This dilemma is usually referred to as the *Tragedy of the Commons* story (Hardin, 1968). This hypothetical example demonstrates the effects of unregulated grazing (of cattle) on a common piece of land, also known as "the commons." In modern economic terms, 'commons' are any shared or unregulated resources to which all individuals have equal and open access, like the atmosphere, roads, or even the fridge of the office. Searching for direct individual profit, most individuals increase their use or exploitation of these common resources, thereby unintentionally causing it to collapse (Hawkes, 1992; Dietz et al., 2003). According to Hardin (1968) and van Vugt et al. (2014) the human mind is shaped to prioritize their personal interests over collective interests because natural selection favors individuals who can gain a personal benefit at the expense of unrelated others. Of course, there are situations under which the collective benefit will be prioritized over that of the induvial. But the conditions under which the human mind is triggered to prioritize the collective good over its own are generally less prevalent (Hardin, 1968).

According to Dawkins (1976), natural selection is the replication of one's genes, which often comes at the expense of the survival of others' genes. Power is thereby often instrumentally used for self-interest at the cost of others. So, survival of the species is not what primarily matters. However, this prioritizing of self-interest is dependent on the relationship of the individual to the group. In tight-knit communities where the individual knows himself to be dependent on the community, his behavior will be in line with this dependency and more likely be in favor of the in-group's interests. When the individual does not feel this connection to an in-group (community), he is probably more likely to prioritize self-interest. Evidence for this strategy is seen in social dilemma research showing that most individuals tend to make selfish choices when they interact with other people in one-shot encounters (Komorita and Parks, 1994; Fehr and Gächter, 2002; van Lange et al., 2013). The evolutionary tendency to let self-interest prevail at the expense of others has direct implications for environmental practice, which often concerns the overexploitation of limited resources, such as the oceans, natural areas, fish stocks, clean air, etc. Consequently, many sustainability problems result from this conflict between personal and collective interests.

4.6.1. Most relevant biases related to personal versus community interest

- Tragedy of the commons (Selfishness and self-interest): the tendency to prioritize one's own interests over the common good of the community (Hardin, 1968).

[4] Governments will want to consider the ethical preconditions and repercussions of these forms of nudging before engaging in it. Though it is a widely applied strategy in our neoliberal system where commercial advertisements are deemed acceptable to nudge the potential customer into buying their product. However, governments should uphold important ethical guidelines that concur with our values of freedom of choice and democracy. For a more in depth study of this, please read, e.g., van Vugt (2009) and Raihani (2013).

- Perverse incentive effect (Cobra effect): the tendency to respond to incentives in a way that best serves our own interests and that does not align with the beneficial goal or idea behind the incentives, which may lead to "perverse behaviors" (Siebert, 2001).
- Anthropocentrism: the tendency to take the own, human perspective as the starting point for interpreting and reasoning about all sorts of things, such as nature and other living animals (Coley and Tanner, 2012).

4.6.2. Interventions to deal with these biases

Key: Introduce and present sustainable options as the most favorable and profitable

- Because we share our genes with our relatives, kinship may be a good motivator of pro-environmental behavior. Pro-environmental appeals may be more effective if they emphasize the interests of our ingroup, children, siblings, and grand-children.
- Create programs where pro-environmental choices result in direct personal (or business) gain, e.g., by proper incentives or rewards, like tax exemptions.
- Create close-knit, stable, and small communities to foster pro-collective behavior and cooperation.
- In all species, behaviors reinforced by rewards or positive feedback tend to be repeated (Thorndike, 1927, 1933), and the more reinforcement, the greater the effect. Therefore, multiple reinforcements on desired social choices increase the chance that this will remain the case or repeat itself in the future.

4.7. Group pressure

Social psychologists have long known that people tend to adapt to the choices and behavior of others (Asch, 1956). Our tendency of following the majority is adaptive since for most species, the costs of individual learning, through trial and error, are substantial (Simon, 1990; Richerson and Boyd, 2006; Sundie et al., 2006; Sloman and Fernbach, 2018). Also for our ancestors, living in uncertain environments it would probably be better to follow and copy others' behavior than figuring things out for yourself (Kameda et al., 2003; Gorman and Gorman, 2016). This is therefore probably an ancient and natural adaptive tendency which may also help maintaining or strengthening a position within the social group (Korteling et al., 2020a). We thus easily follow leaders or people with high status and authority in groups. We adapt to people around us with which we feel connected, but have an aversion against strangers. We have difficulty being indebted to others and we like and support kind, attractive and agreeable people. This can lead, for example, to after-talk and blind copying of the behavior of others and the faithful following of persuasive and charismatic persons. In line with this, it has been found that green practices are more strongly influenced by the behaviors of our peers than by our personal attitudes toward conservation. For example, when people see that their neighbors are not conserving, they tend to increase their own energy consumption as well, even when they had been conserving energy in the past (Schultz et al., 2007). This herd behavior is unconscious, and is mediated by mirror neurons in the brain (Chartrand and Van Baaren, 2009). However, the unconscious nature of this herd behavior is often not acknowledged or even denied by the conformers themselves (Nolan et al., 2008) and is thus hard to battle. Our modern world is built on the basis of an enormous amount of unsustainable methods, tools, practices, and applications, so there is still a long way to go to achieve a sustainable world. Hence, the human tendency to copy the behavior of others and to regard other people's behaviors as the norm and justification of undesirable behavioral choices can be very detrimental to the achievement of sustainable goals.

4.7.1. Most relevant biases related to group pressure

- Bandwagon effect: the tendency to adopt beliefs and behaviors more easily when they have already been adopted by others (Colman, 2003).
- Conformity bias: the tendency to adjust one's thinking and behavior to that of a group standard.
- Ingroup (−outgroup) bias: the tendency to favor one's own group above that of others (Cialdini and Goldstein, 2004).
- Authority bias: the tendency to attribute greater accuracy to the opinion of authority figures (unrelated to its content) and to be more influenced by their opinions (Milgram, 1963).
- Liking bias: the tendency to help or support another person the more sympathetically they feel, which is largely determined by: kindness, attractiveness, and affinity (Cialdini, 2006).
- Reciprocity: the tendency to respond to a positive action with another positive action ("You help me then I help you") and having difficulty being indebted to the other person (Fehr and Gächter, 2002).
- Social proof: the tendency to mirror or copy the actions and opinions of others, causing (groups of) people to converge too quickly upon a single distinct choice (Cialdini, 2006).

4.7.2. Interventions to deal with these biases

Key: Use social norms and peer pressure to encourage sustainable choices and behaviors

- When a behavioral change is requested, it will probably be better to focus peoples' attention on others who already show the desired pro-environmental behavior instead of educating people about the bad behavior of others.
- People can be seduced to choose for a certain option if they see this in many other people. So, present desirable pro-environmental behaviors as behaviors of the majority of the people (or at least large groups) people. Foster, for example, the desired behavioral choices by advertisements suggesting this behavior is already adopted by groups of people.
- Use people with authority, powerful people, and/or attractive people to promote pro-environmental behavior.
- Create feelings of commitment and indebtment for people who make sacrifices for the community in order to foster sustainability.

5. Discussion and conclusion

5.1. Biases and nudges

In the present paper we have described how ingrained cognitive biases in human thinking may counter the development of green policy practices aimed at fostering a more sustainable and livable world. We have focused our study on how the form, content and communication of information affects our decisions and behavior with regard to sustainability. The influence techniques advocated in this paper are informational and psychological interventions, incentives, and/or nudges that could be effective with regard to biased thinking in the context of the current modern world. In general, biased information processing has served us for almost our entire existence (e.g., Haselton et al., 2005; Korteling et al., 2018). However, these natural and intuitive thinking patterns may be very counterproductive for coping with the global and complex problems the world is facing today. The many possible incentives and nudges presented show that there are many ways to deliberately capitalize on biased thinking in people in order to promote more sustainable behavioral choices.

In previous publications we have explained how biases originate from ingrained neuro-evolutionary characteristics of our evolved brain (e.g., Korteling et al., 2018; Korteling and Toet, 2022). This neuro-evolutionary framework provides more fundamental explanations for human decision making than 'explanations' provided by most social- or psychological studies. These latter (social-) psychological explanations are more 'proximate' in terms of "limitations of information processing capacity" (Simon, 1955; Broadbent, 1958; Kahneman, 1973; Norman and Bobrow, 1975; Morewedge and Kahneman, 2010), two metaphorical "Systems of information processing" (Stanovich and West, 2000; Kahneman, 2003; Evans, 2008; Kahneman, 2011), "emotions" (Kahneman and Tversky, 1984; Damasio, 1994), "prospects" prospects (e.g., Kahneman and Tversky, 1979; Mercer, 2005). "lack of training and experience" (Simon, 1992; Klein, 1997, 1998). Our neuro-evolutionary bias framework explains in terms of structural (neural network) and functional (evolutionary) mechanisms the origin of cognitive biases, why they are so systematic, persistent, and pervasive, and why biased thinking feels so normal, natural, and self-evident. Given the inherent/structural ("neural") and ingrained/functional ("evolutionary") character of biases, it seems unlikely that simple education or training interventions would be effective to improve human decision making beyond the specific educational context (transfer) and/or for a prolonged period of time (retention). On the basis of a systematic review of the literature, this indeed appears the case (Korteling et al., 2021). When it comes to solving the problems of the modern world, it will probably be impossible to defeat or eliminate biases in human thinking. Thus, we should always be aware of the pervasive effects of cognitive biases and be modest about our cognitive abilities to solve complex long-term problems in an easy way.

So, the effects on decision making of bias-mitigation training interventions are likely to be rather ineffective, in the same way that it is difficult to get people to change their eating habits by persuading them that chocolate or meat does not taste good. What is more: denying the ultimate and deep-seated neuro-evolutionary causes of the particularities and limitations of human thinking, may hamper adequate development and usage of effective interventions. For example: if governments strive to decrease the demand for energy-inefficient jacuzzi baths, but they ignore the influence of human evolutionary biases, this might lead to an intervention strategy that fails. Perhaps the government would try to persuade people that buying energy-consuming baths is unwise for the future. But in the context of our tendency to discount the value of future consequences, such a strategy on its own is likely to be rather ineffective. It would probably be more effective to use our knowledge of cognitive biases to our advantage. For example, the fact that we compare ourselves to our peers (Social comparison) might lead to a campaign in which the purchase of sustainable solar panels or a sustainable heat pump or fancy e-bike is related to status and prestige. Likewise, it is better to convey pro-environmental messages in a simple, consistent, repetitive, and tangible way and to focus on the consequences (bad or good) of ones choices, rather than on complex intervening processes. Finally, it is better to communicate information about the many aspects of sustainability at different levels of understanding at the same time, i.e., from the instant aspects for the individual to the global consequences for the world of the future.

5.2. The ethics of nudging

Above we have listed tips and tricks to provoke "sustainable decision making." But as we write this, we realize all the more that this knowledge of how biases work, can be used for all kinds of purposes. In the 'wrong' hands, this knowledge about biases can be used to manipulate or incite the population to destructive. That is not even speculative, history has already shown this over and over again. Fossil industries that succeeded in holding back measures against global warming, doctors recommending brands of cigarettes, smear campaigns that led to witch-hunts, and anti-Semitic propaganda during World War II are just a few examples.

There is a serious ethical issue with using our knowledge of biases to our advantage (e.g., Bovens, 2009; Raihani, 2013). Who decides whether it is ethical to nudge citizens and use our knowledge of evolutionary biases to steer the choices and behavior of people? It sometimes may seem obvious that it is a good thing if you want to prevent incitement to hatred and violence, genocide or destructive such as smoking. But there is also a gray area. In the current pandemic, for example, we see that governments are doing their best to silence dissenting voices "for a good cause." But counter voices also represent the basis of a democratic constitutional state, where counter voices must always be welcomed. Can we afford to go beyond our democratic boundaries, by nudging our citizens, for the sake of the climate? Our thought on this is as follows: Democracy means that everyone is allowed to make their voice heard about the goals that you want to achieve as a society. This report is about how to make your voice heard more effectively. It provides tools that everyone (not just politicians and policy makers) can use, for better or for worse. This applies to any instrument, AI, weapons, robots, ICT, etc.… The evil is not in the instrument, but in the purpose for which it is used. If we democratically choose to achieve certain goals, then it can be deemed defendable that governments use those instruments as effectively as possible to achieve those goals. It leaves people still free to choose their own path and goals.

5.3. A vision-based agenda

Politics can ensure that we as humanity behave more sustainably. In that case, our societal and physical environment will have to be organized differently, for example with far-reaching legislation (eg CO2 tax), a different market-oriented economy and a different transport system.

However, these changes are held back by our ingrained preferences for short-term thinking, maintaining the status quo, personal interest, or herd behavior, which may result in fears like losing jobs or losing freedom. These thinking tendencies and fears are exploited by the lobbies of many powerful (e.g. fossil) parties with vested interests. That is why we have to search for ways to get moving as a society. An important part of this is managing well-being, and thereby discovering that there are ways to live sustainably, and also to be happy. This means that, more than ever, there is a need for knowledge and a substantiated vision about the core values that represent us, as humans, and our world, about who we are, how we want to live and where we want to go. This is not just a vision with long-term goals for human well-being, but also one that builds on our natural needs and that takes into account the hidden and inherent systemic risks of the modern, globalized world. This is essential in determining the course and the agenda for the future of humanity.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

The literature search, analysis, conceptual work, and the writing of the manuscript was done by JEK. GP provided knowledge and information concerning sustainability. JM critically reviewed the manuscript several times. All authors contributed to the article and approved the submitted version.

Acknowledgments

The present paper was a deliverable of the Wise Policy Making project and the Responsible Hypernudging project. These Early Research Projects (ERPs) were funded by the Netherlands Organization for Applied Scientific Research (TNO).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adams, B.D., Sartory, J., and Waldherr, S. (2007). *Military Influence Operations: Review of Relevant Scientific Literature. Report No. CR 2007-146.* Toronto: Defence Research and Development Canada.

Alexander, W. H., and Brown, J. W. (2010). Hyperbolically discounted temporal difference learning. *Neural Comput.* 22, 1511–1527. doi: 10.1162/neco.2010.08-09-1080

Arceneaux, K. (2012). Cognitive biases and the strength of political arguments. *Am. J. Polit. Sci.* 56, 271–285. doi: 10.1111/j.1540-5907.2011.00573.x

Arkes, H. R., and Ayton, P. (1999). The sunk cost and Concorde effects: are humans less rational than lower animals? *Psychol. Bull.* 125, 591–600. doi: 10.1037/0033-2909.125.5.591

Arkes, H. R., and Blumer, C. (1985). The psychology of sunk cost. *Organ. Behav. Hum. Decis. Process.* 35, 124–140. doi: 10.1016/0749-5978(85)90049-4

Asch, S. (1956). Studies of independence and conformity. *Psychol. Monogr.* 70, 1–70. doi: 10.1037/h0093718

Baron, J. (1994). *Thinking and Deciding. 2nd Edn.* Cambridge, UK: Cambridge University Press.

Baron, J. (2009). Cognitive biases in moral judgments that affect political behaviour. *Synthese* 172, 7–35. doi: 10.1007/s11229-009-9478-z

Baron, J., Gowda, R., and Kunreuther, H. (1993). Attitudes toward managing hazardous waste: what should be cleaned up and who should pay for it? *Risk Anal.* 13, 183–192. doi: 10.1111/j.1539-6924.1993.tb01068.x

Bellé, N., Cantarelli, P., and Belardinelli, P. (2018). Prospect theory goes public: experimental evidence on cognitive biases in public policy and management decisions. *Public Admin Rev.* 78, 828–840. doi: 10.1111/puar.12960

Benforado, A. (2015). *Unfair: The New Science of Criminal Injustice.* New York: Broadway Books.

Berger, L. S. (2009). *Averting Global Extinction: Our Irrational Society as Therapy Patient.* Plymouth, UK: Jason Aronson.

Biermann, F., Abbott, K., Andresen, S., Baeckstrand, K., Bernstein, S., Betsill, M. M., et al. (2012). Navigating the 23 Anthropocene: improving earth system governance. *Science* 335, 1306–1307. doi: 10.1126/science.1217255

Bird, R., and Smith, E. A. (2005). Signaling theory, strategic interaction, and symbolic capital. *Curr. Anthropol.* 46, 221–248. doi: 10.1086/427115

Boehm, C. (2012). *Moral Origins*. London: Basic Books.

Bovens, L. (2009). "The ethics of nudge" in *Preference Change: Approaches from Philosophy, Economics and Psychology*. eds. T. Grüne-Yanoff and S. O. Hansson (Dordrecht, Netherlands: Springer Sciences), 207–220.

Brickman, P., and Campbell, D. T. (1971). "Hedonic Relativism and Planning the Good Society," in *Adaptation Level Theory*. ed. M. H. Appley. (New York, NY: Academic Press), 287–301.

Broadbent, B. E. (1958). *Perception and communication*. New York: Pergamon Press.

Chartrand, T., and Van Baaren, R. (2009). Human mimicry. *Adv. Exp. Soc. Psychol.* 41, 219–274. doi: 10.1016/S0065-2601(08)00405-X

Choi, W., Hecht, G., and Tayler, W. B. (2012). Lost in translation: the effects of incentive compensation on strategy Surrogation. *Account. Rev.* 87, 1135–1163. doi: 10.2308/accr-10273

Chorus, C. G. (2010). A new model of random regret minimization. *Eur. J. Transp. Infrastruct. Res.* 10. doi: 10.18757/ejtir.2010.10.2.2881

Cialdini, R.D. (2006). *Influence: The Psychology of Persuasion. Revised Edition*. New York: William Morrow.

Cialdini, R. B. (2009). *Influence: Science and Practice* (*Vol. 4*). Boston: Pearson Education.

Cialdini, R. B., and Goldstein, N. J. (2004). Social influence: compliance and conformity. *Annu. Rev. Psychol.* 55, 591–621. doi: 10.1146/annurev.psych.55.090902.142015

Coley, J. D., and Tanner, K. D. (2012). Common origins of diverse misconceptions: cognitive principles and the development of biology thinking. *CBE Life Sci. Educ.* 11, 209–215. doi: 10.1187/cbe.12-06-0074

Colman, A. M. (2003). *Oxford Dictionary of Psychology*. New York, NY, USA: Oxford University Press.

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad. Med.* 78, 775–780. doi: 10.1097/00001888-200308000-00003

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*, New York, NY, USA. G. P. Putnam's Sons.

Dawkins, R. (1976). *The Selfish Gene*. Oxford: Oxford University Press.

Diener, E., and Suh, E. M. (2000). *Culture and Subjective Wellbeing*. Boston: MIT Press.

Dietz, T., Ostrom, E., and Stern, P. C. (2003). The struggle to govern the commons. *Science* 302, 1907–1912. doi: 10.1126/science.1091015

Dobelli, R. (2011). *Die Kunst des Klaren Denkens: 52 Denkfehler die sie Besser Anderen Uberlassen*. Munchen: Karl Hanser Verlag.

Drabek, T. E. (2012). *Human System Responses to Disaster: An Inventory of Sociological Findings*, New York, NY, USA, Springer Verlag.

Eigenauer, J. D. (2018). The problem with the problrm of human irrationality. *Int. J. Educ. Reform* 27, 341–358. doi: 10.1177/105678791802700402

Engler, J. O., Abson, D. J., and von Wehrden, H. (2018). Navigating cognition biases in the search of sustainability. *Ambio* 48, 605–618. doi: 10.1007/s13280-018-1100-5

Evans, J. S. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* 59, 255–278. doi: 10.1146/annurev.psych.59.103006.093629

Fehr, E., and Gächter, S. (2002). Altruistic punishment in humans. *Nature* 415, 137–140. doi: 10.1038/415137a

Festinger, L. (1957). *A Theory of Cognitive Dissonance*, Stanford, CA, USA, Stanford University Press.

Fiske, S. (2004). *Social Beings: Core Motives in Social Psychology*. New York: Wiley and Sons.

Flyvbjerg, B. (2009). Survival of the unfittest: why the worst infrastructure gest built – and what can we do about is. *Oxf. Rev. Econ. Policy* 25, 344–367. doi: 10.1093/oxrep/grp024

Frank, R. (1985). *Choosing the Right Pond: Human and the Quest for Status*. New York: Oxford University Press.

Furnham, A., and Boo, H. C. (2011). A literature review of the anchoring effect. *J. Socio-Econ.* 40, 35–42. doi: 10.1016/j.socec.2010.10.008

Garcia, S. M., Song, H., and Tesser, A. (2010). Tainted recommendations: the social comparison bias. *Organ. Behav. Hum. Decis. Process.* 113, 97–101. doi: 10.1016/j.obhdp.2010.06.002

Gardner, G., and Stern, P. C. (2002). *Environmental Problems and Human Behaviour*. London: Pearson.

Garland, H., and Newport, S. (1991). Effects of absolute and relative sunk costs on the decision to persist with a course of action. *Organ. Behav. Hum. Decis. Process.* 48, 55–69. doi: 10.1016/0749-5978(91)90005-E

Gigerenzer, G., Todd, P. M. and ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford, GB: Oxford University Press.

Gifford, R. (2011). The dragons of inaction: psychological barriers that limit climate change mitigation and adaptation. *Am. Psychol.* 66, 290–302. doi: 10.1037/a0023566

Gigerenzer, G., and Gaissmaier, W. (2011). Heuristic decision making. *Annu. Rev. Psychol.* 62, 451–482. doi: 10.1146/annurev-psych-120709-145346

Godoy, R., Reyes-García, V., Leonard, W. R., Huanca, T., McDade, T., Vadez, V., et al. (2007). Signaling by consumption in a native Amazonian society. *Evol. Hum. Behav.* 28, 124–134. doi: 10.1016/j.evolhumbehav.2006.08.005

Gorman, S. E., and Gorman, J. M. (2016). *Denying to the Grave: Why We Ignore the Facts that Will Save Us*. London, UK: Oxford University Press.

Grabe, M. E., and Bucy, E. P. (2009). *Image Bite Politics: News and the Visual Framing of Elections: News and the Visual Framing of Elections*. Oxford University Press, USA.

Green, L., and Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychol. Bull.* 130, 769–792. doi: 10.1037/0033-2909.130.5.769

Groopman, J. (2007). *How Doctors Think*. New York: Houghton Mifflin.

Hansen, W. G. (2013). *Influence: Theory and Practice*. Montery, California: Naval Postgraduate School.

Harari, Y.N. (2017). *Homo Deus: A Brief History of Tomorrow*. London: Jonathan Cape.

Hardin, G. (1968). Tragedy of the commons. *Science* 162, 1243–1248. doi: 10.1126/science.162.3859.1243

Hardin, G. (1995). *Living with limits: Ecology, economics, and population taboos*. Oxford, UK: Oxford University Press.

Haselton, M. G., and Nettle, D. (2006). The paranoid optimist: an integrative evolutionary model of cognitive biases. *Personal. Soc. Psychol. Rev.* 10, 47–66. doi: 10.1207/s15327957pspr1001_3

Haselton, M. G., Nettle, D., and Andrews, P. W. (2005). "The evolution of cognitive bias" in *The Handbook of Evolutionary Psychology*. ed. D. M. Buss (Hoboken, NJ, USA: John Wiley and Sons Inc).

Hastie, R., and Dawes, R. M. (2001). *Rational Choice in an Uncertain World: The Psychology of Judgement and Decision Making*. Thousand Oaks: Sage.

Hawkes, K. (1992). "Sharing and collective action" in *Evolutionary Ecology and Human Behaviour*. eds. E. Smith and B. Winterhalder (New York: Aldine de Gruyter), 269–300.

Heuer, R. J. (2013). "Cognitive factors in deception and counter deception" in *The Art and Science of Military Deception*. eds. H. Rothstein and B. Whaley (Boston/London: Artech House), 105–133.

IPCC (2013). Summary for Policymakers. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.

IPCC (2014). "Summary for policymakers," in *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. eds. O. Edenhofer, R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, et al. (Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press).

IPCC (2021). Summary for Policymakers. In: The Physical Science Basis. Available at: https://IPCC_AR6_WGI_SPM_final.pdf (Accessed February 02, 2022).

IPCC (2022). Critical Findings of the Sixth Assessment Report (AR6) of Working Group I of the Intergovernmental Panel on Climate Change (IPCC) for Global Climate Change Policymaking: A Summary for Policymakers (SPM) Analysis.

Johnson, E. J., and Goldstein, D. (2003). Do defaults save lives? *Science* 302, 1338–1339. doi: 10.1126/science.1091721

Jost, J. T., and Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *Br. J. Soc. Psychol.* 33, 1–27. doi: 10.1111/j.2044-8309.1994.tb01008.x

Jost, J. T., Banaji, M. R., and Nosek, B. A. (2004). A decade of system justification theory: accumulated evidence of conscious and unconscious bolstering of the status quo. *Polit. Psychol.* 25, 881–919. doi: 10.1111/j.1467-9221.2004.00402.x

Jowett, G, and O'Donnell, V. (1992). *Propaganda and Persuasion*, *2nd*, Newbury Park, CA: Sage Publications, 122–154.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, New Jersey, Prentice-Hall Inc.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *Am. Psychol.* 58, 697–720. doi: 10.1037/0003-066X.58.9.697

Kahneman, D. (2011). *Thinking Fast and Slow*. New York, USA: Farrar, Straus and Giroux.

Kahneman, D., and Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *Am. Psychol.* 64, 515–526. doi: 10.1037/a0016755

Kahneman, D., Krueger, A. B., Schkade, D., Schwarz, N., and Stone, A. A. (2006). You be happier if you were richer: a focusing illusion. *Science* 312, 1908–1910. doi: 10.1126/science.1129688

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge, UK, Cambridge University Press.

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47:263. doi: 10.2307/1914185

Kahneman, D., and Tversky, A. (1984). Choices, values, and frames. *Am. Psychol.* 39, 341–350. doi: 10.1037/0003-066X.39.4.341

Kameda, T., Takezawa, M., and Hastie, R. (2003). The logic of social sharing: an evolutionary game analysis of adaptive norm development. *Personal. Soc. Psychol. Rev.* 7, 2–19. doi: 10.1207/S15327957PSPR0701_1

Kates, R. W., and Parris, T. M. (2003). Long-term trends and a sustainability transition. *Proc. Natl. Acad. Sci.* 100, 8062–8067. doi: 10.1073/pnas.1231331100

Klein, G. (1997). "The recognition-primed decision (RPD) model: looking back, looking forward" in *Naturalistic Decision Making*. eds. C. E. Zsambok and G. Klein (New York, USA: Psychology Press).

Klein, G. (1998). *Sources of Power: How People Make Decisions*, Cambridge, MA, USA, MIT Press.

Komorita, S., and Parks, C. D. (1994). *Social Dilemmas*. Madison, WI: Brown and Benchmark.

Korteling, J. E., Brouwer, A. M., and Toet, A. (2018). A neural network framework for cognitive bias. *Front. Psychol.* 9:1561. doi: 10.3389/fpsyg.2018.01561

Korteling, J.E., and Duistermaat, M. (2018). *Psychological Deception. Report TNO R11532*. Soesterberg: TNO Defence, Safety and Security.

Korteling, J. E., Gerritsma, J., and Toet, A. (2021). Retention and transfer of cognitive bias mitigation interventions: a systematic literature study. *Front. Psychol.* 12:629354. doi: 10.3389/fpsyg.2021.629354

Korteling, J.E., Sassen-van Meer, J., and Toet, A. (2020a). *Neuro-Evolutionary Framework for Cognitive Biases. Rapport TNO 2020 R10611*. Soesterberg: TNO Defence, Safety and Security

Korteling, J.E., Sassen-van Meer, J., and Toet, A. (2020b). *Neuro-Evolutionary Bias Framework. Report TNO 2020 R11451*. Soesterberg: TNO Defence, Safety and Security.

Korteling, J. E., and Toet, A. (2022). "Cognitive biases" in *Encyclopedia of Behavioural Neuroscience*. ed. S. Della Sala. *2nd Edn.* (Elsevier), 610–619.

Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol.* 77, 1121–1134. doi: 10.1037/0022-3514.77.6.1121

Levy, J. S. (2003). Applications of Prospect theory to political science. *Synthese* 135, 215–241. doi: 10.1023/A:1023413007698

Marshall, G. (2015). *Don't Even Think About It: Why Our Brains are Wired to Ignore Climate Change*. Bloomsbury Publishing, USA.

McDermott, R. (2004). Prospect theory in political science: gains and losses from the first decade. *Polit. Psychol.* 25, 289–312. doi: 10.1111/j.1467-9221.2004.00372.x

Meadows, D. (1997). Places to intervene in a system. *Whole Earth* 91, 78–84.

Meadows, D., Randers, J., and Behrens, W.W. (1972). *The Limits to Growth*. New York: Universe Books.

Mercer, J. (2005). Prospect theory and political science. *Annu. Rev. Polit. Sci.* 8, 1–21. doi: 10.1146/annurev.polisci.8.082103.104911

Milgram, S. (1963). Behavioral study of obedience. *J. Abnorm. Soc. Psychol.* 67, 371–378. doi: 10.1037/h0040525

Millennium Ecosystem Assessment. (2005). *Ecosystems and Human Well-Being: Synthesis*. Washington, DC: Island Press.

Miller, G. F. (2009). *Spent: Sex, Evolution, and Consumer Behaviour*. New York: Viking.

Mittone, L., and Savadori, L. (2009). The scarcity bias. *Appl. Psychol.* 58, 453–468. doi: 10.1111/j.1464-0597.2009.00401.x

Modic, D., and Lea, S. E. G. (2013). Scam compliance and the psychology of persuasion. *Soc. Sci. Res. Network* 34. doi: 10.2139/ssrn.2364464

Monat, A., Averill, J., and Lazarus, R. S. (1972). Anticipatory stress and coping reactions under various conditions of uncertainty. *J. Pers. Soc. Psychol.* 24, 237–253. doi: 10.1037/h0033297

Morewedge, C. K., and Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends Cogn. Sci.* 14, 435–440. doi: 10.1016/j.tics.2010.07.004

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2, 175–220. doi: 10.1037/1089-2680.2.2.175

Nolan, J. P., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2008). Normative social influence is underdetected. *Pers. Soc. Psychol. Bull.* 34, 913–923.

Norman, D. A., and Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cogn. Psychol.* 7, 44–64. doi: 10.1016/0010-0285(75)90004-3

O'Sullivan, O. P. (2015). The neural basis of always looking on the bright side. *Dialogues Philos. Ment. Neuro Sci.* 8, 11–15.

OECD (2012). *OECD Environmental Outlook to 2050*. Paris, France.

Ornstein, R., and Ehrlich, P. (1989). *New world, New Mind: Moving Toward Conscious Evolution*. New York: Touchstone Books.

Parker, A. (2003). *In the Blink of an Eye: How Vision Sparked the Big Bang of Evolution*. New York: Basic Books.

Penn, D. J. (2003). The evolutionary roots of our environmental problems: toward a Darwinian ecology. *Q. Rev. Biol.* 78, 275–301. doi: 10.1086/377051

Pinker, S. (2018). *Enlightment Now: The Case for Reason, Science, Humanism, and Progress*. London, GB: Viking.

Plous, S. (1993). *The Psychology of Judgment and Decision Making*. New York: McGraw-Hill.

Powel, T. E. (2017). *Multimodal News Framing Effects. Dissertation*. Amsterdam, NL: University of Amsterdam.

Pronin, E., Lin, D. Y., and Ross, L. (2002). The bias blind spot: perceptions of bias in self versus others. *Personal. Soc. Psychol. Bull.* 28, 369–381. doi: 10.1177/0146167202286008

Raihani, N. J. (2013). Nudge politics: efficacy and ethics. *Front. Psychol.* 4. doi: 10.3389/fpsyg.2013.00972

Richerson, P. J., and Boyd, R. (2006). *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago: Chicago University Press.

Saad, G. (2007). *The Evolutionary Bases of Consumption*. New York: Lawrence Erlbaum Associates.

Samuelson, W., and Zeckhauser, R. (1988). Status quo bias in decision making. *J. Risk Uncertain.* 1, 7–59. doi: 10.1007/BF00055564

Schultz, P. W. (2002). "Inclusion with nature: understanding the psychology of human-nature interactions" in *The psychology of sustainable development*. eds. P. Schmuck and P. W. Schultz (New York: Kluwer), 61–78. doi: 10.1007/978-1-4615-0995-0_4

Schultz, P. W., Nolan, J. P., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychol. Sci.* 18, 429–434. doi: 10.1111/j.1467-9280.2007.01917.x

Shafir, E., and LeBoeuf, R. A. (2002). Rationality. *Annu. Rev. Psychol.* 53, 491–517. doi: 10.1146/annurev.psych.53.100901.135213

Shiller, R. (2015). *Irrational Exuberance*. Princeton: Princeton University Press.

Siebert, H. (2001). *Der Kobra-Effekt. Wie Man Irrwege der Wirtschaftspolitik Vermeidet*. Munich: Deutsche Verlags-Anstalt

Simon, H. A. (1955). A behavioural model of rational choice. *Q. J. Econ.* 69, 99–118. doi: 10.2307/1884852

Simon, H. A. (1990). A mechanism for social selection and successful altruism. *Science* 250, 1665–1668. doi: 10.1126/science.2270480

Simon, H. A. (1992). What is an "explanation" of behaviour? *Psychol. Sci.* 3, 150–161. doi: 10.1111/j.1467-9280.1992.tb00017.x

Sloman, S. A., and Fernbach, P. (2018). *The Knowledge Illusion: Why We Never Think Alone*. Illinois: Penguin.

Slovic, P. (1987). Perception of risk. *Science* 236, 280–285.

Stanovich, K. E., and West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *Behav. Brain Sci.* 23, 645–665.

Steffen, W., Richardson, K., Rockstrom, J., Cornell, S. E., Fetzer, I., Bennett, E. M., et al. (2015). Planetary boundaries: guiding human development on a changing planet. *Science* 347:1259855. doi: 10.1126/science.1259855

Steg, L., and Vlek, C. (2009). Encouraging prosocial behaviour: an integrative review and research agenda. *J. Environ. Psychol.* 29, 309–317. doi: 10.1016/j.jenvp.2008.10.004

Stoknes, P. E. (2015). *What We Think About When We Try Not to Think About Global Warming: Toward a New Psychology of Climate Action*. Vermont: Chelsea Green Publishing.

Sundie, J. M., Cialdini, R. B., Griskevicius, V., and Kenrick, D. T. (2006). "Evolutionary social influence" in *Evolution and Social Psychology*. ed. M. Schaller (New York: Psychology Press), 287–316.

Sunstein, C. R. (2002). Probability neglect: emotions, worst cases, and law. *Yale Law J.* 112, 61–107. doi: 10.2307/1562234

Swim, J. K., Stern, P. C., Doherty, T., Clayton, S., Reser, J., Weber, E., et al. (2011). Psychology's contributions to understanding and addressing global climate change. *Am. Psychol.* 66, 241–250. doi: 10.1037/a0023220

Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*. New York: The random House.

Thaler, R. (1980). Toward a positive theory of consumer choice. *J. Econ. Behav. Organ.* 1, 39–60. doi: 10.1016/0167-2681(80)90051-7

Thaler, R. H., and Sunstein, C. R. (2008): *Nudge–Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, New Haven, CT.

Thorndike, E. L. (1927). The law of effect. *Am. J. Psychol.* 39, 212–222. doi: 10.2307/1415413

Thorndike, E. L. (1933). A proof of the law of effect. *Science* 77, 173–175. doi: 10.1126/science.77.1989.173.b

Toomey, A. H. (2023). Why facts don't change minds: insights from cognitive science for the improved communication of conservation research. *Biol. Conserv.* 278:109886. doi: 10.1016/j.biocon.2022.109886

Tversky, A., and Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cogn. Psychol.* 5, 207–232. doi: 10.1016/0010-0285(73)90033-9

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131. doi: 10.1126/science.185.4157.1124

Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science* 211, 453–458. doi: 10.1126/science.7455683

Uzzell, D. L. (2000). The psycho-spatial dimension of global environmental problems. *J. Environ. Psychol.* 20, 307–318. doi: 10.1006/jevp.2000.0175

van Lange, P. A. M., Balliet, D. P., Parks, C. D., and Vugt, M.van (2013). *Social Dilemmas: The Psychology of Human Cooperation*. Oxford: Oxford University Press.

van Vugt, M. (2009). Averting the tragedy of the commons: using social psychological science to protect the environment. *Curr. Dir. Psychol. Sci.* 18, 169–173. doi: 10.1111/j.1467-8721.2009.01630.x

van Vugt, M., Griskevicius, V., and Schultz, P. W. (2014). Natrurally green: harnessing stone age psychological biases to foster environmental behaviour. *Soc. Issues Policy Rev.* 8, 1–32. doi: 10.1111/sipr.12000

Vis, B. (2011). Prospect theory and political decision making. *Polit. Stud. Rev.* 9, 334–343. doi: 10.1111/j.1478-9302.2011.00238.x

Viscusi, W. K., Magat, W. A., and Huber, J. (1987). An investigation of the rationality of consumer valuation of multiple health risks. *RAND J. Econ.* 18, 465–479. doi: 10.2307/2555636

Wilson, E. O. (2006). *The Creation: An Appeal to Save Life on Earth*. New York: Norton.

Wilson, M., and Daly, M. (2005). Carpe diem: adaptation and devaluing the future. *Q. Rev. Biol.* 80, 55–60. doi: 10.1086/431025

Wilson, T. D., and Gilbert, D. T. (2005). Affective forecasting: knowing what to want. *Curr. Dir. Psychol. Sci.* 14, 131–134. doi: 10.1111/j.0963-7214.2005.00355.x

Yasynska, K. (2019). Can I Trust My Brain? Brainy Sundays. Available at: https://scanberlin.com/2019/09/29/can-i-trust-my-brain (Accessed November 08, 2021).

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 06 March 2023
DOI 10.3389/fpsyg.2023.1127381

Check for updates

OPEN ACCESS

EDITED BY
Dario Monzani,
University of Palermo, Italy

REVIEWED BY
Giulia Ferraris,
University Medical Center Groningen,
Netherlands
Thomas Suslow,
Leipzig University, Germany

*CORRESPONDENCE
Klara Blauth
✉ klara.blauth@uni-bielefeld.de

SPECIALTY SECTION
This article was submitted to
Psychopathology,
a section of the journal
Frontiers in Psychology

RECEIVED 19 December 2022
ACCEPTED 17 February 2023
PUBLISHED 06 March 2023

CITATION
Blauth K and Iffland B (2023) Attentional bias for sad facial expressions in adults with a history of peer victimization. *Front. Psychol.* 14:1127381.
doi: 10.3389/fpsyg.2023.1127381

Attentional bias for sad facial expressions in adults with a history of peer victimization

Klara Blauth* and Benjamin Iffland

Department of Psychology, Bielefeld University, Bielefeld, Germany

Introduction: Previous research has indicated altered attentional processing in individuals with experiences of maltreatment or victimization in childhood and adolescence. The present study examined the impact of child and adolescent experiences of relational peer victimization on attentional processes in adulthood when confronted with emotional facial expressions.

Methods: As part of an online study, a community sample of adults completed a facial dot-probe task. In the present task, pictures of facial expressions displaying four different emotions (anger, disgust, happiness, and sadness) were used.

Results: The results of the hierarchical regression analyses showed that retrospective reports of peer victimization made a significant contribution to the prediction of facilitated orienting processes for sad facial expressions. Experiences of emotional child maltreatment, on the other hand, made a significant contribution to the prediction of attentional biases for angry facial expressions.

Discussion: Our results emphasize the relevance of experiences of emotional and relational maltreatment in childhood and in adolescence for the processing of social stimuli in adulthood. The findings regarding emotional child maltreatment are more indicative of attentional biases in the context of threat detection, whereas the altered attentional processes in peer victimization are more indicative of mood-congruent biases. These altered processes may be active in social situations and may therefore influence future social situations, behavior, feelings, and thus mental health.

KEYWORDS

peer victimization, child maltreatment, attentional bias, emotional facial expressions, dot-probe task

1. Introduction

There are a variety of studies demonstrating the negative impact of maltreatment experiences in a peer context in childhood and adolescence on psychosocial adjustment and particularly mental health (for a review see McDougall and Vaillancourt, 2015). These experiences, also called peer victimization experiences, include different kinds of maltreatment experiences that occur in interactions with peers, e.g., overt forms like physical or verbal violence, or relational maltreatment experiences associated with rejection or exclusion from a social group (De Los Reyes and Prinstein, 2004; Siegel et al., 2009; Sansen et al., 2015). Thus, peer victimization can be distinguished from child maltreatment, where the violence is perpetrated by adults or caregivers, including forms of emotional, physical, or sexual maltreatment (World Health Organization, 1999). Similar to the negative effects of child maltreatment (for a review see Carr et al., 2020), peer victimization is associated with problems in several areas, such as academic achievement or social adjustment (e.g., Schwartz et al., 2005; Juvonen et al., 2011; Takizawa et al., 2014). In addition to their influence on social and economic outcome variables, experiences of peer victimization seem to significantly increase the risk of experiencing various mental disorders like depression, anxiety disorders, PTSD, or substance abuse in childhood and adulthood (e.g., Stapinski et al., 2014; Hébert et al., 2016; Earnshaw et al., 2017). Moreover, a longitudinal study showed an association between frequent victimization and suicide attempts and suicide for girls later in life (Klomek et al., 2009). Thus, the experience of peer victimization in childhood and adolescence has long-term consequences that have a particular impact on mental health even decades after the exposure. Following on from this, studies indicate that experiences of adverse experiences in childhood and adolescence are related to an altered stress response, or structural and functional brain changes which in turn may have an impact on mental health in adulthood (Brendgen et al., 2017; Aults et al., 2019; Quinlan et al., 2020). In addition to physiological factors, altered attentional processes, or attentional biases, have been discussed in the context of traumatic childhood experiences and psychopathology in later life (Fani et al., 2011; Günther et al., 2015; Kelly et al., 2015; Iffland et al., 2019).

Attentional biases refer to the altered attentional focus on stimuli, are influenced by the valence or relevance of a stimulus, and are shaped by individual factors, such as emotional states or psychopathological symptoms (Koster et al., 2004, 2005; Bar-Haim et al., 2007; Cisler and Koster, 2010; Hankin et al., 2010; Peckham et al., 2010). Since they can influence perception and interpretation, and thus cognition and behavior, attentional biases are therefore considered in theories of the development and maintenance of mental disorders (for a review see Cisler and Koster, 2010). Investigating attentional biases in more detail, three different forms of attentional bias can be distinguished (i.e., facilitated attention, difficulties of disengagement, and attentional avoidance; Koster et al., 2004; Cisler and Koster, 2010). Facilitated attention is reflected in the way that emotional stimuli attract attention and thus attention is shifted to these stimuli more quickly. Difficulties in disengagement refer to the extent to which a stimulus attracts attention. This is accompanied by the difficulty in shifting attention from one stimulus to another stimulus. Attentional avoidance is manifested by the avoidance of shifting attention to potentially threatening stimuli and instead directing it to stimuli that are not threatening (Koster et al., 2004; Cisler and Koster, 2010).

Attentional biases have robustly been shown in individuals with depression or various forms of anxiety disorders (for a detailed overview see Bar-Haim et al., 2007; Cisler and Koster, 2010; Peckham et al., 2010). Furthermore, attentional biases have been repeatedly reported in victims of abuse and neglect with and without psychopathology (e.g., Pine et al., 2005; Fani et al., 2011; Romens and Pollak, 2012; Günther et al., 2015; Kelly et al., 2015). Previous research has suggested that experiences of maltreatment appear to influence attentional processes mainly in response to threatening stimuli (Gibb et al., 2009; Iffland and Neuner, 2020). In these studies, abused children had a higher tendency to attend to threatening stimuli, had problems shifting their attention away from cues of anger, and were faster in recognizing anger with less information (for a detailed overview see Jaffee, 2017). For example, using an emotional Stroop task and a dot-probe task, Iffland and Neuner (2022) found that emotional abuse was a significant predictor of attentional biases toward negatively associated neutral faces. In the dot-probe task, this was reflected in facilitated attention to these facial stimuli (Iffland and Neuner, 2022). Other studies found attention avoidance of threatening stimuli associated with child maltreatment (Pine et al., 2005; Kelly et al., 2015). With respect to studies of attentional biases in depression, there is evidence that altered attention allocation in the context of maltreatment does not only refer to a potential threat (Romens and Pollak, 2012; Günther et al., 2015). Günther et al. (2015) examined the connection between child maltreatment experiences and attentional processes using a dot-probe task in adults with a diagnosis of major depression. The authors found that experiences of child maltreatment were associated with altered attention allocation to sad facial expressions. This result was independent of symptom severity. Sustained attention toward sad faces was shown to be a stronger mood-congruent bias in depressed individuals with a history of child maltreatment (Günther et al., 2015). However, there have also been studies that found no evidence of attentional biases to emotional stimuli in general or to negative stimuli when analyzing reaction times within the dot-probe task in maltreated individuals (Fani et al., 2011; Hoepfel et al., 2022). Thus, although the results are not entirely conclusive, there is substantial evidence of attentional bias in the context of child maltreatment experiences.

Similar to child maltreatment perpetrated by adults or caregivers, relational peer victimization was associated with altered attention processes. Particularly, peer abused children showed less interference when confronted with victim-related words in an emotional Stroop task (Rosen et al., 2007). In another study examining adult psychiatric patients and healthy controls, Iffland et al. (2019) reported attentional biases in individuals who experienced relational peer victimization. Independent of the presence of mental illness, peer victimized individuals showed attentional avoidance in response to emotional words. Notably, avoidance was found not only in response to threatening stimuli but to emotional stimuli in general. In addition, Iffland and Neuner (2022) identified attentional biases for neutral faces previously conditioned with negative stimuli in individuals with experiences of relational peer victimization. Specifically, retrospective reports of relational peer victimization made an incremental contribution to the prediction of attentional biases beyond child maltreatment. Yet, they found no evidence of attentional avoidance, but rather an attentional bias toward threatening stimuli (Iffland and Neuner, 2022). Hence, regarding the impact of experiences of peer victimization on attention allocation, findings are not entirely conclusive as they have differed concerning the type of attentional biases and the valence of the stimuli for which biases occur.

To extend existing knowledge regarding the association between relational peer victimization and attentional processes we conducted an online facial dot-probe task (MacLeod et al., 1986) with emotional faces. The present study sought to investigate the influence of relational peer victimization experiences in childhood and adolescence on attentional processes and biases when using stimuli that are relevant in social interactions (emotional facial expressions). Attentional processes were examined in relation to positive and negative stimuli, with negative stimuli distinguished between negative, non-threatening stimuli (sad faces) and potentially threatening stimuli associated with victimization experiences (angry and disgusted faces). Social threat and exclusion are communicated not only through angry facial expressions but also through disgusted facial expressions, as there is an interpersonal aspect of disgust that is elicited by undesirable individuals to protect the social order (Rozin et al., 2008; Tybur et al., 2013). This more nuanced stimulus selection including potentially threatening and non-threatening negative emotions was used to provide a more accurate analysis of attentional processes that extends the findings from previous dot-probe studies in peer-victimized adults (Iffland et al., 2019; Iffland and Neuner, 2022). Drawing from the findings of previous research (Iffland and Neuner, 2022), relational peer victimization was expected to make a significant contribution to the prediction of attentional biases beyond the influence of child maltreatment experiences. We assumed that this would be particularly the case for emotions that are relevant in the context of peer victimization (i.e., anger, disgust). Based on previous results (Iffland and Neuner, 2022) we expected attentional biases to be evident in heightened attention to potentially threatening stimuli.

2. Materials and methods

2.1. Participants

Participants were recruited through the distribution of the participation link or QR code with access to the study *via* social media and flyers. In addition, patients receiving care at two outpatient clinics [*Bielefelder Institut für Psychologische Psychotherapieausbildung* (BIPP) and *Psychotherapeutische Ambulanz der Universität Bielefeld* (PAdUB)] were recruited for participation. The flyer contained information about the aims and methods of the study as well as a notice about the anonymity of the participation. At the beginning of the experiment information on general sociodemographic variables such as age, gender, educational level, and family status were requested. Furthermore, the instrument assessed the presence of mental illnesses, the use of medication, and other physical and psychological health questions in addition to the questionnaires used in this study. The sociodemographic and psychopathological characteristics of the 90 participants who were included in the analyses can be found in Table 1.

2.2. Procedure

Questionnaires were administered using the Qualtrics survey platform. For the experiment, the web version of the program Inquisit 6 (Millisecond software) was used. At the beginning of the study, participants were informed that participation was voluntary and that it was possible to quit the study at any time without penalty. They were also informed that participation would not be remunerated and that confidential information about mental health symptoms and stressful life experiences would be collected. Participation was only possible after participants had given their consent to participate by clicking on a box. After answering the questionnaires, participants were redirected to the Inquisit homepage, from where the Inquisit application could be downloaded. The experiment was designed in such a way that it could be carried out on both computers and mobile devices. The procedure was approved by the Ethics Committee of Bielefeld University.

TABLE 1 Subject soziodemographic and psychopathological characteristics ($N = 90$).

Characteristics	
Gender, % female (*n*)	80.0 (72)
Age, *M* (*SD*)	28.8 (11.1)
Family status, % single (*n*)	38.9 (35)
Educational status (high school or higher), % (*n*)	91.1 (82)
Mental disorder in the past/currently[a], % (*n*)	45.6 (41)/30.0 (27)
Symptoms of depression[b], *M* (*SD*)	14.9 (11.0)
Psychopathology[c], *M* (*SD*)	19.9 (17.2)
Trait anxiety[d], *M* (*SD*)	44.9 (13.2)
Child maltreatment experiences[e], *M* (*SD*)	40.1 (15.0)
Emotional abuse, *M* (*SD*)	10.1 (4.6)
Emotional neglect, *M* (*SD*)	10.1 (4.6)
Physical abuse, *M* (*SD*)	6.3 (2.5)
Physical neglect, *M* (*SD*)	7.1 (2.7)
Sexual abuse, *M* (*SD*)	6.6 (3.8)
Minimization/denial, *M* (*SD*)	0.4 (0.8)
Peer victimization experiences[f], *M* (*SD*)	10.3 (7.5)

[a]based on self-report; [b]Beck Depression Inventory; [c]Symptom Checklist-27; [d]State Trait Anxiety Inventory (Trait); [e]Childhood trauma questionnaire; [f]Fragebogen zu belastenden Sozialerfahrungen.

2.3. Symptoms of psychopathology

To assess general symptoms of psychopathology the Symptom Check List-27 (SCL-27; Hardt and Gerbershagen, 2001) was used. This 27-item questionnaire captures different areas of psychological symptoms (six subscales including depressive, dysthymic, vegetative, agoraphobic, sociophobe symptoms, and symptoms of mistrust). For this sample, there was an excellent internal consistency (Cronbach's $\alpha = 0.94$).

The German version of the Beck Depression Inventory (BDI II; Hautzinger et al., 2006; Kühner et al., 2007) was used to assess current depressive symptomatology over the last 2 weeks. This questionnaire uses 21 items to assess the severity of depressive symptoms on a scale from zero (absent) to four (severely present). The sum value of the items allows conclusions to be drawn about the severity of depressive symptoms (no/minimal, mild, moderate, or severe depressive symptoms). In the present sample, the BDI II showed excellent internal consistency (Cronbach's $\alpha = 0.94$).

The trait subscale of the State-Trait-Anxiety-questionnaire (STAI; Spielberger et al., 1970; Laux et al., 1981) was used to measure trait anxiety. This subscale measures anxiety as a trait by using 20 items rated on a scale from one (almost never) to four (almost always). The STAI showed excellent internal consistency for the present sample (Cronbach's $\alpha = 0.95$).

2.4. Experiences of maltreatment and peer victimization

Experiences of relational peer victimization were assessed by using the *Fragebogen zu belastenden Sozialerfahrungen* (FBS, Adverse Social Experiences Quesstionnaire; Sansen et al., 2013). This questionnaire retrospectively assesses experiences of various forms of relational peer victimization, distinguishing experiences that occurred during childhood (age 6–12) and adolescence (age 13–18) by using 22 items asking about whether a specific social situation was experienced or not. In the present sample, the FBS showed excellent internal consistency (Cronbach's $\alpha = 0.90$). Although the FBS consists of two subscales (separating experiences in childhood and adolescence) it is recommended to use the total score, as there is evidence that it is superior to the subscales in capturing stressful social experiences (Sansen et al., 2013).

For examining experiences of child maltreatment, the German version of the Childhood Trauma Questionnaire (CTQ; Wingenfeld et al., 2010) was used to retrospectively assess different forms of child maltreatment experiences. The CTQ consists of 28 items on five subscales (physical maltreatment, physical neglect, emotional maltreatment, emotional neglect, and sexual abuse). For the total number of items, the CTQ in our sample showed excellent internal consistency (Cronbach's $\alpha = 0.90$). For the CTQ subscales of emotional abuse, emotional neglect, physical abuse, and sexual abuse internal consistency was acceptable to excellent (all $\alpha > 0.79$). As found in previous research (Klinitzke et al., 2012) the physical neglect subscale demonstrated only a questionable internal consistency (Cronbach's $\alpha = 0.60$). In addition to the five subscales, the CTQ captures the tendency to underreport maltreatment experiences with the minimization/denial scale (three items). Values above zero indicate response bias (false negatives) (Bernstein et al., 1994).

2.5. Paradigm and stimuli

For measuring attentional biases, the facial dot-probe paradigm (MacLeod et al., 1986) was used. A fixation cross was presented in the center of the screen for 500 ms. This was followed by the simultaneous and horizontal presentation of two still images for 500 ms. In 80% of the trials, one of the images was an emotional face and the other was a neutral face. In 20% of the trials, both images were neutral. Then a gray dot appeared on one of the two sides of the screen and replaced one of the two images. In congruent trials, the dot replaced the emotional face, in incongruent trials the dot replaced the neutral one. The participants were asked to indicate as quickly and accurately as possible if the dot was presented either on the right or the left side of the screen (by pressing the key 'E' for left and the key 'I' for right on the desktop version of the experiment or by clicking on the right/left side of the screen in the mobile version of the experiment). Four different emotional facial expressions (sad, happy, angry, and disgusted) and neutral facial expressions were used. In addition to the emotional-neutral trials, there were also neutral-neutral trials serving as baseline trials for measuring the different kinds of attentional bias scores. A total of 50 different pictures of 10 actors (five men, five women) were taken from the Radboud Faces Database (Langner et al., 2010). Each actor with each emotion was presented twice. Accordingly, the neutral images were presented more often. The order of trials and the selection of the individual emotions were randomized.

2.6. Data reduction

Drawing from previous studies, the reaction time data were adjusted in several steps (Koster et al., 2004; Bardel et al., 2013; Iffland and Neuner, 2022). Trials in which the location of the dot was incorrectly reported were removed from the trials to be analyzed (1.3% of all trials). No participant had an error rate higher than 25%. In addition, all trials in which subjects had a reaction time of <150 ms or more than 2,000 ms were not included in the analyses (0.1% of all trials). Moreover, individuals whose mean reaction time deviated more than 3 SD from the sample mean reaction time were excluded from the analyses ($n = 1$). In addition, individual trials were removed in which the reaction time deviated more or less than 2 SD from the individual mean reaction time (4.4% of all trials). For measuring the attentional bias scores for each trial type (angry-neutral, sad-neutral, disgust-neutral, happy-neutral) the overall attentional bias score was calculated by subtracting the reaction times for congruent trials (i.e., trials in which the dot replaced the emotional face) from the reaction time for incongruent trials (i.e., trials in which the dot replaced the neutral face). Attention biases are reflected in shorter reaction times for the dot when attention was focused on this area and longer reaction times for the dot when attention was not focused there. Based on the calculation of the score, positive values for the attentional bias score indicated that the attention was on the emotional faces, whereas negative values indicated that the attention of the subjects was not on the emotional face, but the neutral face. To specify altered attentional processes more precisely with respect to the different types of attentional biases, the orientation score and the disengaging score were calculated in addition to the attentional bias score to capture processes of facilitated attention or difficulties of disengagement (Koster et al., 2004). For calculating the orienting score, the reaction time for congruent trials was subtracted from the reaction time for trials in which two neutral faces were presented. This score provides information about whether subjects shifted their attention more quickly to the emotional stimulus. In addition, to gain information about whether subjects had difficulty shifting their attention away from the emotional stimuli, the disengaging score was calculated. For this purpose, reaction times in trials in which two neutral pictures were presented were subtracted from reaction times in incongruent trials.

2.7. Statistical analyses

For sample size estimation a statistical power analysis was calculated for the multiple regression analyses using G Power 3.1 (Faul et al., 2009). Based on previous results (Günther et al., 2015; Iffland and Neuner, 2020, 2022) a medium to large effect size (Cohen, 1988) was assumed (Cohen's $f^2 = 0.25$). Thus, with $\alpha =$

TABLE 2 Pearson correlation coefficients of peer victimization and the different types of child maltreatment experiences and psychopathological measures.

Trial type	Peer victimization[a]	Emotional abuse[b]	Emotional neglect[b]	Physical abuse[b]
	r	*r*	*r*	*r*
Peer victimization	-	-	-	-
Emotional abuse	0.40***	-	-	-
Emotional neglect	0.41***	0.79***	-	-
Physical abuse	0.28**	0.62***	0.52**	-
Psychopathology [c]	0.54***	0.46***	0.47***	0.33**
Trait anxiety [d]	0.42***	0.56***	0.52**	0.27*
Symptoms of depression[e]	0.44***	0.46***	0.45***	0.21*

*p<0.05, **p<0.01, and ***p<0.001; *p* values are FDR-adjusted; [a]Fragebogen zu belastenden Sozialerfahrungen; [b]Childhood trauma questionnaire; [c]Symptom Checklist-27; [d]State trait anxiety inventory (Trait); [e]Beck depression inventory.

0.05, power = 0.95, and the initially planned inclusion of seven predictors (age, psychopathology, emotional abuse, emotional neglect, physical abuse, sexual abuse, peer victimization) the required sample size was $N = 86$.

Statistical analyses were performed using the Statistical Package for the Social Sciences (IBM SPSS Statistics 28). For all analyses, a significance level of $p \leq 0.05$ was used. Correlation analyses and *t*-tests were adjusted for multiple comparisons using false discovery rate (FDR) correction (Benjamini and Hochberg, 1995). To calculate the influence of peer victimization on the different attentional bias scores and to control for the influence of child maltreatment experiences, several sets of hierarchical multiple regression analyses were calculated. Two subscales of the CTQ were not included in the analyses: the sexual abuse subscale due to a lack of variance in our sample, and the physical neglect subscale due to the weak internal consistency and high intercorrelations with other subscales (Klinitzke et al., 2012). For completeness and comparability, the two subscales were nevertheless included in the descriptive statistics. FDR-adjusted Pearson correlation coefficients of peer victimization, the three subscales of child maltreatment, and psychopathological measures are shown in Table 2. To control for the influence of symptoms of psychopathology and age of the participants, the first step of all regression models included the sum score of the SCL-27 and age. Due to the high correlation of the SCL-27 scores with the BDI II scores ($r = 0.84$, $p < 0.001$) and the STAI scores ($r = 0.76$, $p < 0.001$), only the SCL-27 was included in the regression analyses. In a second step, the sum scores of the CTQ subscales were included in the model (i.e., subscales of emotional abuse, emotional neglect, and physical abuse). In a final step, the FBS sum score (i.e., peer victimization) was included as the last predictor in the model. These regression analyses were conducted separately for the individual bias indices and the respective emotion presented. Participants who scored on all three items of the minimization/denial subscale of the CTQ ($n = 3$) were excluded from the analyses (Iffland et al., 2013; Ross et al., 2019). As the pattern of results did not change, the results reported refer to the whole sample. Analyses showed no violation of the multicollinearity assumption (all tolerances ≥ 0.31; all variance inflation factors ≤ 3.28).

TABLE 3 Results of one sample t-tests for the different index scores.

Trial type	*M (SD)*	*t*(89)	*p*	Cohen's *d*
Attentional bias score				
Anger	1.67 (25.04)	0.63	0.634	|0.07|
Disgust	−5.63 (27.15)	−1.97	0.156	|0.21|
Sadness	1.87 (28.99)	0.61	0.591	|0.06|
Happiness	−4.81 (29.74)	−1.53	0.340	|0.16|
Orienting score				
Anger	−0.35 (21.45)	−0.16	0.877	|0.02|
Disgust	−7.20 (22.50)	−3.04	0.018*	|0.32|
Sadness	−3.53 (23.13)	−1.45	0.259	|0.15|
Happiness	−8.55 (25.65)	−3.16	0.024*	|0.33|
Disengaging score				
Anger	2.02 (22.42)	0.86	0.591	|0.09|
Disgust	1.57 (20.34)	0.73	0.621	|0.08|
Sadness	5.40 (25.89)	1.98	0.204	|0.21|
Happiness	3.74 (24.21)	1.47	0.292	|0.16|

*p<0.05; *p* values are FDR-adjusted.

3. Results

A detailed description of the attentional bias index scores for the different emotions, the mean values, standard deviations, and one-sample *t*-tests of the absolute index scores for the presentation of one emotion each are shown in Table 3. The results of the *t*-tests showed that the orienting scores for disgusted and happy faces differed significantly from zero and were negative, indicating attentional avoidance of happy and disgusted facial expressions.

The bivariate Pearson correlation coefficients between maltreatment experiences and the different index scores for each trial type can be found in Table 4. The correlations between the FBS sum score and the different index scores for each emotion were not significant (all FDR corrected *p*'s > 0.05). The analyses showed a positive correlation between the emotional abuse score and the disengaging score for sad-neutral trials as well as a positive correlation between the emotional neglect score and the disengaging score for sad-neutral trials.

3.1. Peer victimization

The hierarchical regression analyses for sad-neutral trials are presented in Table 5. Analyses showed that peer victimization made a significant contribution of variance in the prediction of the orienting score for sad faces. Here, peer victimization was not only the strongest predictor but also the only one with a significant positive association with the orienting score for sad faces. Higher scores on the FBS, and thus more reported peer victimization experiences, were associated with higher scores on the orienting score in the present sample. By including this predictor in the third step, the contribution to variance was 8% [final model: $F_{(6,83)} = 2.34$, adjusted $R^2 = 0.08$, $p = 0.039$]. There was no significant relationship between the index scores and the level of peer victimization experiences for angry faces (see Table 6). Similarly, no significant effects were found in response to disgusted faces for the attentional bias score [final model $F_{(6,83)} = 0.91$, adjusted $R^2 = -0.01$, $p = 0.494$], the orienting score [final model: $F_{(6,83)} = 0.62$, adjusted $R^2 = -0.03$, $p = 0.713$] and the disengaging score [final model: $F_{(6,83)} = 1.47$, adjusted $R^2 = 0.03$, $p = 0.197$]. In addition, there were no significant effects for trials in which happy facial expressions were presented for the attentional bias score, the orienting score, and the disengaging score (see Table 7).

3.2. Further analyses of child maltreatment

Regarding the prediction of attentional biases in angry-neutral trials (see Table 6), the regression models showed that experiences of emotional abuse and emotional neglect were, besides age, the only significant predictors in the final regression model [$F_{(6,83)} = 3.08$, adjusted $R^2 = 0.12$, $p = 0.009$]. The associations with the attentional bias score behaved in opposite ways. Higher scores on the emotional abuse subscale were associated with higher attentional bias scores, whereas higher scores on emotional neglect were associated with lower attentional bias scores on angry faces. For the orienting score, emotional abuse was a significant predictor, with an overall non-significant final model [final model: $F_{(6,83)} = 1.97$, adjusted $R^2 = 0.06$, $p = 0.079$]. Associations between emotional maltreatment experiences and the orienting score for happy faces could also be found in happy-neutral trials (see Table 7). This relationship was inverse to that found for angry faces. Emotional abuse experiences were associated here with lower scores and emotional neglect with higher scores for happy faces. However, the overall model for the orienting score in happy-neutral trials was not significant [final model: $F_{(6,83)} = 1.86$, adjusted $R^2 = 0.06$, $p = 0.098$].

4. Discussion

Given the ambiguous findings on attentional biases in the context of peer victimization, the present work provided new insights into the relationship between relational peer victimization and attentional biases beyond the influence of child maltreatment experiences. In this context, the present study was designed to provide differentiated accounts of attentional biases in the context of maltreatment and peer victimization experiences, thus extending previous research. Consistent with our hypothesis we found altered attentional processes in individuals reporting higher levels of victimization experiences in the present sample. However, this influence was found in sad faces and not, as previously hypothesized, in emotions that were expected to be relevant as threatening stimuli in the context of peer victimization. Furthermore, altered attention processes were found in individuals reporting experiences of emotional maltreatment when confronted with angry facial expressions.

In the present study, the results indicated evidence for facilitated attention to sad facial expressions in individuals with higher levels of relational peer victimization experiences beyond the influence of experiences of child maltreatment. This effect was seen even when controlling for symptoms of psychopathology. These findings were consistent with the results of Günther et al. (2015), who also found facilitated attentional orienting to sad faces in depressed individuals with experiences of child maltreatment when controlling for depressive symptoms. Previous research suggested that attentional biases not only manifest in biased attention regarding potentially threatening stimuli but could also be influenced by a person's mood or are mood-congruent (Koster et al., 2005; Hankin et al., 2010; Romens and Pollak, 2012; Günther et al., 2015). Similarly, previous research revealed the existence of attentional biases in currently depressed or at-risk children and adolescents (e.g., Joormann et al., 2007; Hankin et al., 2010). Accordingly, the presentation of faces in the current study may have triggered negative emotions associated with social interactions, which may have facilitated processing of sad stimuli. In line with this argument, previous research has emphasized the relevance of sadness in the context of victimization and maltreatment. Victims of bullying tend to be insecure and fearful and they are more likely to have a negative view of themselves and rate themselves as stupid or flawed (Olweus, 1994). In a study by Mahady Wilton et al. (2000) the authors observed the behavior of elementary school children and found signs of sadness significantly more often in victims of bullying than in perpetrators, which could be related to the perceived failures in achieving one's goals in social situations. The authors note that sadness signals to the perpetrator that his goal of causing suffering is being met and thus becomes reinforcing, increasing the likelihood of becoming a victim (Mahady Wilton et al., 2000). In addition, previous studies found increased self-reported sadness among victims of bullying (Camodeca and Goossens, 2005; Glew et al., 2005). However, heightened attention for sadness cues is not exclusively associated with experiences of peer victimization. Romens and Pollak (2012) used a mood induction before a dot-probe task with depression-relevant cues and found that children with experiences of physical abuse showed heightened attention for these cues after the induction of sadness.

TABLE 4 Pearson correlation coefficients of the different types of maltreatment experiences and the index scores for each trialtype.

Trial type	Peer victimization[a]	Emotional abuse[b]	Emotional neglect[b]	Physical abuse[b]
	r	*r*	*r*	*r*
Attentional bias score				
Anger	−0.07	0.03	−0.09	−0.04
Disgust	0.01	−0.16	−0.21	−0.09
Sadness	0.18	0.19	0.13	0.15
Happiness	−0.21	−0.13	−0.05	−0.02
Orienting score				
Anger	0.02	0.14	0.02	0.00
Disgust	0.01	−0.09	−0.12	−0.05
Sadness	0.21	−0.13	−0.18	−0.04
Happiness	−0.10	−0.14	−0.02	−0.03
Disengaging score				
Anger	−0.10	−0.10	−0.12	−0.04
Disgust	−0.01	−0.11	−0.15	−0.07
Sadness	0.02	0.32*	0.31*	0.19
Happiness	−0.15	−0.01	−0.04	0.00

*p<0.05; *p* values are FDR-adjusted; [a]Fragebogen zu belastenden Sozialerfahrungen; [b]Childhood trauma questionnaire.

This result is in line with various other studies showing altered responses on a behavioral and neural level in studies using sad faces in participants with various forms of traumatic childhood experiences (for a review see Saarinen et al., 2021). Hence, findings of an altered reaction to sad facial expressions in the wake of peer victimization in the present study may also apply to adverse childhood experiences in general. In conjunction with evidence of mood-congruent bias in depression and at-risk depression (e.g., Joormann et al., 2007; Hankin et al., 2010), the present findings may be indicative of biased information processing in peer victimized individuals that may be relevant in putting individuals at risk for the development of psychopathology. Following Rosen et al. (2007), victims may implicitly associate themselves with victimization which decisively influences cognitions, behavior, and emotions in future social situations.

There was no influence of peer victimization on participant scores when angry or disgusted faces were presented. Further, there was no significant influence of peer victimization on reaction times for happy faces although the results of the one sample *t*-tests suggest that participants generally showed significant avoidance of happy and disgusted faces. Therefore, the findings for the overall sample are not reflected in the analyses for peer victimization. The present results contradict the findings of Iffland et al. (2019) and Iffland and Neuner (2022) who found a significant contribution of peer victimization experiences for attentional biases for potentially threatening stimuli and positive emotional stimuli in their studies. Using a social conditioning task or social evaluative words in these studies, the participants presumably established a reference to themselves in terms of the potentially threatening nature of the stimuli, which may have significantly influenced attentional processes. The simple presentation of emotional faces used in the present study may not activate the social victim schema in a way that leads to higher vigilance for threatening stimuli or emotional stimuli in general. This may lead to the finding that emotions, which were expected to be relevant in the context of peer victimization, were not associated with attentional biases here. We suspect that the mere presentation of emotional faces is more of a projection screen for one's emotional state. Since the results of Iffland et al. (2019) and Iffland and Neuner (2022) also point in different directions concerning the type of attentional biases, it can be assumed that stimulus choice is likely to be crucial for the presence and nature of attentional biases.

Furthermore, analyses showed altered attentional processes that were related to higher levels of emotional childhood maltreatment. These processes were particularly evident for angry faces and suggest that experiences of emotional abuse led to increased attention toward angry faces. The results confirmed the assumption that attention processing of angry faces as potentially threatening stimuli is influenced by adverse childhood experiences (Gibb et al., 2009; Kelly et al., 2015; Iffland and Neuner, 2020). In addition, our findings indicated attentional avoidance of happy facial expressions in individuals reporting emotional abuse experiences, which may be indicative of dysfunctional emotion regulation. In support of this hypothesis, Burns et al. (2010) showed that experiences of emotional abuse were associated with difficulties in emotion regulation. In contrast to experiences of emotional abuse, our results showed that emotional neglect was associated with avoidance of angry faces. These differentiated findings for different subtypes of maltreatment experiences are in line with the results of Iffland and Neuner (2020) who used a face in the crowd task in their study to highlight that attentional processes differ between the two forms of emotional

TABLE 5 Hierarchical multiple regression analyses for sad-neutral trials.

Variable	β	R^2	Adjusted R^2	$\triangle R^2$	F
Attentional bias score					
Step 1		0.03	0.01	0.03	1.34
Age	−0.06				
SCL-27	0.04				
Step 2		0.05	−0.01	0.02	0.88
Emotional abuse	0.15				
Emotional neglect	−0.07				
Physical abuse	0.04				
Step 3		0.06	−0.01	0.01	0.84
Peer victimization	0.11				
Orienting score					
Step 1		0.03	0.01	0.03	1.35
Age	−0.10				
SCL-27	−0.02				
Step 2		0.07	0.02	0.04	1.29
Emotional abuse	−0.08				
Emotional neglect	−0.27				
Physical abuse	0.07				
Step 3		0.15	0.08	0.08	2.34*
Peer victimization	0.33**				
Disengaging score					
Step 1		0.03	0.01	0.03	1.20
Age	0.02				
SCL-27	0.06				
Step 2		0.12	0.06	0.09	2.18
Emotional abuse	0.25				
Emotional neglect	0.17				
Physical abuse	−0.02				
Step 3		0.14	0.07	0.02	2.18
Peer victimization	−0.18				

*p<0.05, **p<0.01; β coeffizients correspond to those of the final model.

maltreatment. They reported a faster detection of negative faces in victims of emotional abuse, whereas slower recognition of negative and neutral faces was more likely in victims of emotional neglect. However, in contrast to the findings of Iffland and Neuner (2020), our findings regarding emotional neglect are less indicative of a general avoidance of emotional faces than of more differentiated processes, possibly involving the avoidance of highly salient stimuli (here angry faces). This is supported by the finding that emotional neglect was associated with an attentional shift toward happy faces. These findings could thus be the result of emotion regulation strategies, which in the context of emotional neglect could be associated with avoidance of threatening stimuli and a shift toward positive stimuli. However, because the final regression model for the orienting score for happy-neutral trials did not reach significance, these conclusions must be viewed with caution. Against our expectations, there was no relationship between maltreatment experiences and reaction times for disgust-neutral trials. Horstmann (2003) showed that disgusted facial expressions are more likely to be interpreted as expressions of emotional experience, whereas anger is more likely to be perceived as having an informative character at the interpersonal level. Future studies should therefore consider the use of disapproval faces as stimuli for social rejection (Burklund et al., 2007). Our results indicate differentiated attentional processes that are influenced by various forms of maltreatment, but particularly in individuals with emotional or relational maltreatment experiences.

TABLE 6 Hierarchical multiple regression analyses for anger-neutral trials.

Variable	β	R^2	Adjusted R^2	$\triangle R^2$	F
Attentional bias score					
Step 1		0.12	0.10	0.12	5.67**
Age	0.37***				
SCL-27	−0.13				
Step 2		0.18	0.13	0.06	3.68**
Emotional abuse	0.42*				
Emotional neglect	−0.40*				
Physical abuse	−0.08				
Step 3		0.18	0.12	0.00	3.08**
Peer victimization	0.06				
Orienting score					
Step 1		0.04	0.01	0.04	1.58
Age	0.10				
SCL-27	−0.32*				
Step 2		0.11	0.06	0.08	2.12
Emotional abuse	0.45*				
Emotional neglect	−0.19				
Physical abuse	−0.12				
Step 3		0.13	0.06	0.01	1.97
Peer victimization	0.14				
Disengaging score					
Step 1		0.08	0.06	0.08	3.58*
Age	0.31**				
SCL-27	0.16				
Step 2		0.12	0.06	0.04	2.20
Emotional abuse	0.04				
Emotional neglect	−0.26				
Physical abuse	0.03				
Step 3		0.12	0.05	0.00	1.85
Peer victimization	−0.06				

*$p<0.05$, **$p<0.01$, and ***$p<0.001$; β coeffizients correspond to those of the final model.

4.1. Limitations

Limitations of the present study must be considered when interpreting the results. One-third of the participants stated that they were currently suffering from a mental disorder. For reasons of anonymity, no information could be collected on whether participants were patients of the outpatient clinics. It cannot be excluded that treatment or current medication influenced the attentional processes or reaction times. However, by including psychopathological symptom severity in the regression models, and by adjusting the reaction time data, we were able to reduce the potential influence on our results. In addition, the data did not allow us to assert causal relationships due to the cross-sectional design of our study. Longitudinal studies for analyzing the relationship to mental health should be addressed in the future. The retrospective assessment of experiences of child maltreatment and peer victimization as self-reports also limits the interpretability of the results as they may be affected by distortions (Baldwin et al., 2019). However, the questionnaires used in the present study have repeatedly shown good reliability and validity and are therefore suitable for the retrospective recording of stressful life experiences (Klinitzke et al., 2012; Sansen et al., 2013). Another limitation of the study is the interpretation of reaction times when using the dot-probe task. However, the reliability can be increased by

TABLE 7 Hierarchical multiple regression analyses for happy-neutral trials.

Variable	β	R^2	Adjusted R^2	$\triangle R^2$	F
Attentional bias score					
Step 1		0.05	0.03	0.05	2.16
Age	−0.23*				
SCL-27	−0.06				
Step 2		0.08	0.02	0.03	1.42
Emotional abuse	−0.29				
Emotional neglect	0.28				
Physical abuse	0.10				
Step 3		0.11	0.05	0.03	1.77
Peer victimization	−0.23				
Orienting score					
Step 1		0.06	0.04	0.06	2.81
Age	−.21				
SCL-27	−.23				
Step 2		0.12	0.07	0.06	2.24
Emotional abuse	−0.38*				
Emotional neglect	0.38*				
Physical abuse	0.10				
Step 3		0.12	0.06	0.00	1.86
Peer victimization	−0.03				
Disengaging score					
Step 1		0.01	−0.02	0.01	0.21
Age	−0.06				
SCL-27	0.17				
Step 2		0.01	−0.05	0.00	0.16
Emotional abuse	0.05				
Emotional neglect	−0.06				
Physical abuse	0.02				
Step 3		0.05	−0.02	0.04	0.74
Peer victimization	−0.25				

*p<0.05; β coeffizients correspond to those of the final model.

the experimental design of the dot-probe task, e.g., by choosing a horizontal instead of a vertical stimulus presentation (Price et al., 2015). Moreover, the findings indicated that the dot-probe paradigm was sensitive enough to allow differentiation between emotions. Nevertheless, our results should be interpreted with caution, especially since the one-sample *t*-tests for the absolute orienting scores are only significant for happy and disgusted facial expressions and not for sad and angry faces. In addition, the results should be interpreted with caution due to the low controllability of the entire study, caused by its realization as an online study. It should be noted that it was not possible to determine which device was used for participation. It cannot be ruled out that the type of device (computer or mobile device) had an influence on the results. In addition, the online study could not control the situational conditions under which the performance took place. However, by adjusting the experimental data, we were able to minimize the influence that the behavior would have had if the instructions had not been followed or if the subjects had been unfocused or distracted. Moreover, results were consistent with the findings of several studies that have used reaction times, and also neural measures (Günther et al., 2015; Saarinen et al., 2021). Future research should nevertheless consider additional measures such as physiological measures to be able to interpret the results of a dot-probe task more reliably. In this context, physiological measurements, and the analysis of event-related potentials could provide more accurate information about attentional processes, since cortical responses can be recorded and analyzed in the range of milliseconds. The analysis of reaction times is limited to a specific point in time (here 500 ms after stimulus onset). So, our results do not provide information about the course of the attentional process. It cannot be excluded that attention has already been shifted. Future studies should therefore include variable presentation durations in addition to physiological outcomes to capture different stages of the attentional process (Chapman et al., 2019). Furthermore, it should be noted that neutral facial expressions were chosen as baseline. This may have influenced the results, as there is some evidence that individuals with experience of maltreatment perceive neutral faces as negative (Pollak et al., 2000). Nevertheless, our results indicate differences in attentional processes with respect to negative and neutral facial expressions, yet future work could consider the use of calm faces instead of neutral faces (Kelly et al., 2015).

5. Conclusion

In line with previous results, our study showed that experiences of relational peer victimization and emotional child maltreatment in childhood and adolescence influence attentional processes in adulthood. Higher levels of peer victimization were associated with facilitated attention to sad facial expressions in our sample. The results are thus indicative of mood-congruent attentional biases in individuals who have experienced relational peer violence. In addition, altered attentional processes for angry faces were present in participants with higher levels of emotional child maltreatment experiences. Adverse childhood experiences, particularly experiences of emotional maltreatment and relational peer victimization, can thus be considered relevant to the development of cognitive schemata that continue to be activated in adulthood, and therefore can potentially influence new experiences, feelings, thoughts in social situations, and thus presumably mental health.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Committee of Bielefeld University. Written informed consent was not provided because of the design as an online study. The participants gave their informed consent by clicking on a box.

Author contributions

KB contributed to the conception and design of the work, the acquisition, analysis, interpretation of the data, drafted, revised, approved the manuscript, and ensures the accuracy and integrity of any part of the work. BI was the chief investigator for this study, contributed to the conception of the study, supervised data analyses, participated in the interpretation of the data, and critically revised the manuscript for important intellectual content. All authors contributed to the article and approved the submitted version.

Funding

We acknowledge support for the publication costs by the Open Access Publication Fund of Bielefeld University and the Deutsche Forschungsgemeinschaft (DFG).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aults, C. D., Machluf, K., Sellers, P. D., and Jones, N. A. (2019). Adolescent girls' biological sensitivity to context: heart rate reactivity moderates the relationship between peer victimization and internalizing problems. *Evolut. Psychol. Sci.* 5, 178–185. doi: 10.1007/s40806-018-0176-2

Baldwin, J. R., Reuben, A., Newbury, J. B., and Danese, A. (2019). Agreement between prospective and retrospective measures of childhood maltreatment: a systematic review and meta-analysis. *JAMA Psychiatry* 76, 584–593. doi: 10.1001/jamapsychiatry.2019.0097

Bardel, M.-H., Woodman, T., Perreaut-Pierre, E., and Barizien, N. (2013). The role of athletes' pain-related anxiety in pain-related attentional processes. *Anxiety Stress Coping* 26, 573–583. doi: 10.1080/10615806.2012.757306

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., and van IJzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. *Psychol. Bull.* 133, 1–24. doi: 10.1037/0033-2909.133.1.1

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bernstein, D. P., Fink, L., Handelsman, L., Foote, J., Wenzel, K., Sapareto, E., et al. (1994). Initial reliability and validity of a new retrospective measure of child abuse and neglect. *Am. J. Psychiatry* 151, 1132–1136. doi: 10.1176/ajp.151.8.1132

Brendgen, M., Ouellet-Morin, I., Lupien, S., Vitaro, F., Dionne, G., and Boivin, M. (2017). Does cortisol moderate the environmental association between peer victimization and depression symptoms? A genetically informed twin study. *Psychoneuroendocrinology* 84, 42–50. doi: 10.1016/j.psyneuen.2017.06.014

Burklund, L. J., Eisenberger, N. I., and Lieberman, M. D. (2007). The face of rejection: rejection sensitivity moderates dorsal anterior cingulate activity to disapproving facial expressions. *Soc. Neurosci.* 2, 238–253. doi: 10.1080/17470910701391711

Burns, E. E., Jackson, J. L., and Harding, H. G. (2010). Child maltreatment, emotion regulation, and posttraumatic stress: The impact of emotional abuse. *J. Aggress. Maltreat. Trauma* 19, 801–819. doi: 10.1080/10926771.2010.522947

Camodeca, M., and Goossens, F. A. (2005). Aggression, social cognitions, anger and sadness in bullies and victims. *J. Child Psychol. Psychiatry* 46, 186–197. doi: 10.1111/j.1469-7610.2004.00347.x

Carr, A., Duff, H., and Craddock, F. (2020). A systematic review of reviews of the outcome of noninstitutional child maltreatment. *Trauma Violence Abuse* 21, 828–843. doi: 10.1177/1524838018801334

Chapman, A., Devue, C., and Grimshaw, G. M. (2019). Fleeting reliability in the dot-probe task. *Psychol. Res.* 83, 308–320. doi: 10.1007/s00426-017-0947-6

Cisler, J. M., and Koster, E. H. (2010). Mechanisms of attentional biases towards threat in anxiety disorders: an integrative review. *Clin. Psychol. Rev.* 30, 203–216. doi: 10.1016/j.cpr.2009.11.003

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

De Los Reyes, A., and Prinstein, M. J. (2004). Applying depression-distortion hypotheses to the assessment of peer victimization in adolescents. *J. Clin. Child Adolesc. Psychol.* 33, 325–335. doi: 10.1207/s15374424jccp3302_14

Earnshaw, V. A., Elliott, M. N., Reisner, S. L., Mrug, S., Windle, M., Emery, S. T., et al. (2017). Peer victimization, depressive symptoms, and substance use: a longitudinal analysis. *Pediatrics* 139, 3426. doi: 10.1542/peds.2016-3426

Fani, N., Bradley-Davino, B., Ressler, K. J., and McClure-Tone, E. B. (2011). Attention bias in adult survivors of childhood maltreatment with and without posttraumatic stress disorder. *Cognit. Ther. Res.* 35, 57–67. doi: 10.1007/s10608-010-9294-2

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/BRM.41.4.1149

Gibb, B. E., Schofield, C. A., and Coles, M. E. (2009). Reported history of childhood abuse and young adults' information-processing biases for facial displays of emotion. *Child Maltreat.* 14, 148–156. doi: 10.1177/1077559508326358

Glew, G. M., Fan, M.-Y., Katon, W., Rivara, F. P., and Kernic, M. A. (2005). Bullying, psychosocial adjustment, and academic performance in elementary school. *Arch. Pediatr. Adolesc. Med.* 159, 1026–1031. doi: 10.1001/archpedi.159.11.1026

Günther, V., Dannlowski, U., Kersting, A., and Suslow, T. (2015). Associations between childhood maltreatment and emotion processing biases in major depression: results from a dot-probe task. *BMC Psychiatry* 15, 123. doi: 10.1186/s12888-015-0501-2

Hankin, B. L., Gibb, B. E., Abela, J. R. Z., and Flory, K. (2010). Selective attention to affective stimuli and clinical depression among youths: role of anxiety and specificity of emotion. *J. Abnorm. Psychol.* 119, 491–501. doi: 10.1037/a0019609

Hardt, J., and Gerbershagen, H. (2001). Cross-validation of the SCL-27: a short psychometric screening instrument for chronic pain patients. *Eur. J. Pain* 5, 187–197. doi: 10.1053/eujp.2001.0231

Hautzinger, M., Keller, F., and Kühner, C. (2006). *Beck Depressions Inventar: Revision (BDI-II)*. Frankfurt: Harcourt Test Services.

Hébert, M., Langevin, R., and Daigneault, I. (2016). The association between peer victimization, PTSD, and dissociation in child victims of sexual abuse. *J. Affect. Disord.* 193, 227–232. doi: 10.1016/j.jad.2015.12.080

Hoepfel, D., Günther, V., Bujanow, A., Kersting, A., Bodenschatz, C. M., and Suslow, T. (2022). Experiences of maltreatment in childhood and attention to facial emotions in healthy young women. *Sci. Rep.* 12, 1–12. doi: 10.1038/s41598-022-08290-1

Horstmann, G. (2003). What do facial expressions convey: Feeling states, behavioral intentions, or actions requests? *Emotion* 3, 150. doi: 10.1037/1528-3542.3.2.150

Iffland, B., Brähler, E., Neuner, F., Häuser, W., and Glaesmer, H. (2013). Frequency of child maltreatment in a representative sample of the german population. *BMC Public Health* 13, 1–7. doi: 10.1186/1471-2458-13-980

Iffland, B., and Neuner, F. (2020). Varying cognitive scars - differential associations between types of childhood maltreatment and facial emotion processing. *Front. Psychol.* 11, 732. doi: 10.3389/fpsyg.2020.00732

Iffland, B., and Neuner, F. (2022). Peer victimization influences attention processing beyond the effects of childhood maltreatment by caregivers. *Front. Psychol.* 13, 784147. doi: 10.3389/fpsyg.2022.784147

Iffland, B., Weitkämper, A., Weitkämper, N. J., and Neuner, F. (2019). Attentional avoidance in peer victimized individuals with and without psychiatric disorders. *BMC Psychology* 7, 12. doi: 10.1186/s40359-019-0284-1

Jaffee, S. R. (2017). Child maltreatment and risk for psychopathology in childhood and adulthood. *Annu. Rev. Clin. Psychol.* 13, 525–551. doi: 10.1146/annurev-clinpsy-032816-045005

Joormann, J., Talbot, L., and Gotlib, I. H. (2007). Biased processing of emotional information in girls at risk for depression. *J. Abnorm. Psychol.* 116, 135–143. doi: 10.1037/0021-843X.116.1.135

Juvonen, J., Yueyan, W.ang, and Espinoza, G. (2011). Bullying experiences and compromised academic performance across middle school grades. *J. Early Adolesc.* 31, 152–173. doi: 10.1177/0272431610379415

Kelly, P. A., Viding, E., Puetz, V. B., Palmer, A. L., Mechelli, A., Pingault, J.-B., et al. (2015). Sex differences in socioemotional functioning, attentional bias, and gray matter volume in maltreated children: A multilevel investigation. *Dev. Psychopathol.* 27, 1591–1609. doi: 10.1017/S0954579415000966

Klinitzke, G., Romppel, M., Häuser, W., Brähler, E., and Glaesmer, H. (2012). The german version of the childhood trauma questionnaire (CTQ): psychometric characteristics in a representative sample of the general population. *PPmP Psychotherapie Psychosomatik Medizinische Psychologie* 62, 47–51. doi: 10.1055/s-0031-1295495

Klomek, A. B., Sourander, A., Niemel,ä, S., Kumpulainen, K., Piha, J., Tamminen, T., et al. (2009). Childhood bullying behaviors as a risk for suicide attempts and completed suicides: a population-based birth cohort study. *J. Am. Acad. Child Adolesc. Psychiatry* 48, 254–261. doi: 10.1097/CHI.0b013e318196b91f

Koster, E. H., Crombez, G., Verschuere, B., and De Houwer, J. (2004). Selective attention to threat in the dot probe paradigm: differentiating vigilance and difficulty to disengage. *Behav. Res. Ther.* 42, 1183–1192. doi: 10.1016/j.brat.2003.08.001

Koster, E. H. W., De Raedt, R., Goeleven, E., Franck, E., and Crombez, G. (2005). Mood-congruent attentional bias in dysphoria: maintained attention to and impaired disengagement from negative information. *Emotion* 5, 446–455. doi: 10.1037/1528-3542.5.4.446

Kühner, C., Bürger, C., Keller, F., and Hautzinger, M. (2007). Reliabilität und validität des revidierten Beck-Depressionsinventars (BDI-II): Befunde aus deutschsprachigen stichproben. *Nervenarzt* 78, 651–656. doi: 10.1007/s00115-006-2098-7

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., and van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognit. Emot.* 24, 1377–1388. doi: 10.1080/02699930903485076

Laux, L., Glanzmann, P., Schaffner, P., and Spielberger, C. (1981). *Das State-Trait-Angstinventar: STAI*. Weinheim: Beltz.

MacLeod, C., Mathews, A., and Tata, P. (1986). Attentional bias in emotional disorders. *J. Abnorm. Psychol.* 95, 15–20. doi: 10.1037/0021-843X.95.1.15

Mahady Wilton, M. M., Craig, W. M., and Pepler, D. J. (2000). Emotional regulation and display in classroom victims of bullying: Characteristic expressions of affect, coping styles and relevant contextual factors. *Soc. Dev.* 9, 226–245. doi: 10.1111/1467-9507.00121

McDougall, P., and Vaillancourt, T. (2015). Long-term adult outcomes of peer victimization in childhood and adolescence: Pathways to adjustment and maladjustment. *Am. Psychol.* 70, 300–310. doi: 10.1037/a0039174

Olweus, D. (1994). Bullying at school: basic facts and effects of a school based intervention program. *J. Child Psychol. Psychiatry* 35, 1171–1190. doi: 10.1111/j.1469-7610.1994.tb01229.x

Peckham, A. D., McHugh, R. K., and Otto, M. W. (2010). A meta-analysis of the magnitude of biased attention in depression. *Depress Anxiety* 27, 1135–1142. doi: 10.1002/da.20755

Pine, D. S., Mogg, K., Bradley, B. P., Montgomery, L., Monk, C. S., McClure, E., et al. (2005). Attention bias to threat in maltreated children: Implications for vulnerability to stress-related psychopathology. *Am. J. Psychiatry* 162, 291–296. doi: 10.1176/appi.ajp.162.2.291

Pollak, S. D., Cicchetti, D., Hornung, K., and Reed, A. (2000). Recognizing emotion in faces: developmental effects of child abuse and neglect. *Dev. Psychol.* 36, 679. doi: 10.1037/0012-1649.36.5.679

Price, R. B., Kuckertz, J. M., Siegle, G. J., Ladouceur, C. D., Silk, J. S., Ryan, N. D., et al. (2015). Empirical recommendations for improving the stability of the dot-probe task in clinical research. *Psychol. Assess.* 27, 365–376. doi: 10.1037/pas0000036

Quinlan, E. B., Barker, E. D., Luo, Q., Banaschewski, T., Bokde, A. L. W., Bromberg, U., et al. (2020). Peer victimization and its impact on adolescent brain development and psychopathology. *Mol. Psychiatry* 25, 3066–3076. doi: 10.1038/s41380-018-0297-9

Romens, S. E., and Pollak, S. D. (2012). Emotion regulation predicts attention bias in maltreated children at-risk for depression: emotion regulation in maltreated children. *J. Child Psychol. Psychiatry* 53, 120–127. doi: 10.1111/j.1469-7610.2011.02474.x

Rosen, P. J., Milich, R., and Harris, M. J. (2007). Victims of their own cognitions: Implicit social cognitions, emotional distress, and peer victimization. *J. Appl. Dev. Psychol.* 28, 211–226. doi: 10.1016/j.appdev.2007.02.001

Ross, N. D., Kaminski, P. L., and Herrington, R. (2019). From childhood emotional maltreatment to depressive symptoms in adulthood: The roles of self-compassion and shame. *Child Abuse Neglect* 92, 32–42. doi: 10.1016/j.chiabu.2019.03.016

Rozin, P., Haidt, J., and McCauley, C. R. (2008). "Disgust," in *Handbook of Emotions*, eds M. Lewis, J. M. Haviland-Jones, and L. F. Barrett (New York, NY: The Guilford Press), 757–776.

Saarinen, A., Keltikangas-Järvinen, L., Jääskeläinen, E., Huhtaniska, S., Pudas, J., Tovar-Perdomo, S., et al. (2021). Early adversity and emotion processing from faces: a meta-analysis on behavioral and neurophysiological responses. *Biol. Psychiatry* 6, 692–705. doi: 10.1016/j.bpsc.2021.01.002

Sansen, L., Iffland, B., Catani, C., and Neuner, F. (2013). Entwicklung und evaluation des fragebogens zu belastenden sozialerfahrungen in der peergroup (FBS). *Zeitschrift für Klinische Psychologie und Psychotherapie* 42, 34–44. doi: 10.1026/1616-3443/a000184

Sansen, L. M., Iffland, B., and Neuner, F. (2015). The trauma of peer victimization: psychophysiological and emotional characteristics of memory imagery in subjects with social anxiety disorder: the trauma of peer victimization. *Psychophysiology* 52, 107–116. doi: 10.1111/psyp.12291

Schwartz, D., Gorman, A. H., Nakamoto, J., and Toblin, R. L. (2005). Victimization in the peer group and children's academic functioning. *J Educ Psychol*. 97, 425–435. doi: 10.1037/0022-0663.97.3.425

Siegel, R. S., La Greca, A. M., and Harrison, H. M. (2009). Peer victimization and social anxiety in adolescents: Prospective and reciprocal relationships. *J. Youth Adolesc*. 38, 1096–1109. doi: 10.1007/s10964-009-9392-1p

Spielberger, C. D., Gorsuch, R. L., and Lushene, R. E. (1970). *STAI. Manual for the State-Trait Anxiety Inventory*. Edina: Consulting Psychologists Press.

Stapinski, L. A., Bowes, L., Wolke, D., Pearson, R. M., Mahedy, L., Button, K. S., et al. (2014). Peer victimization during adolescence and risk for anxiety disorder in adulthood: a prospective cohort study. *Depress Anxiety* 31, 574–582. doi: 10.1002/da.22270

Takizawa, R., Maughan, B., and Arseneault, L. (2014). Adult health outcomes of childhood bullying victimization: evidence from a five-decade longitudinal british birth cohort. *Am. J. Psychiatry* 171, 777–784. doi: 10.1176/appi.ajp.2014.13101401

Tybur, J. M., Lieberman, D., Kurzban, R., and DeScioli, P. (2013). Disgust: evolved function and structure. *Psychol. Rev*. 120, 65. doi: 10.1037/a0030778

Wingenfeld, K., Spitzer, C., Mensebach, C., Grabe, H., Hill, A., Gast, U., et al. (2010). Die deutsche version des childhood trauma questionnaire (CTQ): erste befunde zu den psychometrischen kennwerten. *PPmP Psychotherapie Psychosomatik Medizinische Psychologie* 60, 442–450. doi: 10.1055/s-0030-1247564

World Health Organization (1999). *Report of the consultation on child abuse prevention, 29-31 march 1999*. Technical report, Geneva, Schweiz.

TYPE Original Research
PUBLISHED 28 March 2023
DOI 10.3389/fpsyg.2023.1086699

OPEN ACCESS

EDITED BY
Dario Monzani,
University of Palermo, Italy

REVIEWED BY
Ilana Ritov,
Hebrew University of Jerusalem, Israel
Silvia Francesca Maria Pizzoli,
University of Milan, Italy

*CORRESPONDENCE
Marc Wyszynski
✉ m.wyszynski@uni-bremen.de

SPECIALTY SECTION
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

RECEIVED 01 November 2022
ACCEPTED 06 March 2023
PUBLISHED 28 March 2023

CITATION
Wyszynski M and Diederich A (2023) Individual differences moderate effects in an Unusual Disease paradigm: A psychophysical data collection lab approach and an online experiment. *Front. Psychol.* 14:1086699.
doi: 10.3389/fpsyg.2023.1086699

Individual differences moderate effects in an Unusual Disease paradigm: A psychophysical data collection lab approach and an online experiment

Marc Wyszynski[1]* and Adele Diederich[2]

[1]Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany,
[2]Department of Psychology, University of Oldenburg, Oldenburg, Germany

We report two studies investigating individual intuitive-deliberative cognitive-styles and risk-styles as moderators of the framing effect in Tversky and Kahneman's famous Unusual Disease problem setting. We examined framing effects in two ways: counting the number of frame-inconsistent choices and comparing the proportions of risky choices depending on gain-loss framing. Moreover, in addition to gain-loss frames, we systematically varied the number of affected people, probabilities of surviving/dying, type of disease, and response deadlines. Study 1 used a psychophysical data collection approach and a sample of 43 undergraduate students, each performing 480 trials. Study 2 was an online study incorporating psychophysical elements in a social science approach using a larger and more heterogeneous sample, i.e., 262 participants performed 80 trials each. In both studies, the effect of framing on risky choice proportions was moderated by risk-styles. Cognitive-styles measured on different scales moderated the framing effect only in study 2. The effects of disease type, probability of surviving/dying, and number of affected people on risky choice frequencies were also affected by cognitive-styles and risk-styles but different for both studies and to different extents. We found no relationship between the number of frame-inconsistent choices and cognitive-styles or risk-styles, respectively.

KEYWORDS

individual differences, framing effects, cognitive-style, risk-style, thinking-style, cognitive-experiential self-theory, framing susceptibility, frame-inconsistent choice

1. Introduction

Since Tversky and Kahneman (1981) seminal paper on framing, numerous studies have shown that decisions under risk are often influenced by the way the decision problem is presented. This phenomenon, known as framing effect, violates the normative principle of description invariance; that is, a decision must not depend on the way how it is presented. Presumably, the most famous and most applied example for framing risky choice alternatives is Tversky and Kahneman (1981) Unusual Disease Problem.[1] The problem describes two programs to combat a hypothetical disease that is expected to kill 600 people in either a

[1] According to the World Health Organization best practices for the naming of new human infectious diseases, we use a more contemporary term without labeling the disease with a country or region of origin.

positive or a negative frame. In the positive (negative) frame, 200 people can be saved (400 will die) for sure with program A (C), or 600 people will be saved (will die) with a probability of 1/3 (2/3) with program B (D). Most of the participants chose program A in the positive frame and program D in the negative frame. The framing effect in Unusual Disease Problems has repeatedly been demonstrated by more than 40 studies (see e.g., Kühberger, 1998; Levin et al., 1998; Kühberger et al., 1999; Piñon and Gambara, 2005; Steiger and Kühberger, 2018, for meta-analytic reviews).

The framing effect is typically accounted for by prospect theory (Kahneman and Tversky, 1979), but more recently the notion of dual processes have been brought into play. According to this approach, framing effects result from the interplay of two different systems of reasoning. One system, generically called "System 1," includes fast and intuitive processes, whereas the other system, often called "System 2," is described in terms of slow and deliberative processing (see e.g., Chaiken and Trope, 1999; Stanovich and West, 2000; Evans, 2008; Mukherjee, 2010; Guo et al., 2017; Roberts et al., 2021). Framing effects mainly emerge in the fast and intuitive System 1, and they tend to disappear when the slow and deliberative System 2 is engaged (see e.g., Sloman, 1996; Kahneman and Frederick, 2002, 2005). Empirical findings support these hypotheses: stronger framing effects are observed when participants are put under time pressure (Guo et al., 2017; Diederich et al., 2018, 2020; Wyszynski et al., 2020; Roberts et al., 2021; Wyszynski and Diederich, 2022), and weaker framing effects occur when people are forced to use deliberative reasoning (Miller and Fagley, 1991; Takemura, 1994; Sieck and Yates, 1997; Almashat et al., 2008).

If experimental manipulations inducing intuitive or deliberative processing can affect the strength of the framing effect then it is possible that the decision-makers individual style of processing information intuitively or deliberately may also moderate framing effects (Stanovich, 1999; Evans, 2008; Mandel and Kapler, 2018).

The cognitive-experiential self-theory (CEST; Epstein, 1994) originally introduced as a global theory of personality (Epstein, 1973) assumes a rational and an experiential system. In both systems people have constructs about the self and the world, referred to as schemata (rational) and beliefs (experiential). The experiential system has been linked to heuristics. Furthermore, CEST assumes "important individual differences in the relative degree and effectiveness with which individuals use the two modes of information processing" (Epstein, 1994, p. 719). Several scales based on CEST to measure those differences have been constructed. For instance, Epstein et al. (1996) developed the Rational-Experiential Inventory (REI) that consists of a modified version of the Need for Cognition scale (NFC, Cacioppo and Petty, 1982) and the Faith in Intuition scale (FI, Epstein et al., 1996). NFC is a measure of deliberative-rational cognitive-style. In particular, it "refers to an individual's tendency to engage in and enjoy effortful cognitive endeavors" (Cacioppo et al., 1984, p. 306). Decision makers with a high NFC score are expected to be less susceptible to the effect of framing. A person's FI reflects its intuitive-experiential processing (Epstein et al., 1996), which is characterized by a rapid, holistic, and emotional cognitive-style. Decision-makers who score high in FI are expected to produce more framing effects than those with lower FI scores.

Another concept of deliberative thinking-style is Actively Open-Minded Thinking (AOT, Baron, 1993). AOT style is characterized by the tendencies "to weight new evidence against a favored belief, to spend sufficient time on a problem before giving up, and to consider carefully the opinions of others in forming one's own" (Haran et al., 2013, p. 189). Higher AOT is associated with better decision-making performance, i.e., producing fewer framing effects. The Stimulating-Instrumental Risk Inventory (SIRI, Zaleskiewicz, 2001) measures individual risk-styles based on rational-experiential processing modes. Stimulating risk is associated with experiential risk-style and the enjoyment of risk. It may lead to faster, less analytical, and more heuristic decisions. Instrumental risk-taking relates to the rational system. High instrumental risk-takers are expected to analyze the characteristics and values of a risky choice carefully and, therefore, produce fewer framing effects.

Individual differences have been linked to framing susceptibility but the results are mixed. Some studies indicated that individual differences in intuitive-deliberative cognitive-styles and risk-styles moderate framing effects in the expected way (NFC: e.g., LeBoeuf and Shafir, 2003; Simon et al., 2004; Björklund and Bäckström, 2008; Peng et al., 2019; AOT: e.g., West et al., 2008; Erceg et al., 2022; Rachev et al., 2022), and other research, however, failed to identify a significant relationship (NFC: e.g., Corbin, 2015; Fatmawati, 2015; Stark et al., 2017; Mandel and Kapler, 2018; FI: e.g., Levin et al., 2002; Shiloh et al., 2002; Björklund and Bäckström, 2008; Stark et al., 2017; AOT: e.g., Erceg et al., 2022, study 2; Mandel and Kapler, 2018; SIRI: e.g., Mahoney et al., 2011).

These studies, however, differ substantially in the framing effect interpretations, characteristics of the decision problems presented in the experiments, sample compositions, and study designs, which may explain the discrepancy of their results. In particular, several interpretations of framing effects have been used. Peng et al. (2019) and Rachev et al. (2022) used the "resistance to framing" component of the Adult Decision Making Competence (ADMC) scale (Bruine de Bruin et al., 2007). This involves one unusual disease-like decision problem framed as gain or loss. West et al. (2008) and Mandel and Kapler (2018) counted the frame-(in)consistent choices participants made in one risky decision problem. Mandel and Kapler (2018) counted a "frame-consistent" choice if participants chose the sure option in the gain frame, and the risky option in the loss frame in a between-subjects design. West et al. (2008) counted a "frame-inconsistent" choice when participants chose the sure option in one frame and the risky option in the other frame in a within-subjects study. Other studies evaluated individual differences in proportions of choosing the sure and the risky option depending on the framing of the decision problem (e.g., Shiloh et al., 2002; Simon et al., 2004). Different framing interpretations may account for differences in strength of the framing effect and its correlation with psychometric instruments.

Furthermore, certain characteristics describing the decision problem, such as probabilities, magnitude of outcome, problem domain, and different time limits for making a choice, have been shown to influence risky choice additionally to framing (see Kühberger et al., 1999; Mahoney et al., 2011; Diederich et al., 2018, for overviews). Only a few studies investigating the impact of cognitive-styles and/or risk-styles on risky choice framing effects varied one or more problem-describing characteristics in their experiments (e.g., Mahoney et al., 2011; Corbin, 2015). None of them report any results about the relationship between problem-describing characteristics and individual differences in risky choices.

Whether or not cognitive-style and/or risk-style moderate the framing effect may further depend on the sample composition. In previous studies, many samples were composed of undergraduate or graduate university students. Using student samples could be seen as a kind of pre-selection or screening because student samples are more homogeneous and may provide a limited range of psychometric scores measured using a particular instrument as compared to a community or online sample (see, e.g., Peterson, 2001).

Finally, the study design may not have been optimal in several cases. While some studies varied the framing manipulation within participants (e.g., Levin et al., 2002; Mahoney et al., 2011; Peng et al., 2019; Erceg et al., 2022; Rachev et al., 2022), other studies relied on between-subjects designs where a particular decision problem is described by different frames and each participant responds to only one of these frames (Shiloh et al., 2002; Simon et al., 2004; Björklund and Bäckström, 2008; Fatmawati, 2015; Stark et al., 2017; Mandel and Kapler, 2018). However, several researchers pointed out that a within-subjects design is more appropriate when investigating framing effects on the individual level (Frisch, 1993; Baron, 2010; Appelt et al., 2011; Mahoney et al., 2011; Aczel et al., 2018). It allows analyzing an individual's susceptibility to framing effects based on certain individual characteristics such as cognitive-styles and risk-styles.

A key challenge in investigating framing effects using within-subjects designs is the transparency of framing manipulation. Once participants notice the similarity between frames, they may tend to give the same response in both frames (Aczel et al., 2018). The common way of dealing with this problem is adding intervening steps between the two frames, for instance, by inserting a temporal break (e.g., Levin et al., 2002; Parker and Fischhoff, 2005), inserting filling questions (e.g., Stanovich and West, 1998; LeBoeuf and Shafir, 2003; Li and Liu, 2008), or masking the frames by presenting different problems in random order (e.g., Frisch, 1993). However, framing effect strengths are often smaller in within-subjects studies than in between-subjects designs (Piñon and Gambara, 2005; Aczel et al., 2018). This difference is still commonly explained by the higher transparency of manipulations in within-subjects designs (Kahneman and Frederick, 2005).

To overcome these problems, Mahoney et al. (2011) introduced an alternative approach using a within-subjects design: The Unusual Disease Problem varied with respect to the specific disease, the number of affected people, and probabilities of surviving/dying to create five unique choice problems, each framed as gain and loss. They found strong framing effects. However, the results did not support their hypothesis that individual cognitive-styles and risk-styles moderate the framing effect.

To shed some more light on the mixed results in previous research, we have the following goals. First, we seek to extend the within-subjects study of Mahoney et al. (2011) by using a psychophysical data collection approach in experiment 1. That is, instead of presenting few trials to many participants as in a typical social science approach, here fewer participants perform many more trials. This method had successfully been used in other framing studies (Guo et al., 2017; Diederich et al., 2020; Roberts et al., 2021; Wyszynski and Diederich, 2022). Second, we include two different interpretations of the framing effect: a narrow interpretation, i.e., comparing the number of frame-inconsistent choices between participants; and a wide one, i.e., comparing the proportions of risky choices made by the participants in the two frames. Third, we include variables defining the choice problems as explanatory variables. Fourth, we seek to replicate the results of our first experiment using an online-sample to overcome a potential homogeneity issue of student samples. For the online experiment, we incorporate the psychophysical approach from experiment 1 into a social science approach requiring a larger sample size in favor of fewer trials per participant. The combined design has three advantages for our study: (1) a larger and more heterogeneous sample provides a broader range of psychometric cognitive-style and risk-style scores; (2) the correlations between frame-inconsistent choices and scores measured with psychometric instruments are expected to be more stable in larger samples (Schönbrodt and Perugini, 2013); and (3) due to the smaller number of trials, participants are less likely to drop out during the online session.

2. Experiment 1

The first experiment was done in a lab using a quasi psychophysical approach. Participants were asked to choose either the sure or the risky (gamble) option in a series of Unusual Disease Problems. Choice and response time data are based on Diederich et al. (2018) who investigated several determinants of risky decision making utilizing a sample of students receiving monetary compensation. Similar to Mahoney et al. (2011), the study used three different diseases embedded into two frames. Details on the number of affected people, probabilities, and response deadline variations are described in the following. For the current study, we elected scores on different psychometric instruments to examine the influence of cognitive-style and risk-style on choice behavior.

2.1. Materials

In addition to the framing manipulations, i.e., presenting each trial in a gain and a loss frame, Diederich et al. (2018) included four variables (characteristics) describing the choice problem: outcomes, probabilities of surviving/dying, problem domain, and time limits.

Outcomes: The outcomes of the decision were described as the number of people affected by a certain disease. Diederich et al. (2018) defined two major categories for the number of affected people, called *Scope* here. Category Small included the values 20, 40, 60, and 80. To minimize a possible impact of prominent numbers on risky choice, each value was flanked by ±1 resulting in four triplets of values (19, 20, 21; 39, 40, 41; 59, 60, 61; 79, 80, 81). For category Large, these numbers were multiplied by 100.

Probabilities: The probability indicated for a particular choice problem describes the affected peoples' chance of survival/death. Probabilities of surviving/dying varied on four levels. The particular values were 0.3, 0.4, 0.6, and 0.7.

Problem domain: The problem domain was varied by including three different versions of the Unusual Disease Problem (scenarios). For the control condition of the disease variable, the scenario described an outbreak of an unusual infectious disease (category: Infectious). The other two Unusual Disease Problem scenarios were about a new agent to treat leukemia (category: Leukemia) and a new agent to treat AIDS (category: AIDS). The full texts of the disease scenarios can be found in the Supplementary material.

Time limits: Two response deadlines were included. A short time limit of 1 s and a longer time limit of 3 s.

For a given Scope, the twelve numbers of affected people were paired with the probabilities to 48 combinations (12 × 4) per frame resulting in 96 individual test trials. The sure option for each trial was created to match the expected value of the gamble option. In addition, 24 catch trials (12 per frame) were constructed to assess accuracy and engagement in the task. The catch trials had two non-equivalent choice options. One option had a significantly larger expected value than the other option. For the catch trials, a probability of 0.9 (0.1) for risky options was paired with the expected value of the number of affected people multiplied by 0.1 (0.9). The sure option was preferable to the risky option for 12 catch trials (6 per frame), and vice versa for the other 12 catch trials (for details see Diederich et al., 2018). Altogether, 96 test trials plus 24 catch trials make 120 trials presented in one block. Furthermore, a block of trials was embedded in one disease category and one level of time limits.

2.2. Measures

We measured cognitive-styles with two different inventories. First, similarly to Mahoney et al. (2011), we used the 40-items Rational-Experiential Inventory (REI-40), with the rational-analytic (RA) and the experiential-intuitive (EX) sub-scales (Pacini and Epstein, 1999). RA thinking is equivalent to the concept of Need for Cognition (NFC). It is measured by an adapted version of the original NFC instrument (Cacioppo and Petty, 1982). EX thinking is basically equivalent to the Faith in Intuition (FI) concept (Epstein et al., 1996). Participants rated all items on a 5-point Likert scale that ranged from 1 ("definitely not true of myself") to 5 ("definitely true of myself"). We observed a reliability of RA and EX of $\alpha = 0.86$ and $\alpha = 0.84$, respectively.

Second, we used the 7-item short form of the Actively Open-Minded Thinking (AOT-7) scale as used in Haran et al. (2013), who investigated the role of AOT in the acquisition, accuracy, and calibration of information. Participants rated all items on a 7-point Likert scale from 1 ("completely disagree") to 7 ("completely agree"). In the current study, the reliability of the AOT scale was $\alpha = 0.7$.

We measured risk-styles with the Stimulating-Instrumental Risk Inventory (SIRI; Zaleskiewicz, 2001), which is composed of two sub-scales, the stimulating-risk sub-scale (ST) and the instrumental-risk sub-scale (IN). Participants have to self-assess their attitudes to 17 statements (10 ST, 7 IN) using a 5-point Likert scale from 1 ("does not describe me at all") to 5 ("describes me very well"). In the current study, the reliability was $\alpha = 0.74$ for the ST scale and $\alpha = 0.58$ for the IN scale.

The questionnaires, as they were used in this study, are found in the Supplementary material.

2.3. Design and procedure

The study had a mixed design. Three diseases and two levels of Scope were paired to six combinations. Each subject was exposed to two different diseases, one with Small and the other with Large Scope. The remaining factors were balanced within subjects. Each participant completed 480 trials in two sessions with two blocks of 120 trials, the first block of trials with a 3 s deadline and the second with a 1 s deadline. Note that within a given session, Disease and Scope conditions were the same. Participants had 5-min breaks between blocks and sessions.

The experimental trials started by showing the number of affected people for the corresponding trial. The subsequent screen showed the choice options (visualized by pie charts) and time limit for that particular trial. A response had to be made within the given time limit. The last screen provided feedback about the outcome of the choice. After offset of the screen, the next trial started (for screenshots and details see Supplementary material and Diederich et al., 2018). Participants filled the REI after the first session, the AOT before the second session, and the SIRI after the second session. Questions of each scale were presented in random order.

2.4. Data processing and statistical methods

For each instrument, we normalized the values recorded for the participants by subtracting the smallest measurable value of the instrument (I_{min}) from the value recorded for each participant (I_i) and divide the result by the highest measurable value of the instrument (I_{max}) minus I_{min}: $I_{norm} = \frac{I_i - Imin}{I_{max} - I_{min}}$.

We quantified the number of frame-inconsistent choices (FIC) of each participant by comparing the responses to gain-framed trials with those given to the identical counterpart in the loss frame. We counted a FIC when a participant's response to otherwise identical trials varied depending on the framing as gain or loss.

We first evaluated the data using descriptive statistics and Pearson correlations between the number of FIC and the normalized values of the instruments.

To analyze the effects of framing, choice problem characteristics, and individual differences on the proportion of choosing the gamble, we used generalized linear mixed models (GLMM; family: binomial, bound optimization: quadratic approximation) with random intercept variance across participants and sequence of stimuli presented (trial sequence). For the statistical analysis, we used the computing environment R (version

4.0.3; packages: "lme4," "descr," "Hmisc," "psych," "simr"; Bates et al., 2014; Green and MacLeod, 2016; Aquino, 2018; R Core Team, 2018; Revelle, 2020; Harrell, 2021).[2]

All models included the relative frequency of choosing the risky option as the dependent variable. Frame (Loss; Gain), Scope of affected people, with categories Small (basic values: 20, 40, 60, 80) and Large (100 times the Small values), Probabilities of surviving/dying (<0.5; >0.5), Disease (Infectious disease; Leukemia; AIDS), and Time (1 s; 3 s limit) were included as explanatory variables. The first categories served as references. Since the scores of some of the instruments are expected to be highly correlated with each other, a model including all instruments would be affected by the problem of multicollinearity. Therefore, we executed the model separately for each of the five instruments (main effects models), i.e., the sub-scales of the REI (RA and EX), the AOT, and the sub-scales of the SIRI (ST and IN). Furthermore, to investigate the relationship between a person's test score and the impact of the explanatory variables on risky choice, we included two-way interactions of the instrument scores by each explanatory variable in the models (interaction models).

A *post-hoc* sensitivity analysis indicating the smallest detectable effect sizes (using the R package "simr"; Green and MacLeod, 2016) is shown the Supplementary material (Supplementary Tables S1–S3).

2.5. Participants

Fifty-five undergraduates (26 female, 29 male) of Jacobs University Bremen participated in two experiment sessions (age: 18–26 years; median = 20; English speakers). Altogether, each participant performed 480 trials (384 test trials; 96 catch trials). The experiment lasted for about 90 min. See Diederich et al. (2018) for details.

2.6. Results

Of the 55 participants, 12 (7 females) have been excluded due to an unusually high number of catch trial failures (14 inferior responses in one block). Of the remaining 16,512 test trials (43 × 386), 80 trials were timeouts and were also removed from the data set. Thus, the following analysis is based on a total of 16,432 trials. In 51.1% of valid trials, the risky option was chosen. Overall, participants chose the risky option more often in loss trials (60.1%) than in gain trials (39.9%), indicating a framing effect (for details see Diederich et al., 2018). Probabilities and Scope had an impact on choice behavior: (1) The larger the probability of surviving/dying in the scenario was the higher the proportion of the risky choice option, and (2) the fewer people were affected (Scope: Small), the higher the proportion of the risky choice option (for details see Diederich et al., 2018).

TABLE 1 Experiment 1: correlations between FIC and scores of risk-style (stimulating and instrumental risk) and cognitive-style (rational thinking, experiential thinking, and actively open-minded thinking style).

	FIC	ST	IN	RA	EX
ST	0.24				
p	*0.121*				
IN	0.16	**0.47**			
p	*0.315*	*0.002*			
RA	0.07	**0.43**	0.27		
p	*0.646*	*0.004*	*0.075*		
EX	−0.15	−0.27	0.02	−0.13	
p	*0.327*	*0.074*	*0.884*	*0.398*	
AOT	−0.19	**−0.31**	**−0.42**	−0.18	0.09
p	*0.214*	*0.046*	*0.005*	*0.257*	*0.552*

FIC, frame-inconsistent choice; ST, stimulating risk; IN, instrumental risk; RA, rational thinking; EX, experiential thinking; AOT, actively open-minded thinking. Statistically significant correlations ($p < 0.05$) are printed in bold and *p*-values are italicized.

2.6.1. Individual differences in frame-inconsistent choices

The number of frame-inconsistent choices (FIC) ranged from 8 to 64 (overall: mean = 43.2, SD = 15.9) among the participants. That is, the average proportion of FIC was 67%. Note that the FIC proportions varied between the conditions of time limit (72% for 1 s, and 62% for 3 s time limit). The individual scores measured using the instruments varied across a moderate range. Details and normalized scores are found in the Supplementary material (Supplementary Table S4). We observed statistically significant correlations between scores of the following scales: ST and IN, ST and RA, ST and AOT, and IN and AOT. However, none of the instruments correlated significantly with FIC (see Table 1).

2.6.2. Individual differences in choice proportions

The main effects GLMM analyses (see Supplementary Tables S5–S9) showed no significant relationship between the scores measured using the instruments and proportions of choosing the gamble option. However, we found significant effects for Frame, Scope, and Probabilities but not for Disease and Time in each main effects model.

In the following, we report the results of the interaction effects GLMM analyses, separate for each instrument. Note that we interpret interactions even if the main effects were not significant. It is well possible that effects have canceled out due to the specific response behavior of participants with different risk-styles and cognitive-styles.

2.6.2.1. Rational-experiential thinking

Table 2 shows the results of the GLMM analyses and Figure 1 illustrates significant interaction effects. We interpret the findings as follows:

[2] Note that previous research often analyzed data with an ANOVA approach (e.g., Shiloh et al., 2002; LeBoeuf and Shafir, 2003; Mahoney et al., 2011). We use GLMMs since they have been shown to be more flexible, accurate, powerful, and suited for (categorical) data analysis (Kristensen and Hansen, 2004; Jaeger, 2008; Yu et al., 2022).

TABLE 2 Experiment 1: Generalized linear mixed models, Interactions: rational and experiential thinking-style.

Rational thinking-style				
Fixed effects:	**Est.**	**SE**	***z*-value**	***p*-value**
(Intercept)	0.215	0.552	0.389	0.697
RA	−1.609	1.056	−1.523	0.128
Frame (Gain)	−1.160	0.151	−7.705	<0.001
Scope (Large)	−0.611	0.168	−3.637	<0.001
Prob. (>0.5)	1.740	0.155	11.219	<0.001
Leukemia	−1.018	0.246	−4.132	<0.001
AIDS	−0.530	0.216	−2.455	0.014
Time (3 s)	−0.025	0.141	−0.179	0.858
RA × Frame	−0.395	0.294	−1.342	0.180
RA × Scope	1.117	0.316	3.537	<0.001
RA × Prob.	2.084	0.302	6.890	<0.001
RA × Leukemia	1.949	0.456	4.275	<0.001
RA × AIDS	0.895	0.420	2.131	0.033
RA × Time	0.160	0.273	0.586	0.558
Random effects:		**SD (Est.)**		
Trial seq. (Intercept)		0.024		
Subject (Intercept)		0.967		
Experiential thinking-style				
Fixed effects:	**Est.**	**SE**	***z*-value**	***p*-value**
(Intercept)	−1.747	0.932	−1.873	0.061
EX	1.697	1.300	1.306	0.192
Frame (Gain)	−1.427	0.253	−5.649	<0.001
Scope (Large)	0.589	0.272	2.168	0.030
Prob. (> 0.5)	3.670	0.260	14.114	<0.001
Leukemia	−0.001	0.370	−0.001	0.999
AIDS	1.017	0.325	3.128	0.002
Time(3 s)	0.077	0.237	0.324	0.746
EX × Frame	0.112	0.351	0.319	0.749
EX × Scope	−1.002	0.381	−2.629	0.009
EX × Prob.	−1.275	0.362	−3.519	<0.001
EX × Leukemia	−0.107	0.511	−0.210	0.834
EX × AIDS	−1.595	0.455	−3.508	<0.001
EX × Time	−0.033	0.330	−0.099	0.921
Random effects:		**SD (Est.)**		
Trial seq. (Intercept)		0.024		
Subject (Intercept)		0.967		

16,432 observations from $n = 43$ participants indicated in a series of 120 trials per block.

Frame: The GLMM analyses revealed no significant interaction effects of RA by Frame and EX by Frame. That is, the framing effect, i.e., divergence in proportions of choosing the gamble between gain and loss frames, was not moderated by rational or experiential thinking-style.

Scope: The effect of Scope on choosing the gamble was moderated by RA and EX. In particular, the proportions of gambling were lower for Scope Large than for Scope Small for individuals with lower RA scores (about <0.5). However, it increased with RA scores for Scope Large but not for Scope Small. The GLMM analysis of the EX scores showed that gambling increased with EX scores for Scope Small but not for Scope Large. That is, the effect of Scope reverses with increasing RA scores, and it becomes stronger with increasing EX scores.

Probabilities: Both RA and EX moderated the effect of Probabilities on choosing the gamble option. Participants chose the gamble more often for Probabilities >0.5, and they chose the sure option more often for Probabilities <0.5. Gambling proportions decreased with RA scores and increased with EX scores for Probabilities <0.5, and they increased with RA scores and decreased with EX scores for Probabilities >0.5. That is, the effect of Probabilities is getting stronger with increasing RA scores and it becomes weaker with increasing EX scores.

Disease: The effect of Disease types on the proportion of choosing the gamble option varied for individuals with different RA and EX scores, respectively. For Infectious disease, gambling increased with increasing RA and EX scores. For AIDS, however, it decreased with increasing RA and EX scores. Moreover, the GLMM revealed that gambling increased even stronger with RA scores for Leukemia than for Infectious disease.

Time: No significant interaction effects of RA by Time and EX by Time were observed.

2.6.2.2. Actively open-minded thinking

Table 3 shows the interaction results when including AOT scores in the GLMM analysis. Figure 2 illustrates significant interactions.

Frame: The GLMM showed no interaction effect of AOT by Frame. That is, AOT did not serve as a moderator of the framing effect in the current study.

Scope: Participants chose the gamble option more often for Scope Small than for Scope Large (main effect; see Supplementary Table S7). For Scope Small, the proportion of choosing the gamble decreased with increasing AOT scores, and for Scope Large it increased with increasing AOT scores. That is, the effect strength of Scope becomes smaller with increasing AOT scores.

Probabilities: For Probabilities >0.5, the risky option was chosen more often, whereas for Probabilities <0.5, the sure option was chosen more often. As for the other instruments, we found a significant interaction effect of AOT by Probabilities: Proportions of choosing the gamble decreased with increasing AOT scores for Probabilities <0.5, and they increased with AOT scores for Probabilities >0.5. As observed for rational thinking-style, the effect of Probabilities is getting stronger with increasing AOT scores.

Disease: Participants with normalized AOT scores around 0.5. and 0.6, which are lower AOT scores measured in the sample used for this study, chose the sure option more often for Leukemia and the gamble more often for the Infectious disease. The gambling frequency increased with AOT scores for Leukemia, and it decreased with increasing AOT scores for the Infections Disease.

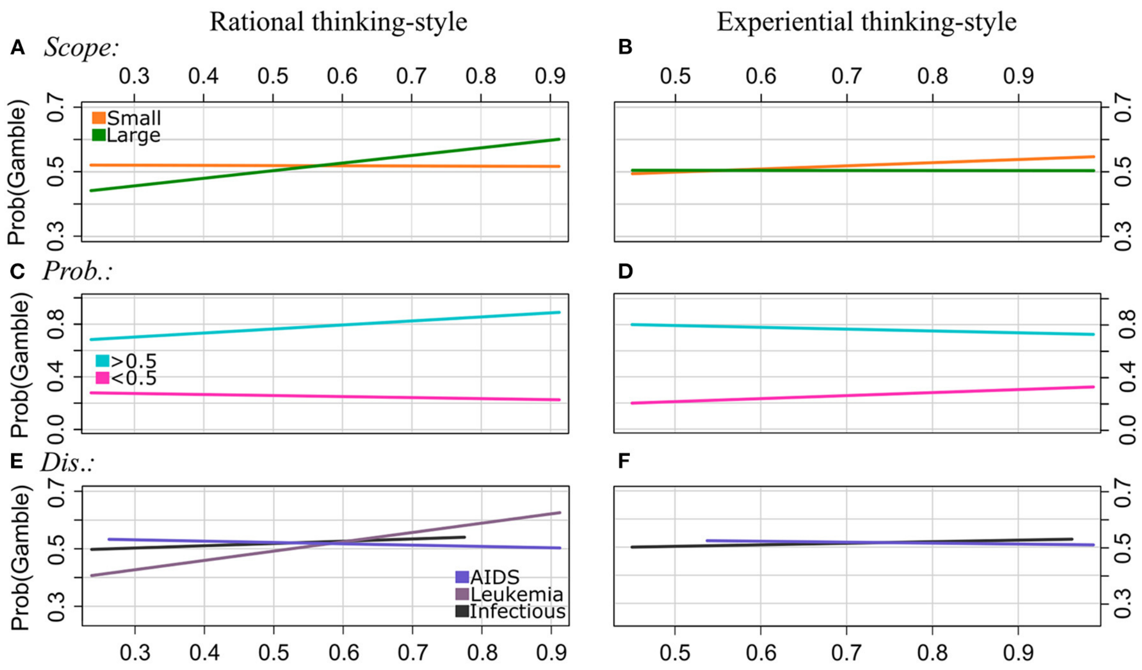


FIGURE 1
Experiment 1: Regression lines of the proportions of choosing the gamble option as a function of rational thinking-style (left column) and experiential thinking-style (right column), separately for the levels of Scope [patterns **(A, B)**], Probabilities **(C, D)**, and Disease **(E, F)**. Note that we applied a smaller range of values on the y-axis for the plots of the patterns **(A, B, D, E)** to illustrate the interaction effect more clearly.

TABLE 3 Experiment 1: Generalized linear mixed model, Interactions: actively open-minded thinking-style.

Fixed effects:	Est.	SE	*z*-value	*p*-value
(Intercept)	1.000	1.135	0.881	0.378
AOT	−2.213	1.531	−1.445	0.148
Frame (Gain)	−0.922	0.282	−3.274	0.001
Scope (Large)	−0.935	0.383	−2.440	0.015
Prob. (>0.5)	−0.754	0.288	−2.619	0.009
Leukemia	1.154	0.555	2.078	0.038
AIDS	−0.351	0.388	−0.906	0.365
Time (3 s)	0.092	0.268	0.341	0.733
AOT × Frame	−0.603	0.387	−1.559	0.119
AOT × Scope	1.200	0.521	2.302	0.021
AOT × Prob.	4.860	0.398	12.218	<0.001
AOT × Leukemia	−1.643	0.748	−2.196	0.028
AOT × AIDS	0.483	0.522	0.926	0.355
AOT × Time	−0.052	0.367	−0.143	0.887
Random effects:	**SD (Est.)**			
Trial seq. (Intercept)	<0.001			
Subject (Intercept)	1.014			

16,432 observations from $n = 43$ participants indicated in a series of 120 trials per block.

Time: There were no significant interactions between AOT and Time.

2.6.2.3. Stimulating-instrumental risk-style

Results of the GLMM analyses are shown in Table 4. Figure 3 displays significant interaction effects. We interpret the findings as follows:

Frame: The strength of the framing effect increases with scores of both stimulating risk-style and instrumental risk-style. In particular, the proportion of choosing the gamble option increases with ST scores in gain and loss frames. In loss frames, however, gambling proportions increase more strongly with ST scores than in gain frames. For the IN scale, there is a tendency of decreasing gambling proportions with increasing IN scores in the gain frame and increasing gambling proportions with increasing IN scores in the loss frame.

Scope: ST moderated the effect of Scope on risky choices. Overall, participants were less likely to choose the risky option for the Large Scope category than for Scope Small. However, the strength of the effect of Scope becomes smaller with increasing ST scores, and it even reverses at higher ST scores. The interaction effect of IN by Scope was not significant.

Probabilities: The gamble option was chosen more often for Probabilities >0.5, whereas, for Probabilities <0.5, the sure option was chosen more often. In the latter category (<0.5), the proportion of choosing the gamble increased with ST and IN scores. For high probabilities (>0.5), however, gambling frequency increased weaker or was relatively stable across participants with different scores of ST and IN, respectively. That is, the strength of the effect of Probabilities on choice behavior decreased with increasing scores of stimulating and instrumental risk-style.

Disease type: The GLMM analysis revealed a statistically significant interaction effect between the disease type "AIDS" and both risk-styles (ST and IN, respectively). For the reference

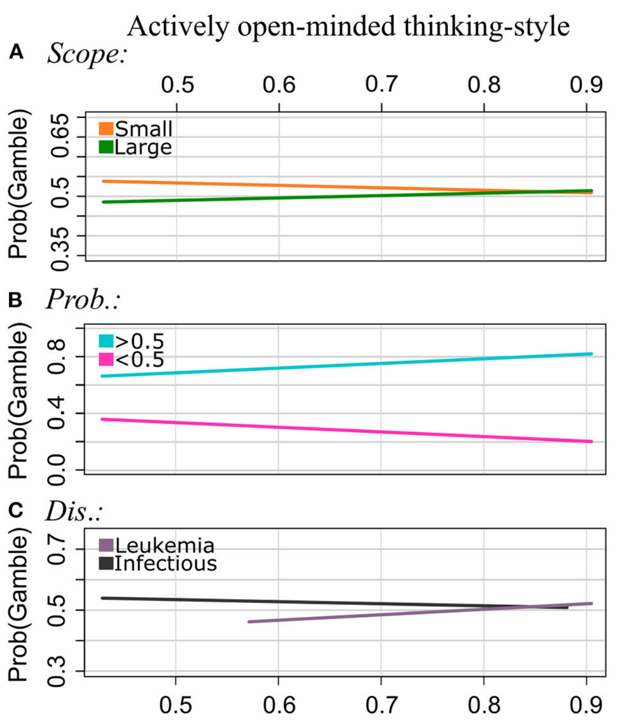


FIGURE 2
Experiment 1: Regression lines of the proportion of choosing the gamble option as a function of AOT, separately for the levels of Scope [pattern **(A)**], Probabilities **(B)**, and Disease **(C)**. Note that we applied a smaller range of values on the y-axis for the plots of the patterns **(A–C)** to illustrate the interaction effect more clearly.

TABLE 4 Experiment 1: Generalized linear mixed models, Interactions: stimulating and instrumental risk-style.

Stimulating risk-style				
Fixed effects:	Est.	SE	*z*-value	*p*-value
(Intercept)	−1.330	0.431	−3.087	0.002
ST	1.985	0.998	1.990	0.047
Frame (Gain)	−0.824	0.120	−6.885	<0.001
Scope (Large)	−0.814	0.134	−6.059	<0.001
Prob. (>0.5)	3.581	0.127	28.180	<0.001
Leukemia	−0.285	0.176	−1.619	0.105
AIDS	−0.636	0.161	−3.964	<0.001
Time (3 s)	0.021	0.114	0.185	0.853
ST × Frame	−1.285	0.274	−4.693	<0.001
ST × Scope	1.664	0.301	5.535	<0.001
ST × Prob.	−1.946	0.287	−6.777	<0.001
ST × Leukemia	0.497	0.409	1.213	0.225
ST × AIDS	1.175	0.352	3.339	<0.001
ST × Time	0.081	0.259	0.314	0.753
Random effects:		**SD (Est.)**		
Trial seq. (Intercept)		0.033		
Subject (Intercept)		0.979		
Instrumental risk-style				
Fixed effects:	Est.	SE	*z*-value	*p*-value
(Intercept)	−1.865	0.906	−2.058	0.040
IN	1.981	1.363	1.454	0.146
Frame (Gain)	0.412	0.246	1.676	0.094
Scope (Large)	−0.483	0.262	−1.844	0.065
Prob. (>0.5)	4.129	0.256	16.126	<0.001
Leukemia	−0.050	0.313	−0.161	0.872
AIDS	−0.847	0.375	−2.262	0.024
Time (3 s)	−0.332	0.234	−1.419	0.156
IN × Frame	−2.654	0.368	−7.214	<0.001
IN × Scope	0.564	0.390	1.448	0.148
IN × Prob.	−2.021	0.380	−5.317	<0.001
IN × Leukemia	−0.073	0.472	−0.155	0.877
IN × AIDS	1.118	0.548	2.037	0.042
IN × Time	0.583	0.348	1.676	0.094
Random effects:		**SD (Est.)**		
Trial seq. (Intercept)		0.026		
Subject (Intercept)		1.002		

16,432 observations from $n = 43$ participants indicated in a series of 120 trials per block.

category, i.e., “Infectious” disease, the proportion of choosing the gamble option tends to increase with increasing scores of ST and IN, respectively. However, for the “AIDS” disease category, gambling frequency was relatively stable across participants with different ST scores, and it slightly decreased with increasing IN scores.

Time: There were no significant interactions of ST by Time and IN by Time.

2.7. Summary and discussion

To investigate individual differences in susceptibility to risky choice framing, we used a psychophysical data collection approach and five different scales for measuring individual differences in cognitive-style and risk-style. We included two different interpretations of the framing effect. A narrow one, i.e., we compared the number of frame-inconsistent choices each participant made, and a wide one, i.e., we compared the proportions of choosing the gamble as a function of framing and other variables defining the choice problems. Overall, we found a high average proportion of frame-inconsistent choices (67%) and a strong effect of framing on the proportion of choosing the gamble (Gain: 40%; Loss: 60%).

The number of frame-inconsistent choices did not significantly correlate with the scores of any psychometric instrument used to measure cognitive-styles and risk-styles in the current study. Our findings are consistent with the results of Mandel and Kapler (2018), who investigated the impact of cognitive-styles (AOT

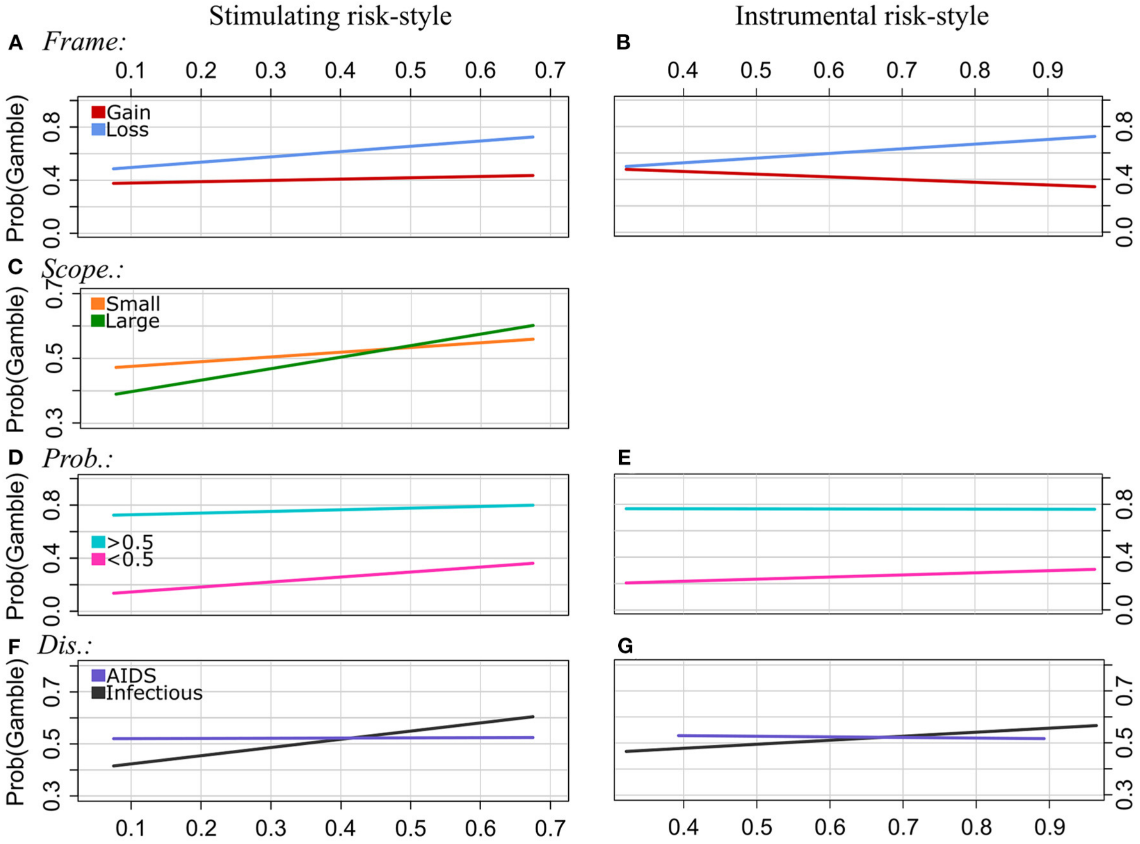


FIGURE 3
Experiment 1: Regression lines of the proportions of choosing the gamble option as a function of stimulating risk-style (left column) and instrumental risk-style (right column), separately for the levels of Frame [pattern **(A, B)**], Scope [**(C)**, effect of EX by Scope was n.s.], Probabilities **(D, E)**, and Disease **(F, G)**. Note that we applied a smaller range of values on the y-axis for the plots of the patterns **(C, F, G)** to illustrate the interaction effect more clearly.

and NFC) on the susceptibility to framing effects applying a narrow interpretation of frame-(in)consistent choices. In their between-subjects experiment, they counted a "frame-consistent choice" when a participant chose the sure option in the positive frame condition or the risky option in the negative frame condition. Neither AOT nor NFC correlated significantly with the number of frame-consistent choices.[3] However, other studies showed statistically significant ($p < 0.05$) correlations between a measure of frame-(in)consistent choices and cognitive-styles (i.e., AOT, NFC; see e.g., Björklund and Bäckström, 2008; West et al., 2008; Peng et al., 2019; Erceg et al., 2022; Rachev et al., 2022), contradicting our results. According to the classification by Cohen (1988), these correlations are small to moderate. Furthermore, previous research on individual differences in framing susceptibility has paid most attention to measures of cognitive-style (Mandel and Kapler, 2018), and the relationship between risk-style and a narrow interpretation of the framing effect such as a measure of frame-(in)consistent choices has not been investigated so far.

For the wide framing effect interpretation, we found no impact of cognitive-style on framing effect strength which supports the majority of previous research applying a similar wide framing effect interpretation (Levin et al., 2002; Shiloh et al., 2002; LeBoeuf and Shafir, 2003; Björklund and Bäckström, 2008; Mahoney et al., 2011; Stark et al., 2017). Note that a few studies found a relationship between cognitive-style and framing effect strength in the wide interpretation. In particular, LeBoeuf and Shafir (2003), study 2 and Simon et al. (2004) found weaker framing effects for individuals with higher NFC scores.

In the current study, only stimulating risk-style and instrumental risk-style moderated the effect of framing on proportions of choosing the gamble. As expected, the framing effect becomes stronger as the scores of stimulating-risk increase. However, we observed the same pattern for instrumental risk-style, which is the opposite relationship than expected. High instrumental risk-style is theoretically associated with more deliberative risk-taking and, therefore, lower susceptibility to cognitive biases such as the framing effect (Zaleskiewicz, 2001). Mahoney et al. (2011), who also measured risk-style based on the intuitive-deliberative processing approach using the SIRI, found no significant moderator effects of stimulating risk-style and instrumental risk-style on framing effect strength. Note that the relationship between risk-style and framing effect strength has been investigated by previous studies using other

[3] Mandel and Kapler (2018) also measured numeracy and cognitive-ability in their study. They found, that only cognitive-ability (i.e., the Cognitive Reflection Task; Frederick, 2005) correlated with the number frame consistent choices.

concepts of risk-style (e.g., group polarization, risk-avoidance). The findings here are mixed (see Mahoney et al., 2011, for a review).

Moreover, we found that the effect of different outcomes (numbers of affected people; called Scope here) on risky choice behavior was moderated by rational thinking, experiential thinking, actively open-minded thinking, and stimulating risk-style. In line with the theory, we observed the effect of Scope to become stronger with increasing scores of experiential thinking-style, and it becomes weaker with increasing AOT scores. However, the other significant moderator effects were inconsistent with the basic assumptions of the scales. In particular, the effect of Scope reverses with increasing scores of rational thinking-style and stimulating risk-style. That is, individuals with lower scores chose the gamble less often, and those with higher scores chose the gamble more often for the large Scope than for the small one.

Each scale moderated the effects of probabilities of surviving/dying on choice behavior. Contrary to the theoretical implications, the effect of probabilities, i.e., selecting the sure option more often for probabilities <0.5 and the gamble more often for probabilities >0.5, becomes stronger with increasing scores of rational thinking-style and actively open-minded thinking-style, and it becomes weaker with increasing scores of experiential thinking-style. For the risk-style measures, we observed that the effect of probabilities becomes weaker with increasing scores. That is, only instrumental-risk moderated the effect of probabilities as expected.

The problem domain (different disease problems) was also moderated by each scale. In line with the theory, the effect of different disease problems on risky choice was found to become weaker with increasing AOT scores. However, the other measures of cognitive-style and risk-style moderated the effect of Disease in a different way than theoretically predicted. In particular, for the rational thinking-style, we expected that differences in the proportions of gambling between the three diseases will become smaller with increasing scores. For the experiential thinking-style, we expected to observe the opposite (i.e., differences in gambling proportion become larger with increasing scores). However, we observed that the frequency of choosing the gamble was lowest for Leukemia, higher for the Infectious disease, and highest for AIDS among individuals with the lowest scores of rational thinking-style. We observed the reversed order among individuals with the highest scores of rational thinking. Similarly, gambling proportions were lower for Leukemia than for AIDS among low experiential thinkers and lower for AIDS than for Leukemia among high experiential thinkers. A similar pattern emerged for the risk-style measures: For AIDS, the proportions of choosing the gamble was about the same (about 50%) for individuals with different scores of stimulating and instrumental risk-style. However, for the Infectious disease, individuals with low scores chose the sure option more often, and those with high scores chose the risky option more often.

We found no relationship between Time limits and any of the psychometric measures included in the current study.

Note that the current investigation is based on a reanalysis of existing data. However, the original study was not designed to measure correlations between frame-inconsistent choices and individual differences. Although the sensitivity analysis revealed a strong statistical power for the GLMM analyses, correlations require a much higher sample size to be stable (Schönbrodt and Perugini, 2013). Moreover, the lowest normalized scores of EX and AOT measured in the current study are 0.45, and 0.43 indicating a lack of participants with low and very low scores of these instruments. The small range of scores may be due to the sample size, which was determined for the analysis of particular effects in the original study, and the homogeneity of the sample. Using a student sample may result in a smaller range of scores for particular psychometric measures.

3. Experiment 2

Experiment 2 somewhat combines the social and psychophysical data collection approach. The experiment was conducted online. The composition of online samples might be more heterogeneous (e.g., age, education, profession) as compared to a student sample. We further increased the sample size to stabilize the correlations between frame-inconsistent choices and risk-style and cognitive-style, respectively. The general setup of the experiment was similar to the first one with a few exceptions described in the following. The study was conducted using Amazon MTurk and the online survey software EFS from TIVIAN. The statistical methods are the same as before.

3.1. Materials

We used the same three disease problem scenarios as in experiment 1 (unusual infectious disease, leukemia, and AIDS). As before, the Scope categories were Small and Large, with Small including only the values 20, 40, 60, and 80; for condition Scope Large, these numbers are multiplied by 100. The probabilities of surviving/dying were 0.3, 0.4, 0.6, and 0.7. For a given Scope, the 16 combinations (4 values × 4 probabilities) were framed as gains and losses, resulting in 32 test trials. In addition, 8 catch trials (4 per frame) were constructed to assess accuracy and engagement in the task. In four of the eight catch trials, participants were required to choose the sure option that offers a 100% chance to save all affected people. The risky option, however, involved a probability of 0.3 to save all people (no one will be saved with a probability of 0.7). In the other four catch trials, participants were required to choose the risky option that involved a probability of 0.7 to save all affected people. The sure option offered a 100% chance that no one will be saved. Based on pretesting, we allowed the participants to make two catch trial failures. The third catch trial failure led to the termination of the experiment.

One experimental block consisted of one of the three disease problem scenarios with 32 test trials and 8 catch trials. Different from experiment 1, we did not include different deadlines.

As in experiment 1, we used the same three measures (SIRI, REI, AOT). However, we replaced the 40-items REI with the shorter 10-items REI-short (Epstein et al., 1996). The reliability of the scales was $\alpha = 0.83$ for the ST scale, $\alpha = 0.76$ for the IN scale, $\alpha = 0.71$ for the RA scale, $\alpha = 0.84$ for the EX scale, and $\alpha = 0.81$ for the AOT scale. Furthermore, we added attention checks to each scale (one to the AOT scale, and three to the REI and SIRI scale, respectively) where participants were asked to give a particular rating (e.g., "please rate this item with '4'"). An attention check failure terminated the experiment.

3.2. Design and procedure

The design was similar to the one used in experiment 1, however, with fewer trials per participant. Instead of completing 480 trials, each participant completed 80 trials in two blocks of 40 trials. Disease and Scope combinations varied between the two blocks. All trials had a response deadline of 5 s. Responses that were too slow (timeouts) were recorded as missing values and had no consequences for the participant. Timeouts in catch trails, however, were recorded as catch trial failures.

Upon inclusion in the study, participants first received basic information about the study. They were then introduced in the experimental procedure. The task was explained using two example trials (one per frame) with the components (e.g., choice options) labeled with explanatory comments. After participants remained for at least 60 s on the explanation page, they performed five practice trials. The first four practice trials included comments explaining the display. The first two practice trials had no response deadline. In practice trials 3 and 4, participants had to respond within the 5 s deadline. In case of a timeout, they were asked to repeat the corresponding practice trial. Practice trial 5 demonstrated how a test trial is displayed (i.e., explanatory comments disappeared).

Each block started with displaying the respective disease problem scenario. The procedure of the experimental trials and the display were similar to those in experiment 1 with the following modifications: (1) The screen displaying the number of affected patients was presented for 2 s (instead of 2.5 s). (2) The choice options were additionally labeled according to the frames ("patients survive" or "patients die," respectively). There were no labels in experiment 1 (framing was only indicated by different gray shades). (3) The remaining time for a trial was indicated by a clock (instead of bars) counting down the seconds starting from 5 (screenshots and details can be found in the Supplementary material). (4) Participants were asked to use a standard computer mouse or an comparable input device for indicating their choice (instead of the left and right arrow-key of the keyboard).

Participants completed the AOT after the first block, the SIRI and the REI after the second block. Finally, they were asked for their age and gender. On the final page, participants received an individual, randomly generated code required to get the participation fee from MTurk.

3.3. Participants

We determined the sample size to match the valid observations in experiment 1 (384 test trials of 43 participants results in 16,512 test trials). The online experiment includes 64 test trials, which then require 258 participants. We requested participants on Amazon MTurk. They received a hyperlink that directed to the online experiment.

TABLE 5 Experiment 2: Main effects model.

Fixed effects:	Est.	SE	z-value	p-value
(Intercept)	−0.867	0.158	−5.498	<0.001
Frame (Gain)	−0.727	0.040	−18.212	<0.001
Scope (Large)	−0.025	0.040	−0.630	0.529
Prob. (>0.5)	0.900	0.056	22.454	<0.001
Leukemia	−0.065	0.056	−0.1.155	0.248
AIDS	−0.069	0.057	−1.210	0.226
Random effects:	**SD (Est.)**			
Subjects (Intercept)	2.377			
Trial seq. (Intercept)	0.027			

16,592 observations provided by $n = 262$ participants in a series of 40 trials per block.

The online experiment was open for participation on Amazon MTurk from August 17th to 19th, and on August 24th 2021. MTurk workers were not required to meet any additional qualifications to participate (i.e., minimum HIT approval rate, language, location). On the fourth day, 1,327 workers accepted the HIT (human intelligence task) for participation. In total, 262 (117 female, 141 male, 4 preferred not to say) participants completed the experiment. The mean age was 32.73 years (median: 30, range: 20–64, SD: 9.64, $n = 1$ preferred not to say). Participants gave their informed consent prior to their inclusion in the study. The average completion time was 20 min and 27 s. Participants were paid a fixed amount of $4.70.

3.4. Results

Of the 1,327 individuals who accepted the HIT on MTurk for participation, 1,065 dropped out at the first pages showing the instructions, gave an incorrect response to an attention testing scale item, or failed more than two catch trials (see Supplementary material for an exploratory analysis of reasons for exclusion). We included data from the remaining 262 participants who finished the experiment. Of the 16,768 (262 × 64) test trials, 176 were timeouts and treated as missing values. Thus, the analysis is based on 16,592 trials. In 40.1% of the trials, the risky option was chosen. Overall, participants chose the risky option more often in loss trials (45.8%) than in gain trials (34.5%), indicating a framing effect. Furthermore, participants chose the risky option more often when Probabilities were large (>0.5: 47.1%) compared to small (<0.5: 33.1%). We found no effect of Scope and Disease on risky choice (see Table 5).

3.4.1. Individual differences in frame-inconsistent choices

Participants made between 0 and 32 frame-inconsistent choices (FIC) with a mean of 11.27 (SD = 9.99). That is, the mean proportion of FIC was 35%. Scores measured by the psychometric instruments varied across a wide range. Details

TABLE 6 Experiment 2: correlations between FIC and values of risk-style (stimulating and instrumental risk) and cognitive-style (rational thinking, experiential thinking, and actively open-minded thinking style).

	FIC	ST	IN	RA	EX
ST	0.08				
p	*0.225*				
IN	0.09	**0.69**			
p	*0.131*	*<0.001*			
RA	0.01	**−0.20**	−0.05		
p	*0.874*	*<0.001*	*0.380*		
EX	0.05	**0.51**	**0.52**	**−0.24**	
p	*0.367*	*<0.001*	*<0.001*	*<0.001*	
AOT	−0.10	**−0.61**	**−0.40**	**0.52**	**−0.53**
p	*0.1210*	*<0.001*	*<0.001*	*<0.001*	*<0.001*

FIC, frame-inconsistent choice; ST, stimulating risk; IN, instrumental risk; RA, rational thinking; EX, experiential thinking; AOT, actively open-minded thinking. Statistically significant correlations ($p < 0.05$) are printed in bold and *p*-values are italicized.

and normalized values are found in the Supplementary material (Supplementary Table S4). We found no significant correlations between the instruments and the number of frame-inconsistent choices (see Table 6). However, we found a high number of significant correlations between the scores of the instruments: ST correlated positively with IN and EX, and it correlated negatively with RA and AOT. IN correlated positively with EX, and negatively with AOT. RA correlated positively with AOT, and negatively with EX. EX correlated negatively with AOT.

3.4.2. Individual differences in choice proportions

We found no impact of cognitive-styles and risk-styles on risky choices. The main effects models showed significant effects of Frame and Probabilities on the proportion of preferring the gamble over the sure option (see Supplementary Tables S10–S14). As before, we show all interactions, separate for each instrument. Note that we interpret significant interaction effects even when the main effects were not significant. It is well possible that effects have been canceled out due to the specific response behavior depending on individual cognitive-style or risk-style.

3.4.2.1. Rational-experiential thinking

The interaction effect analysis of rational and experiential thinking-styles revealed significant effects for EX by Frame, RA and EX by Scope, RA and EX by Probabilities, and RA by Disease. Table 7 shows the results and Figure 4 illustrates significant interaction effects. We interpret the results as follows:

TABLE 7 Experiment 2: Generalized linear mixed models, Interactions: rational and experiential thinking-style.

Rational thinking-style				
Fixed effects:	Est.	SE	*z*-value	*p*-value
(Intercept)	−1.046	0.507	−2.065	0.039
RA	0.336	0.843	0.398	0.691
Frame (Gain)	−0.695	0.130	−5.366	<0.001
Scope (Large)	−0.389	0.131	−2.970	0.003
Prob. (>0.5)	0.498	0.130	3.827	<0.001
Leukemia	−0.128	0.182	−0.702	0.483
AIDS	−0.774	0.188	−4.109	<0.001
RA × Frame	−0.059	0.214	−0.277	0.782
RA × Scope	0.647	0.217	2.976	0.003
RA × Prob.	0.699	0.216	3.240	0.001
RA × Leukemia	0.089	0.296	0.301	0.763
RA × AIDS	1.246	0.311	4.005	<0.001
Random effects:		**SD (Est.)**		
Subject (Intercept)		2.354		
Trial seq. (Intercept)		0.029		
Experiential thinking-style				
Fixed effects:	Est.	SE	*z*-value	*p*-value
(Intercept)	−0.986	0.556	−1.774	0.076
EX	0.183	0.863	0.212	0.832
Frame (Gain)	−1.561	0.143	−10.919	<0.001
Scope (Large)	0.099	0.141	0.702	0.483
Prob. (>0.5)	1.635	0.144	11.381	<0.001
Leukemia	−0.002	0.196	−0.013	0.990
AIDS	0.144	0.209	0.691	0.489
EX × Frame	1.340	0.220	6.104	<0.001
EX × Scope	−0.202	0.217	−0.933	0.351
EX × Prob.	−1.177	0.221	−5.339	<0.001
EX × Leukemia	−0.096	0.297	−0.323	0.747
EX × AIDS	−0.345	0.326	−1.057	0.291
Random effects:		**SD (Est.)**		
Subject (Intercept)		2.379		
Trial seq. (Intercept)		0.021		

16,592 observations provided by $n = 262$ participants in a series of 40 trials per block.

Frame: We found a significant interaction effect between EX and Frame. The frequency of choosing the risky option decreased with increasing EX scores in loss frames, and it increased with EX scores in gain frames. That is, the strength of the framing effect decreases with increasing EX scores. The interaction of RA by Frame was not statistically significant. Scope: The effect of Scope on choosing the gamble option was moderated by RA. In particular, gambling increased with higher RA scores in both categories of Scope. However, it increased stronger for Large than for Small Scope. EX did not moderate the effect of Scope.

Probabilities: Both RA and EX moderated the effect of Probabilities on choosing the risky option. Gambling proportions increased with RA scores and EX scores for Probabilities <0.5, and they increased even stronger with RA scores and decreased with EX scores for Probabilities >0.5. That is, the effect of Probabilities is getting stronger with increasing RA scores, and it becomes weaker with increasing EX scores.

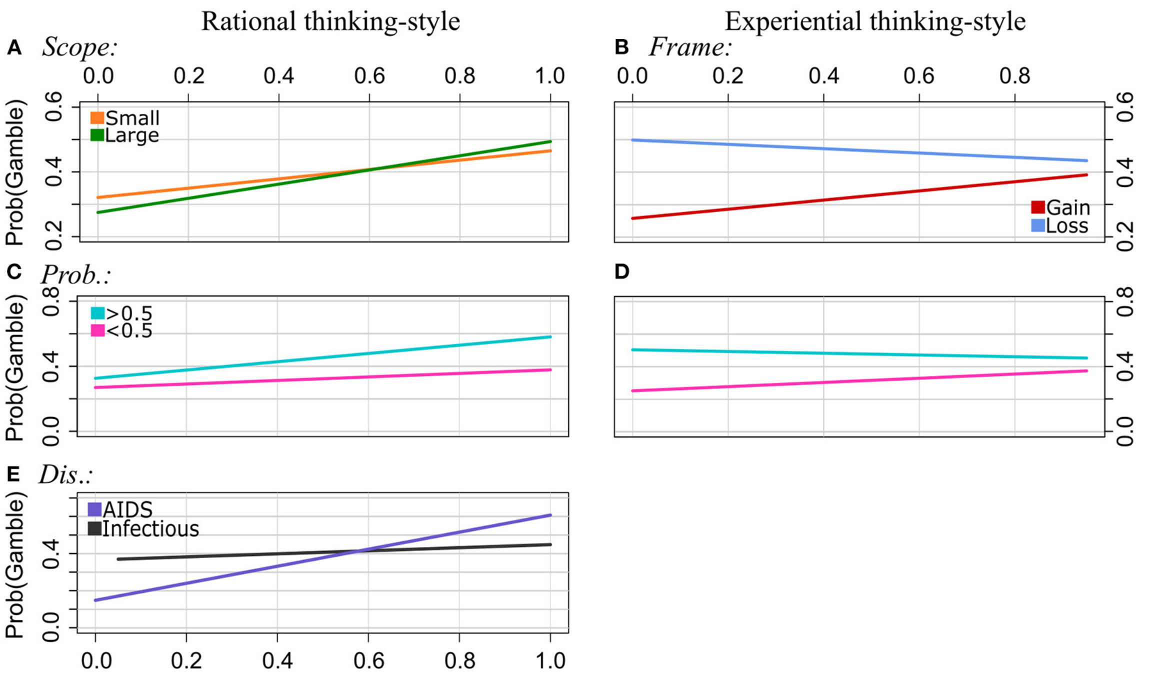


FIGURE 4

Experiment 2: Regression lines of the proportions of choosing the gamble option as a function of rational thinking-style (left column) and experiential thinking-style (right column), separately for the levels of Scope [pattern **(A)**, interaction effects of EX by Scope was n.s.], Frame [**(B)**, interaction effects of RA by Frame was n.s.], Probabilities **(C, D)**, and Disease [**(E)**, interaction effect of EX by Disease was n.s.]. Note that we applied a smaller range of values on the y-axis for the plots of the patterns **(A, B, E)** to illustrate the interaction effects more clearly.

Disease: We found a significant interaction effect between the RA scale and Disease. Specifically, the choice pattern for the Infectious disease was differed from that for AIDS. The gambling frequency increased with RA scores for both diseases. However, it increased even stronger with RA for AIDS than for the Infectious disease. No significant interaction effects between EX and the diseases were found.

3.4.2.2. Actively open-minded thinking

Table 8 shows the results of the interaction effects model when including actively open-minded thinking scores. We found significant interaction effects for AOT by Frame, AOT by Scope, and AOT by Probabilities. Significant effects are illustrated in Figure 5.

Frame: The higher the AOT scores, the more often the gamble was chosen in gain frames and in loss frames. However, the increase was steeper in the loss condition. That is, the framing effect becomes stronger with increasing AOT values.

Scope: The proportions of choosing the gamble option increased with increasing AOT scores in both categories of Scope. However, this increase was stronger for Scope Large than for Scope Small.

Probabilities: Gambling strongly increased with AOT scores for Probabilities (>0.5). This effect was much weaker for Probabilities (<0.5). That is, the effect of Probabilities on risky choices becomes stronger with increasing AOT scores.

No interactions of AOT by Disease were found.

TABLE 8 Experiment 2: Generalized linear mixed model, Interactions: actively open-minded thinking-style.

Fixed effects:	Est.	SE	*z*-value	*p*-value
(Intercept)	−1.215	0.545	−2.231	0.026
AOT	0.569	0.858	0.664	0.507
Frame (Gain)	−0.436	0.141	−3.104	0.002
Scope (Large)	−0.332	0.140	−2.371	0.018
Prob. (> 0.5)	−0.017	0.141	−0.120	0.905
Leukemia	−0.152	0.194	−0.785	0.432
AIDS	0.002	0.203	0.008	0.993
AOT × Frame	−0.483	0.222	−2.174	0.030
AOT × Scope	0.508	0.220	2.305	0.021
AOT × Prob.	1.509	0.225	6.719	<0.001
AOT × Leukemia	0.155	0.309	0.501	0.616
AOT × AIDS	−0.114	0.317	−0.359	0.719
Random effects:	**SD (Est)**			
Subjects (Intercept)	2.369			
Trial seq. (Intercept)	0.025			

16,592 observations provided by $n = 262$ participants in a series of 40 trials per block.

3.4.2.3. Stimulating Instrumental Risk Inventory (SIRI)

The results of the interaction effect models show that both stimulating and instrumental risk-style moderate the

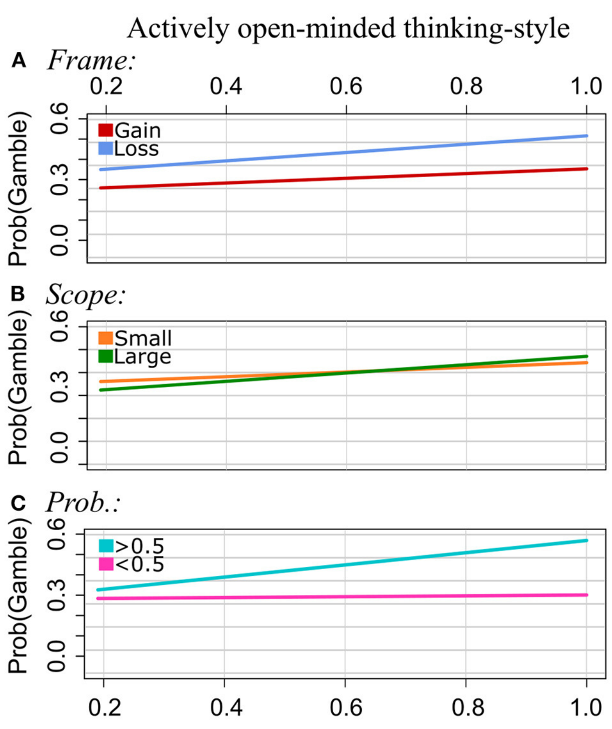


FIGURE 5
Experiment 2: Regression lines of the proportion of choosing the gamble option as a function of AOT, separately for the levels of Frame [pattern **(A)**], Scope **(B)**, and Probabilities **(C)**.

effects of Frame and Probabilities. Based on the statistical significance shown in Table 9, we interpret the interaction effect as follows:

Frame: Participants chose the gamble less often in the gain frame than in the loss frame. As illustrated in Figure 6, gambling increased in the gain frame with ST and IN scores. In the loss frame, however, gambling proportions did not change substantially with ST scores, and they slightly decreased with increasing IN scores. The findings suggest that framing effects become weaker with increasing stimulating and instrumental risk-style.

Probabilities: For Probabilities (>0.5), the proportions of choosing the risky option were relatively stable for participants with different scores of ST and IN. However, gambling increased with IN and EX scores when probabilities were low (<0.5). That is, the strength of the effect of Probabilities on choice behavior decreased with increasing scores of stimulating and instrumental risk-style (Figure 6).

TABLE 9 Experiment 2: Generalized linear mixed models, Interactions: stimulating and instrumental risk-style.

Stimulating risk-style				
Fixed effects:	Est.	SE	*z*-value	*p*-value
(Intercept)	−1.101	0.435	−2.530	0.011
ST	0.493	0.895	0.551	0.582
Frame (Gain)	−1.356	0.116	−11.690	<0.001
Scope (Large)	0.030	0.114	0.262	0.794
Prob. (>0.5)	1.579	0.117	13.486	<0.001
Leukemia	0.106	0.158	0.671	0.502
AIDS	−0.205	0.174	−1.177	0.239
ST × Frame	1.366	0.235	5.801	<0.001
ST × Scope	−0.108	0.233	−0.463	0.644
ST × Prob.	−1.472	0.237	−6.206	<0.001
ST × Leukemia	−0.374	0.321	−1.165	0.244
ST × AIDS	0.293	0.357	0.821	0.412
Random effects:		**SD (Est.)**		
Subject (Intercept)		2.385		
Trial seq. (Intercept)		0.023		
Instrumental risk-style				
Fixed effects:	Est.	SE	*z*-value	*p*-value
(Intercept)	−0.778	0.654	−1.190	0.234
IN	−0.166	1.029	−0.161	0.872
Frame (Gain)	−2.416	0.179	−13.502	<0.001
Scope (Large)	−0.033	0.175	−0.189	0.850
Prob. (>0.5)	1.815	0.178	10.170	<0.001
Leukemia	0.091	0.237	0.383	0.702
AIDS	−0.485	0.281	−1.730	0.084
IN × Frame	2.692	0.276	9.745	<0.001
IN × Scope	0.028	0.272	0.104	0.917
IN × Prob.	−1.450	0.276	−5.257	<0.001
IN × Leukemia	−0.266	0.371	−0.718	0.473
IN × AIDS	0.652	0.431	1.514	0.130
Random effects:		**SD (Est.)**		
Subject (Intercept)		2.409		
Trial seq. (Intercept)		0.022		

16,592 observations provided by $n = 262$ participants in a series of 40 trials per block.

3.5. Summary and discussion

The risky choice framing of the choice options as gains and losses and the probabilities of surviving/dying influenced choice behavior: Participants chose the gamble option more often in the loss frame than in the gain frame, and they chose it more often for probabilities >0.5 than for probabilities <0.5. In contrast to experiment 1, no effect of the number of affected people (Scope) was found. That is, the social science approach involving psychophysical elements was able to replicate the effects of framing and different probabilities, but it failed to replicate the effect of Scope on choice behavior.

We found no significant correlations between our narrow framing effect interpretation, i.e., frame-inconsistent choices, and

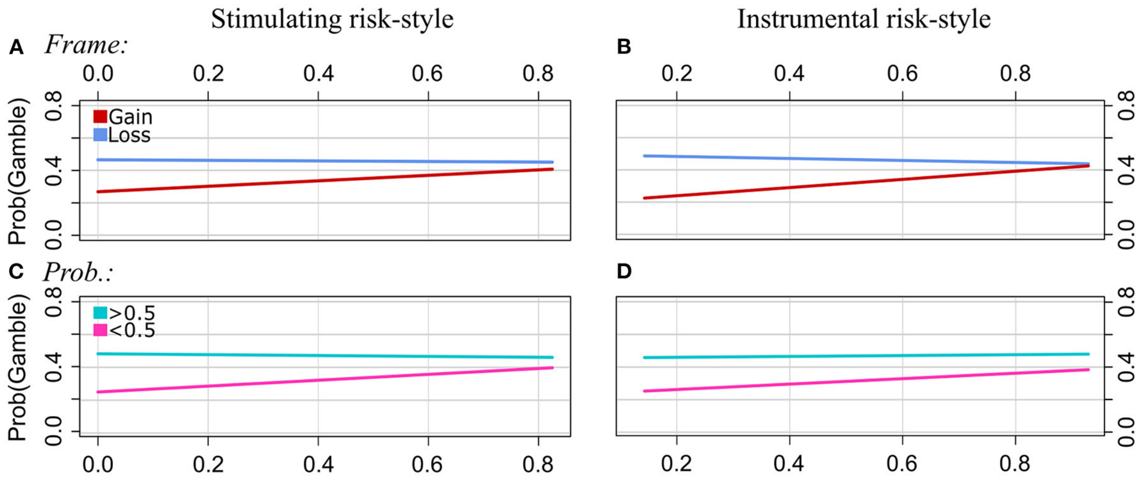


FIGURE 6
Experiment 2: Regression lines of the proportions of choosing the gamble option as a function of stimulating risk-style (left column) and instrumental risk-style (right column), separately for the levels of Frame [patterns **(A, B)**] and Probabilities **(C, D)**.

the scores measured using the psychometric instruments (REI, AOT, SIRI), supporting the results of experiment 1.

For the wide framing effect interpretation, i.e., the difference in proportions of choosing the gamble between the frames, we found that stimulating and instrumental risk-style, and experiential and actively open-minded thinking served as moderators of the framing effect in experiment 2. This finding is different from the results of experiment 1, where we found only stimulating and instrumental risk-style moderating the framing effect. In experiment 2, only instrumental risk-style moderated the framing effect as predicted by the theory. The other moderator effects have the opposite direction. In particular, the strength of the framing effect decreased with increasing scores of stimulating and instrumental risk-style. These are the opposite effects as observed in experiment 1. Moreover, the framing effect strength decreased with increasing scores of experiential thinking-style, and it increased with scores of actively open-minded thinking-style.

Furthermore, the results of experiment 2 show that, apart from the framing, other effects influencing the proportion of choosing the gamble option are moderated by cognitive-styles and risk-styles. As before, we found that some scales moderated the effects differently than one would expect from their underlying assumptions. In particular, the relationship between rational thinking-style and the effect of Scope was similar in both experiments. The effect of Scope reverses with increasing scores of rational thinking-style. Participants who scored low in rational thinking chose the gamble less often, and those scoring high chose the gamble more often for the Large than for Small Scope. Moreover, the same relationship was observed for the moderator effect of actively open-minded thinking-style on Scope in experiment 2. However, this was different from the finding in experiment 1, where the effect strength of Scope decreased with increasing scores of actively open-minded thinking. Note that, due to a higher heterogeneity of the sample composition, we observed a wider range of cognitive-style scores in experiment 2. Specifically, the sample of experiment 2 includes more participants with low scores of rational, experiential, and actively open-minded thinking. This finding might be an explanation for the failed replication of the main effect of Scope. It may simply have been canceled out due to the reversed effect direction for participants with low scores.

As in experiment 1, all measures of cognitive-style and risk-style moderated the effect of probabilities of surviving/dying for the (hypothetical) affected people. The findings replicate those of experiment 1. That is, only instrumental risk-style moderated the effect in a way that is in line with the theoretical assumptions. The other moderators show the opposite effect direction than expected according to the theory: The strength of the effect increased with scores of rational and actively open-minded thinking-style, and it decreased with increasing scores of experiential thinking-style and stimulating risk-style.

We found a relationship between disease problems and rational thinking-style in experiment 2. As compared to the Infectious Disease condition, gambling increased with scores of rational thinking for the AIDS problem. In experiment 1, however, we found that gambling decreased with the scores for the same scenario. Moreover, no other significant moderator effects of a scale on the effect of Disease on risky choice were found in experiment 2. This result is different from our findings in experiment 1, where we found that each of the scales moderated that particular effect in some way.

4. General discussion and conclusions

In the current study, we investigated the impact of individual intuitive and deliberative processing styles, i.e., rational, experiential, and actively open-minded thinking-style (Baron, 1993; Epstein, 1998), and stimulating and instrumental risk-styles (Zaleskiewicz, 2001), on the strength of risky choice framing effects. Previous research on that has shown mixed results, which might be explained by the large variety of methodological implementations. In particular, study designs varied (within vs. between-subject designs), framing effects have been interpreted in different ways, variables describing the decision problem beyond the framing of the choice options have been mostly ignored so far, and the studies often used student samples which are more homogeneous than, e.g., community or online samples.

We report two experiments involving elements of psychophysical data collection. For both experiments, we evaluated framing effects using two different interpretations: a narrow one, that is, we counted and compared the number of frame-inconsistent choices (FIC) participants made, and a wide interpretation, that is, we looked at the proportions of choosing the risky option depending on framing. Furthermore, our analysis considered other variables describing the decision problem, such as outcomes, probabilities, problem domains, and response time constraints. Experiment 1 was conducted in the lab using a student sample, and experiment 2 was conducted online using a more heterogeneous (with respect to e.g., age, education, and profession) online sample.

Once again, the psychophysical data collection approach has been shown to be an excellent method for measuring framing effects (see also, e.g., Guo et al., 2017; Diederich et al., 2020; Roberts et al., 2021; Wyszynski and Diederich, 2022). In addition to framing, we found other effects influencing the proportion of choosing the gamble. The findings of the strongest effects, i.e., framing and surviving/dying probability of the hypothetical people affected by a disease, could be replicated in experiment 2 using an approach that combined the social science approach with psychophysical elements. In both studies, participants chose the gamble more often in loss than in gain trials and for probabilities higher than 0.5 as compared to probabilities lower than 0.5. However, the less pronounced effect of Scope on choice behavior, i.e., fewer risky choices when more hypothetical people are affected by a disease, could not be replicated in our second experiment.

Results of both experiments consistently show no relationship between the number of FIC and cognitive-styles or risk-styles. The findings are in line with Mandel and Kapler (2018), who found no moderating effect of need for cognition (equivalent to rational thinking-style) and actively open-minded thinking on a similarly narrow interpretation of framing effects. However, other studies showed small (according to the classification of Cohen, 1988) but significant positive correlations suggesting the number of FIC to decrease with increasing scores of rational (LeBoeuf and Shafir, 2003; Björklund and Bäckström, 2008; Peng et al., 2019) and actively open-minded thinking-style (West et al., 2008; Erceg et al., 2022; Rachev et al., 2022). Note that the relationships between FIC (or a similar measure) and experiential thinking-style, and between FIC and stimulating-instrumental risk-styles have not been investigated so far. Taken together, our findings and those of previous research show, at best, a small effect of cognitive-styles and risk-styles based on the intuitive-deliberative processing approach on the number of FIC (or a similarly narrow interpretation of risky choice framing effects).

A different picture emerges when examining individual differences in cognitive-style and risk-style on framing effects in the wide interpretation, i.e., the impact of risky choice framing on the proportion of choosing the gamble: Although we did not find any impact of rational, experiential, and actively open-minded thinking-style on framing effect strength in experiment 1, experiment 2 showed that risky choice framing effects become weaker with increasing scores of experiential thinking-style and stronger with scores of actively open-minded thinking-style. According to the classification of Cohen (1988), the effect sizes are large and medium, respectively. Inconsistent with theoretical assumptions, the vast majority of previous research investigating rational-experiential thinking-style as moderator of the influence of framing on choice behavior have not found a direct relationship (see e.g., Levin et al., 2002; Shiloh et al., 2002; Björklund and Bäckström, 2008; Mahoney et al., 2011; Stark et al., 2017). Our findings on rational-thinking style support these studies. However, Simon et al. (2004) found a small (according to Cohen, 1988) moderator effect of rational thinking-style: in line with the theory, they observed stronger framing effects for individuals with lower need for cognition scores. Moreover, Mahoney et al. (2011) found in one of the five decision problems they used in their study that the strength of the framing effect increased with experiential thinking-style supporting the theory. In contrast, our second study revealed a strong relationship exhibiting the opposite direction (i.e., framing effect strength decreased with experiential thinking-style).

In contrast to previous research (Mahoney et al., 2011), we found in both experiments that stimulating and instrumental risk-style moderated the framing effect. However, the results of our two experiments are different: For both risk-styles, experiment 2 showed stronger framing effects for increasing scores; we found the opposite results (weaker framing effects with increasing scores) in experiment 1.

Furthermore, we found that the scores of the psychometric instruments we used to measure cognitive-styles and risk-styles moderated the effects of other problem-describing characteristics on risky choice. However, as for the framing effect, the directions of the effects, i.e, whether they were stronger or weaker for particular scale scores, were often different from what we expected according to the basic assumptions of the scales and also between our experiments. Some of the discrepancies could be explained by the more heterogeneous sample composition in experiment 2, where the scores of the instruments were measured on a broader range. It is also possible that other effects not considered in the current studies, as well as further interaction effects (e.g., 3-way-interactions), influence choice behavior. For instance, we know from previous studies that short time limits for making the risky choice enhance the framing effect (e.g., Guo et al., 2017; Diederich et al., 2018, 2020; Wyszynski and Diederich, 2022). In the current study, the effect of time limits was not moderated by scores of the psychometric instruments (see experiment 1), but it is well possible that individual cognitive-styles and risk-styles influence the relationship between time and framing or other interactions. However, the analysis of three-way interactions was not part of the current investigation, but they are worth being explored in future research. Moreover, it should also be questioned whether the scales actually measured precisely the individual differences they were supposed to measure.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: Open Science Framework (https://osf.io/thgdz/).

Ethics statement

The studies involving human participants were reviewed and approved by Jacobs University Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

Author contributions

MW: conceptualization, methodology, software, formal analysis, investigation, data curation, writing—original draft, writing—review and editing, visualization, and project administration. AD: conceptualization, methodology, validation, resources, writing—original draft, writing—review and editing, supervision, and funding acquisition. Both authors contributed to the article and approved the submitted version.

Funding

This work was supported by Deutsche Forschungsgemeinschaft grant DFG FOR2104 (Need-based justice and distributive procedures), DI 506/13-1 and DI 506/13-2 to AD.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1086699/full#supplementary-material

References

Aczel, B., Szollosi, A., and Bago, B. (2018). The effect of transparency on framing effects in within-subject designs. *J. Behav. Decis. Mak.* 31, 25–39. doi: 10.1002/bdm.2036

Almashat, S., Ayotte, B., Edelstein, B., and Margrett, J. (2008). Framing effect debiasing in medical decision making. *Patient Educ. Couns.* 71, 102–107. doi: 10.1016/j.pec.2007.11.004

Appelt, K. C., Milch, K. F., Handgraaf, M. J. J., and Weber, E. U. (2011). The Decision Making Individual Differences Inventory and guidelines for the study of individual differences in judgment and decision-making research. *Judgm. Decis. Mak.* 6, 252–262. doi: 10.1017/S1930297500001455

Aquino, J. (2018). *descr: Descriptive Statistics*. R Package Version 1.1.4.

Baron, J. (1993). Why teach thinking? - An essay. *Appl. Psychol.* 42, 191–214. doi: 10.1111/j.1464-0597.1993.tb00731.x

Baron, J. (2010). Looking at individual subjects in research on judgment and decision making (or anything). *Acta Psychol. Sin.* 42, 88–98. doi: 10.3724/SP.J.1041.2010.00088

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv Preprint.* arXiv:1406.5823. doi: 10.18637/jss.v067.i01

Björklund, F., and Bäckström, M. (2008). Individual differences in processing styles: validity of the Rational–Experiential Inventory. *Scand. J. Psychol.* 49, 439–446. doi: 10.1111/j.1467-9450.2008.00652.x

Bruine de Bruin, W., Parker, A. M., and Fischhoff, B. (2007). Individual differences in adult decision-making competence. *J. Pers. Soc. Psychol.* 92, 938–956. doi: 10.1037/0022-3514.92.5.938

Cacioppo, J. T., and Petty, R. E. (1982). The need for cognition. *J. Pers. Soc. Psychol.* 42, 116–131. doi: 10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., and Feng Kao, C. (1984). The efficient assessment of need for cognition. *J. Pers. Assess.* 48, 306–307. doi: 10.1207/s15327752jpa4803_13

Chaiken, S., and Trope, Y. (1999). *Dual-Process Theories in Social Psychology*. New York, NY: Guildford Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn*. New York, NY: Routledge.

Corbin, J. C. (2015). *Fuzzy-Trace Theory and Risky Choice Framing: An Individual Differences Approach*. Ph.D. Thesis, Cornell University, Ithaca, NY.

Diederich, A., Wyszynski, M., and Ritov, I. (2018). Moderators of framing effect in variations of the Asian Disease problem: time constraint, need, and disease type. *Judgm. Decis. Mak.* 13, 529–546. doi: 10.1017/S1930297500006574

Diederich, A., Wyszynski, M., and Traub, S. (2020). Need, frame, and time constraints in risky decision making. *Theory Decis.* 89, 1–37. doi: 10.1007/s11238-020-09744-6

Epstein, S. (1973). The self-concept revisited: or a theory of a theory. *Amer. Psychol.* 28, 404–416. doi: 10.1037/h0034679

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *Amer. Psychol.* 49, 709–724. doi: 10.1037/0003-066X.49.8.709

Epstein, S. (1998). "Cognitive-experiential self-theory," in *Advanced Personality, The Plenum Series in Social/Clinical Psychology*, eds D. F. Barone, M. Hersen, and V. B. Van Hasselt (Boston, MA: Springer US), 211–238. doi: 10.1007/978-1-4419-8580-4_9

Epstein, S., Pacini, R., Denes-Raj, V., and Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *J. Pers. Soc. Psychol.* 71, 390–405. doi: 10.1037/0022-3514.71.2.390

Erceg, N., Galić, Z., and Bubić, A. (2022). Normative responding on cognitive bias tasks: some evidence for a weak rationality factor that is mostly explained by numeracy and actively open-minded thinking. *Intelligence* 90, 101619. doi: 10.1016/j.intell.2021.101619

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* 59, 255–278. doi: 10.1146/annurev.psych.59.103006.093629

Fatmawati, I. (2015). "The moderating effects of need for cognition on framed message promoting electricity energy saving behavior," in *The 2015 International Conference of Management Sciences*, (Universitas Muhammadiyah Yogyakarta), 54–68.

Frederick, S. (2005). Cognitive reflection and decision making. *J. Econ. Perspect.* 19, 25–42. doi: 10.1257/089533005775196732

Frisch, D. (1993). Reasons for framing effects. *Organ. Behav. Hum. Decis. Process.* 54, 399–429. doi: 10.1006/obhd.1993.1017

Green, P., and MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol. Evol.* 7, 493–498. doi: 10.1111/2041-210X.12504

Guo, L., Trueblood, J. S., and Diederich, A. (2017). Thinking fast increases framing effects in risky decision making. *Psychol. Sci.* 28, 530–543. doi: 10.1177/0956797616689092

Haran, U., Ritov, I., and Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgm. Decis. Mak.* 8, 188–201. doi: 10.1017/S1930297500005921

Harrell, F. E. Jr. (2021). *Hmisc: Harrell Miscellaneous*. R Package Version 4.6-0.

Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007

Kahneman, D., and Frederick, S. (2002). "Representativeness revisited: attribute substitution in intuitive judgment," in *Heuristics and Biases: The Psychology of Intuitive Judgment*, eds T. Gilovich, D. Griffin, and D. Kahneman (Cambridge: Cambridge University Press), 49–81.

Kahneman, D., and Frederick, S. (2005). "A model of heuristic judgment," in *The Cambridge Handbook of Thinking and Reasoning*, eds K. J. Holyoak and R. G. Morrison (Cambridge: Cambridge University Press), 267–293.

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econ. J. Econ. Soc.* 47, 263–291. doi: 10.2307/1914185

Kristensen, M. and Hansen, T. (2004). Statistical analyses of repeated measures in physiological research: a tutorial. *Adv. Physiol. Educ.* 28, 2–14. doi: 10.1152/advan.00042.2003

Kühberger, A. (1998). The influence of framing on risky decisions: a meta-analysis. *Organ. Behav. Hum. Decis. Process.* 75, 23–55. doi: 10.1006/obhd.1998.2781

Kühberger, A., Schulte-Mecklenbeck, M., and Perner, J. (1999). The effects of framing, reflection, probability, and payoff on risk preference in choice tasks. *Organ. Behav. Hum. Decis. Process.* 78, 204–231. doi: 10.1006/obhd.1999.2830

LeBoeuf, R. A. and Shafir, E. (2003). Deep thoughts and shallow frames: on the susceptibility to framing effects. *J. Behav. Decis. Mak.* 16, 77–92. doi: 10.1002/bdm.433

Levin, I. P., Gaeth, G. J., Schreiber, J., and Lauriola, M. (2002). A new look at framing effects: distribution of effect sizes, individual differences, and independence of types of effects. *Organ. Behav. Hum. Decis. Process.* 88, 411–429. doi: 10.1006/obhd.2001.2983

Levin, I. P., Schneider, S. L., and Gaeth, G. J. (1998). All frames are not created equal: a typology and critical analysis of framing effects. *Organ. Behav. Hum. Decis. Process.* 76, 149–188. doi: 10.1006/obhd.1998.2804

Li, S., and Liu, C.-J. (2008). Individual differences in a switch from risk-averse preferences for gains to risk-seeking preferences for losses: can personality variables predict the risk preferences? *J. Risk Res.* 11, 673–686. doi: 10.1080/13669870802086497

Mahoney, K., Buboltz, W., Levin, I., Doverspike, D., and Svyantek, D. (2011). Individual differences in a within-subjects risky-choice framing study. *Pers. Individ. Dif.* 51, 248–257. doi: 10.1016/j.paid.2010.03.035

Mandel, D. R., and Kapler, I. V. (2018). Cognitive style and frame susceptibility in decision-making. *Front. Psychol.* 9:1461. doi: 10.3389/fpsyg.2018.01461

Miller, P. M., and Fagley, N. S. (1991). The effects of framing, problem variations, and providing rationale on choice. *Pers. Soc. Psychol. Bull.* 17, 517–522. doi: 10.1177/0146167291175006

Mukherjee, K. (2010). A dual system model of preferences under risk. *Psychol. Rev.* 117, 243–255. doi: 10.1037/a0017884

Pacini, R., and Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *J. Pers. Soc. Psychol.* 76, 972–987. doi: 10.1037/0022-3514.76.6.972

Parker, A. M., and Fischhoff, B. (2005). Decision-making competence: external validation through an individual-differences approach. *J. Behav. Decis. Mak.* 18, 1–27. doi: 10.1002/bdm.481

Peng, J., Feng, T., Zhang, J., Zhao, L., Zhang, Y., Chang, Y., et al. (2019). Measuring decision-making competence in Chinese adults. *J. Behav. Decis. Mak.* 32, 266–279. doi: 10.1037/t76708-000

Peterson, R. A. (2001). On the use of college students in social science research: insights from a second-order meta-analysis. *J. Consum. Res.* 28, 450–461. doi: 10.1086/323732

Piñon, A., and Gambara, H. (2005). A meta-analytic review of framming effect: risky, attribute and goal framing. *Psicothema* 17, 325–331.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Rachev, N. R., Geiger, S. J., Vintr, J., Kirilova, D., Nabutovsky, A., and Nelsson, J. (2022). Actively open-minded thinking, bullshit receptivity, and susceptibility to framing. *Eur. J. Psychol. Assess.* 38, 440–451. doi: 10.1027/1015-5759/a000685

Revelle, W. (2020). *psych: Procedures for Psychological, Psychometric, and Personality Research.* R Package Version 2.0.12. Northwestern University, Evanston.

Roberts, I. D., Teoh, Y. Y., and Hutcherson, C. A. (2021). Time to pay attention? Information search explains amplified framing effects under time pressure. *Psychol. Sci.* 33, 90–104. doi: 10.1177/09567976211026983

Schönbrodt, F. D., and Perugini, M. (2013). At what sample size do correlations stabilize? *J. Res. Pers.* 47, 609–612. doi: 10.1016/j.jrp.2013.05.009

Shiloh, S., Salton, E., and Sharabi, D. (2002). Individual differences in rational and intuitive thinking styles as predictors of heuristic responses and framing effects. *Pers. Indiv. Dif.* 32, 415–429. doi: 10.1016/S0191-8869(01)00034-4

Sieck, W., and Yates, J. F. (1997). Exposition effects on decision making: choice and confidence in choice. *Organ. Behav. Hum. Decis. Process.* 70, 207–219. doi: 10.1006/obhd.1997.2706

Simon, A. F., Fagley, N. S., and Halleran, J. G. (2004). Decision framing: moderating effects of individual differences and cognitive processing. *J. Behav. Decis. Mak.* 17, 77–93. doi: 10.1002/bdm.463

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychol. Bull.* 119, 3–22. doi: 10.1037/0033-2909.119.1.3

Stanovich, K. E. (1999). *Who is Rational? Studies of Individual Differences in Reasoning.* New York, NY: Psychology Press. doi: 10.4324/9781410603432

Stanovich, K. E., and West, R. F. (1998). Individual differences in rational thought. *J. Exp. Psychol. Gen.* 127, 161–188. doi: 10.1037/0096-3445.127.2.161

Stanovich, K. E., and West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *Behav. Brain Sci.* 23, 645–665. doi: 10.1017/S0140525X00003435

Stark, E., Baldwin, A. S., Hertel, A. W., and Rothman, A. (2017). The role of rational and experiential processing in influencing the framing effect. *J. Soc. Psychol.* 157, 308–321. doi: 10.1080/00224545.2016.1198301

Steiger, A., and Kühberger, A. (2018). A meta-analytic re-appraisal of the framing effect. *Z. Psychol.* 226, 45–55. doi: 10.1027/2151-2604/a000321

Takemura, K. (1994). Influence of elaboration on the framing of decision. *J. Psychol.* 128, 33–39. doi: 10.1080/00223980.1994.9712709

Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science* 211, 453–458. doi: 10.1126/science.7455683

West, R. F., Toplak, M. E., and Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: associations with cognitive ability and thinking dispositions. *J. Educ. Psychol.* 100, 930–941. doi: 10.1037/a0012842

Wyszynski, M., and Diederich, A. (2022). Keep your budget together! Investigating determinants on risky decision-making about losses. *PLoS ONE* 17:e0265822. doi: 10.1371/journal.pone.0265822

Wyszynski, M., Diederich, A., and Ritov, I. (2020). Gamble for the needy! Does identifiability enhances donation? *PLoS ONE* 15:e0234336. doi: 10.1371/journal.pone.0234336

Yu, Z., Guindani, M., Grieco, S. F., Chen, L., Holmes, T. C., and Xu, X. (2022). Beyond t test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research. *Neuron* 110, 21–35. doi: 10.1016/j.neuron.2021.10.030

Zaleskiewicz, T. (2001). Beyond risk seeking and risk aversion: personality and the dual nature of economic risk taking. *Eur. J. Pers.* 15(S1), S105–S122. doi: 10.1002/per.426

TYPE Original Research
PUBLISHED 30 March 2023
DOI 10.3389/fpsyg.2023.1132168

OPEN ACCESS

EDITED BY
Sergio Da Silva,
Federal University of Santa Catarina,
Brazil

REVIEWED BY
Jyrki Suomala,
Laurea University of Applied Sciences,
Finland
Bin Liu,
Hainan University,
China

*CORRESPONDENCE
Jay Hegdé
✉ jhegde@augusta.edu

SPECIALTY SECTION
This article was submitted to
Cognitive Science,
a section of the journal
Frontiers in Psychology

RECEIVED 26 December 2022
ACCEPTED 07 February 2023
PUBLISHED 30 March 2023

CITATION
Branch F and Hegdé J (2023) Toward a more nuanced understanding of probability estimation biases.
Front. Psychol. 14:1132168.
doi: 10.3389/fpsyg.2023.1132168

Toward a more nuanced understanding of probability estimation biases

Fallon Branch and Jay Hegdé*

Department of Neuroscience and Regenerative Medicine, Medical College of Georgia, Augusta University, Augusta, GA, United States

In real life, we often have to make judgements under uncertainty. One such judgement task is estimating the probability of a given event based on uncertain evidence for the event, such as estimating the chances of actual fire when the fire alarm goes off. On the one hand, previous studies have shown that human subjects often significantly misestimate the probability in such cases. On the other hand, these studies have offered divergent explanations as to the exact causes of these judgment errors (or, synonymously, biases). For instance, different studies have attributed the errors to the neglect (or underweighting) of the prevalence (or base rate) of the given event, or the overweighting of the evidence for the individual event ('individuating information'), etc. However, whether or to what extent any such explanation can fully account for the observed errors remains unclear. To help fill this gap, we studied the probability estimation performance of non-professional subjects under four different real-world problem scenarios: (i) Estimating the probability of cancer in a mammogram given the relevant evidence from a computer-aided cancer detection system, (ii) estimating the probability of drunkenness based on breathalyzer evidence, and (iii & iv) estimating the probability of an enemy sniper based on two different sets of evidence from a drone reconnaissance system. In each case, we quantitatively characterized the contributions of the various potential explanatory variables to the subjects' probability judgements. We found that while the various explanatory variables together accounted for about 30 to 45% of the overall variance of the subjects' responses depending on the problem scenario, no single factor was sufficient to account for more than 53% of the explainable variance (or about 16 to 24% of the overall variance), let alone all of it. Further analyses of the explained variance revealed the surprising fact that no single factor accounted for significantly more than its 'fair share' of the variance. Taken together, our results demonstrate quantitatively that it is statistically untenable to attribute the errors of probabilistic judgement to any single cause, including base rate neglect. A more nuanced and unifying explanation would be that the actual biases reflect a weighted combination of multiple contributing factors, the exact mix of which depends on the particular problem scenario.

KEYWORDS

base rate neglect, cognitive rules of thumb, individuating information, inverse fallacy, judgement and decision-making under uncertainty, miss rate neglect, representativeness heuristic

Introduction

In everyday life, ordinary people and trained professionals alike often encounter situations where they must estimate the probability of an event using imperfect evidence for the event. If the lawn is wet in the morning, what are chances that it rained during the previous night (Pearl, 1988)? What is the probability that there is an intruder in your yard if the dog barks? If someone is positively identified in a police lineup, how likely is it that this person is the actual culprit? What are the chances that the patient actually has cancer when a physician diagnoses one? Obviously, errors in estimating these probabilities can have significant real-world consequences.

A large number of previous studies have examined how well human subjects solve this problem in a wide variety of contexts (Kahneman and Tversky, 1973; Eddy, 1982; Kahneman et al., 1982; Fischhoff and Bar-Hillel, 1984; Thompson and Schumann, 1987; Bar-Hillel, 1991; Koehler, 1996; Villejoubert and David, 2002; Raacke, 2005; Kalinowski et al., 2008; Mandel, 2014; Raab and Gigerenzer, 2015; Dahlman et al., 2016). While these studies understandably vary in the exact task they used, they typically have the following design: The subjects are presented with a problem scenario, including the actual binary outcome (e.g., a patient is positively diagnosed with cancer or not) and the three underlying probabilistic factors: (i) true positive rate of the diagnosis, i.e., the probability that the patient actually has cancer given a positive diagnosis, (ii) false positive rate, the patient does not actually have cancer, and the diagnosis was a 'false alarm', and (iii) the prevalence, or base rate, of cancer in the given patient population. The subjects are then asked to estimate the actual probability of the outcome given the evidence for the outcome (e.g., probability that the patient actually has cancer given the diagnosis). The studies then compare the subjects' probability estimates with the corresponding theoretically expected probabilities (see General Methods below for technical details).

Using this general approach, previous studies have consistently found that human subjects substantially misestimate the probabilities. That is, the subjects' estimates typically deviate substantially from the theoretically expected probabilities (Eddy, 1982; Kahneman et al., 1982; Koehler, 1996; Mandel, 2014; Raab and Gigerenzer, 2015). Actually, for most real-world scenarios where the base rate is low, the subjects tend to *overestimate* the probability (Eddy, 1982; Kahneman et al., 1982; Koehler, 1996; Mandel, 2014; Raab and Gigerenzer, 2015).

An obvious next question is why. About this, previous studies have offered widely differing explanations: One longstanding view has been that these errors arise because the subjects attach too little weight to (or 'underweight', or neglect) the underlying prevalence, or base rate, of the event (Kahneman and Tversky, 1973; Fischhoff and Bar-Hillel, 1984; Bar-Hillel, 1991). This is why these judgements have been referred to as base rate fallacy, base rate neglect, or base rate bias (Kahneman and Tversky, 1973; Fischhoff and Bar-Hillel, 1984; Thompson and Schumann, 1987; Koehler, 1996; Dahlman et al., 2016). Some studies have also attributed the judgement errors to *overweighting* (i.e., attaching too much importance to) the evidence for a given individual event (or 'individuating' information; Kahneman and Tversky, 1973; Bar-Hillel, 1980; Kahneman and Tversky, 1982); the inverse fallacy (Villejoubert and David, 2002; Raacke, 2005; Kalinowski et al., 2008); and the so-called 'miss rate neglect', which actually refers to the neglect of false positive rates (Dahlman et al., 2016). On the one hand, few studies have explicitly claimed that any of these individual causes fully account for all of the observed errors. For instance, even those studies that attribute the estimation errors to base rate neglect stop short of explicitly offering base rate neglect as the *sole* explanation. On the other hand, it remains unclear as to whether and to what extent base rate neglect or any other aforementioned cause can, by itself fully account for the empirically observed errors.

The present study seeks to help fill this gap by focusing on a simple, straightforward question: When subjects estimate the probability of an event using the aforementioned established task paradigm, how much do various predictor variables contribute to the subjects' estimated probabilities? We addressed this question using multiple different problem scenarios, and replicated the aforementioned biases in each case. We then quantitatively evaluated the extent to which the various potential causes contributed to the observed biases in each case. While we make no claims that our findings are the final word on this topic (see Discussion), we do show that there are principled reasons to call into question the prevailing explanations of what causes the observed biases.

General methods

Participants

The present study consisted of four mutually independent experiments. All procedures used in each experiment were approved in advance by the Institutional Review Board (IRB) of Augusta University, Augusta, GA, United States, where the experiments were carried out. Subjects were recruited using IRB-approved ads posted on various campus sites. All the subjects who participated in this research were adult volunteers with normal or corrected-to-normal vision, and provided informed consent prior to participating in the study. All were non-professional subjects, in the sense that none of the subjects had any known expertise in the task used in any of the four experiments, and that no subject was recruited, included, or excluded based on their education, training, or expertise. A total of 23 different subjects (mean age, 22.23 years ±4.23 [SD], excluding one subject whose age was not available; 16 women and one non-binary person) participated in this study. Some subjects participated in more than one experiment (see Supplementary material).

Procedure

As noted above, accurately judging the probability of an actual outcome or event A (e.g., actual cancer) given binary evidence B for the event (e.g., diagnosis of cancer) requires one to jointly evaluate the following four pieces of information:

1. The prevalence, or base rate $p(A)$ of the event,
2. The true positive rate, i.e., hit rate or $p(B|A)$, which denotes the probability of observing the evidence. B given that event A has actually occurred,
3. The false positive rate, i.e., false alarm rate, $p(B|\text{-}A)$, which denotes the probability of observing the evidence B given that event A has not actually occurred, and
4. Whether or not the evidence indicates the event has occurred, i.e., $B=1$ or $B=0$. Given the aforementioned four pieces of information, the expected probability of the event A given that the evidence for the event had been observed, i.e., $B=1$, is precisely specified by the Bayesian formula

$$p(A|B)=\left[p(A)p(B|A)\right]/\left[p(A)p(B|A)+p(-A)p(B|-A)\right]. \tag{1a}$$

The expected probability that the underlying event has not occurred given that evidence for the event has not been observed, i.e., $B=0$, is given by

$$p(-A|-B) = \left[p(-A)\,p(-B|-A)\right] / \left[p(A)\,p(B|A) + p(-A)\,p(B|-A)\right]. \qquad (1b)$$

We used the above equations to calculate the theoretically expected probability for each given combination of input values for the equations (Eddy, 1982; Raab and Gigerenzer, 2015). It is important to emphasize, however, that our study neither required the subjects to estimate the probabilities in this fashion, nor did it assume that they did. That is, our study neither required the subjects to carry out mathematical calculations in their heads, nor assumed that this is how subjects do the task at hand.

Because the present study aimed to characterize the factors that underlie previously reported errors in probability estimation, we needed to reproduce the underlying errors in our study. For this reason, we simply adopted the task paradigm used in the influential study by Eddy (2005) and many others since [for a review, see Koehler, 1996]. Note that this study did not aim to, nor does it claim to, address the so-called 'ecological validity' of this task paradigm (Spellman, 1996).

Task paradigm

During each trial, subjects were simultaneously given the above four items of information on a computer screen. For instance, in the context of Experiment 1 below, $p(A)$ was the base rate of breast cancer; $p(B|A)$ and $p(B|\text{-}A)$ were the hit and false-alarm rates of a hypothetical CAD (computer-assisted diagnosis) system, and B was the binary decision of the system (see the Methods under the individual experiments below for details).

The meaning of each term was explained to the subjects interactively using both written and verbal explanations. We interactively ascertained that the subjects accurately understood the meanings of the terms prior to proceeding with the trials. Subjects were not provided any information whatsoever about the expected probabilities or approaches, Bayesian or otherwise, to carrying out the task.

Using only the information provided, subjects had to estimate, using a mouse-driven on-screen slider, the percent chance that the given event had actually occurred (also see individual experiments below). Subjects were afforded *ad libitum* opportunity to view the on-screen information and enter their response. They received no feedback.

The various rates and probabilities were presented both as fractions of 1 (e.g., 0.005) and as the corresponding 'natural' frequencies (e.g., 5 in 1000). This is because previous studies (Hoffrage and Gigerenzer, 1998; Hoffrage et al., 2015), and our preliminary work (Sevilla and Hegdé, 2017), have shown that some subjects are more comfortable with natural frequencies. Before the actual data collection, subjects underwent practice trials until they indicated they were fully familiar with all aspects of the task. The data from the practice trials were discarded.

Data analysis

We analyzed the data using scripts custom-written in the R language (R_Core_Team, 2019). We carried out parametric statistical tests of significance where appropriate, and randomization-based tests of significance (Manly, 2007) otherwise. Where necessary, we corrected for multiple comparisons using the false discovery rate (FDR) method (Benjamini and Hochberg, 1995).

Power analyses

These analyses were carried out using the R library *pwr*. Before initiating the present study, we carried out *a priori* power analyses to determine the subject recruitment target. To do this, we used the empirically observed fit of the data from a pilot study (Branch et al., 2022) as the expected fit of the model (see below), and calculated the total number of trials (pooled across all subjects). The results indicated that at least 63 trials (pooled across all subjects and repetitions) would be needed to achieve a statistical power of 0.90. *A posteriori* power analyses using the actual data indicated that our data achieved a power of >0.95 for the regression analyses in each of the four experiments in this study.

Generalized linear mixed modeling

We used GLMM to determine the contribution of the various predictor variables to the subjects' reported probabilities. GLMM is the appropriate modeling approach when the predictor variables are 'mixed', in that one or more variables are factorial or categorical (e.g., the binary decision of the system, in our case), and others are continuous (e.g., base rate; Dean and Nielsen, 2007; Berridge and Crouchley, 2011; Fox and Fox, 2016). GLMM has been used extensively for this purpose in psychological research (Dean and Nielsen, 2007; Berridge and Crouchley, 2011; Fox and Fox, 2016; Bono et al., 2021). In this report, we follow the recommended practices of reporting GLMM results [Bono et al., 2021; also see Cooper and American Psychological Association (2018)].

We carried out GLMM in two stages. We first constructed an exploratory model, which we will refer to as the "Initial Model," in which we included as predictor variables all the primary independent variables in the given experiment and their pairwise interactions. For Experiments 1 through 3, the primary independent variables were base rate, false alarm rate, and the binary decision of the system. Hit rate was not included as a variable, because the hit rate was not varied in these experiments. The hit rate was varied in Experiment 4, and was included in the modeling of the results for Experiment 4.

Our modeling approach was designed to safeguard against the common pitfalls of regression modeling of real-world data (Aggarwal and Ranganathan, 2017; Ranganathan and Aggarwal, 2018). We will note many of the features of our approach in this section, and will highlight additional ones in context in the Results section of various experiments as appropriate, and will discuss the limitations of our approach in the General Discussion section.

One of the potential pitfalls of GLMM in particular, and of multiple regression in general, arises when the predictor (or independent) variables are mutually correlated, i.e., the nominally independent variables are not actually independent (Aggarwal and Ranganathan, 2017; Ranganathan and Aggarwal, 2018). Note that this caveat does not apply to our experiments, because the predictor variables were truly independent in that they were varied independently of each other. Note also that the fact that two or more predictor variables may have a joint influence on the response variable is not the same as the predictor variables being mutually correlated (Jaccard and Turrisi, 2003). In our models, such joint influences are captured by the interaction between predictor variables (Jaccard and Turrisi, 2003).

Analysis of the relative importance of the independent variables: *lmg* statistic

The relative importance of predictor variables was assessed using the standard *lmg* statistic (Lindeman et al., 1980; Grömping, 2006). This is a well-established statistical analysis that can better assess the relative importance (or, equivalently, the relative contribution) of the predictor variables better than the conventional linear regression metrics, e.g., when the predictor variables covary (Lindeman et al., 1980; Grömping, 2006).

Model selection

We used standard model selection procedures (Draper and Smith, 1998; Burnham et al., 2002) to evaluate the aforementioned Initial Model to determine the most parsimonious version of this model that accounted for greatest possible amount of the information in the data. Model selection is the standard approach to minimizing overfitting effects, one of the common pitfalls of multiple regression (Hegdé, 2021).

We will refer to the model ultimately selected in this fashion as the "Final Model." Specifically, we used the aforementioned Initial Model as the input to a stepwise model selection algorithm that used the Akaike Information Criterion or AIC (Venables and Ripley, 2003). While model selection was carried independently of the aforementioned *lmg* analysis (and vice versa), the results of the two analyses were largely consistent with each other (not shown).

Note that the above modeling procedures make no assumptions about how the subjects arrived at their probability estimations. Note, in particular, that our models do not, however indirectly, utilize Equations 1a and 1b above. Instead, our models are data driven, our methods simply determine the model that best fits the empirical data at hand. Note also that GLMM modeling neither assumes nor requires that the underlying relationship between the predictor variables on the one hand and the response variables on the other is linear (Dean and Nielsen, 2007; Berridge and Crouchley, 2011; Fox and Fox, 2016). On the other hand, the GLMM approach does make certain standard assumptions about the nature of the underlying data (Dean and Nielsen, 2007; Berridge and Crouchley, 2011; Fox and Fox, 2016). In general, data in all four experiments adequately met these assumptions (data not shown). In particular, the residuals were normally distributed in all four experiments (not shown), indicating that the linear models adequately captured the underlying relationship between the independent variables vs. response variables (Fox and Fox, 2016; Fox and Weisberg, 2019).

Relative contribution index

We calculated RCI values individually for each of the variables retained in the Final Model. We defined RCI value for the given variable *i* as

$$\mathrm{RCI} = lmg_{i,actual} \,/\, lmg_{i,random} \tag{2}$$

where $lmg_{i,\ actual}$ was the actual *lmg* value for the given variable.

To calculate the $lmg_{i,\ random}$ value, we randomly reshuffled the values of each variable *i* across trials. We then refitted the same model to the randomized data and re-calculated the *lmg* value for each variable *i*. We repeated this process 1,000 times, calculated the *lmg* value for each variable *i*. The mean *lmg* value for a given variable *i* across the randomization was defined as the $lmg_{i,\ random}$ value for that variable. The uncorrected 95% confidence interval (CI) was defined as the 5th and the 95th percentiles the 1,000 $lmg_{i,\ random}$ values. The *p* value for the corresponding one-tailed alternative hypothesis was defined as the proportion of times the $lmg_{i,\ random}$ value was higher (or lower) than the $lmg_{i,\ actual}$ value (Manly, 2007). These *p* values were corrected for multiple comparison using the FDR method (Benjamini and Hochberg, 1995).

Note that the above RCI analysis implicitly uses the null hypothesis that all the predictor variables contribute equally to the observed probability estimates and tests this hypothesis against the empirical data. This is a principled approach, especially because the aforementioned previous studies of neglect implicitly assume that the proper estimation requires equal weighting (Kahneman and Tversky, 1973; Fischhoff and Bar-Hillel, 1984; Thompson and Schumann, 1987; Koehler, 1996; Dahlman et al., 2016).

Experiment 1: Estimating the probability of cancer in a mammogram based on CAD system evidence

Methods

Thirteen subjects (10 women; mean age, 19.67 years ±1.67) participated in this experiment. Subjects were simultaneously given four items of information on a computer screen:

1. The prevalence, or base rate, of breast cancer in the given cohort of patients [i.e., $p(A)$ in Eqs. 1a,b above],
2. The hit rate $p(B|A)$ of a hypothetical CAD system for breast cancer detection,
3. The false alarm rate $p(B|\text{-}A)$ of the system, and
4. The binary decision of the system as to whether or not a given mammogram was positive for cancer. No mammogram was shown. That is, the subjects had to estimate the probability that the given unseen mammogram was positive for cancer based solely on the above four items of information.

During this Experiment, we held the hit rate constant at 1.0, and systematically varied the remaining three variables, and measured its effect on the subjects' estimated probabilities of cancer. During any given trial, values for each of the three variables were randomly drawn from the corresponding repertoire of possible values: two possible values of the base rate (0.05 or 0.005), five possible values of the false alarm rate (0.05, 0.25, 0.5, 0.75 and 0.95), and two possible values for the binary decision of the CAD system (0 or 1, corresponding to whether the mammogram was positive or negative for cancer, respectively). Note that the values of the four variables varied independently from one trial to the next. Each possible combination of these values was tested exactly once during each block of 20 trials. Subjects performed 1 or 2 blocks each. Data were pooled across subjects.

It is worth noting that the data we present in this experiment are entirely independent of the data we have presented in a comparable previous study that was designed to address a different issue (Branch et al., 2022). That is, the data in the two studies were collected independently of each other using non-overlapping sets of subjects. Moreover, task parameters used in the previous study were different from those used in this experiment.

Results

The cancer probability estimates pooled across all subjects are plotted as a function of the corresponding theoretically expected probabilities in Figure 1A, where each plotting symbol denotes the reported probability estimate from an individual subject during a single trial (see legend for details). The plotting symbols corresponding to the two decisions of the CAD system (i.e., that the given mammogram is positive or negative for cancer) are denoted as a red circle or green triangle, respectively (see key at bottom right of Figure 1). Each vertical column represents the data points for a single theoretically expected probability.

Two qualitative aspects of these results are worth noting. First, the subjects generally misestimated the probability of cancer, as denoted by the fact that the estimates (*red circles* and *green triangles*) differed substantially from the theoretically expected probabilities ('*X*' symbols and the *diagonal*). If the subjects had estimated the probability correctly, all their estimates would overlap the X symbol in the given column. Instead, the subjects' estimates deviated substantially from the theoretically correct estimates. Across all subjects, the maximum and minimum difference between the reported vs. expected percent probabilities were 1.0 and − 0.21, respectively. The average difference was 0.33 ± 0.29 (standard deviation).

Second, the estimated values typically were *overestimates*, as denoted by the fact that most of the estimates were above the diagonal. The overestimates were highly significant (1-tailed paired *t*-test, $t = 26.60$, $df = 519$, $p < 2.2^{-16}$). This systematic bias straightforwardly indicates that the subjects failed to estimate the probabilities accurately. Intriguingly, the subjects' overestimates were significantly larger when the mammogram was deemed positive for cancer than when they were deemed negative (1-tailed *t*-test, $t = 8.61$, $df = 516.15$,

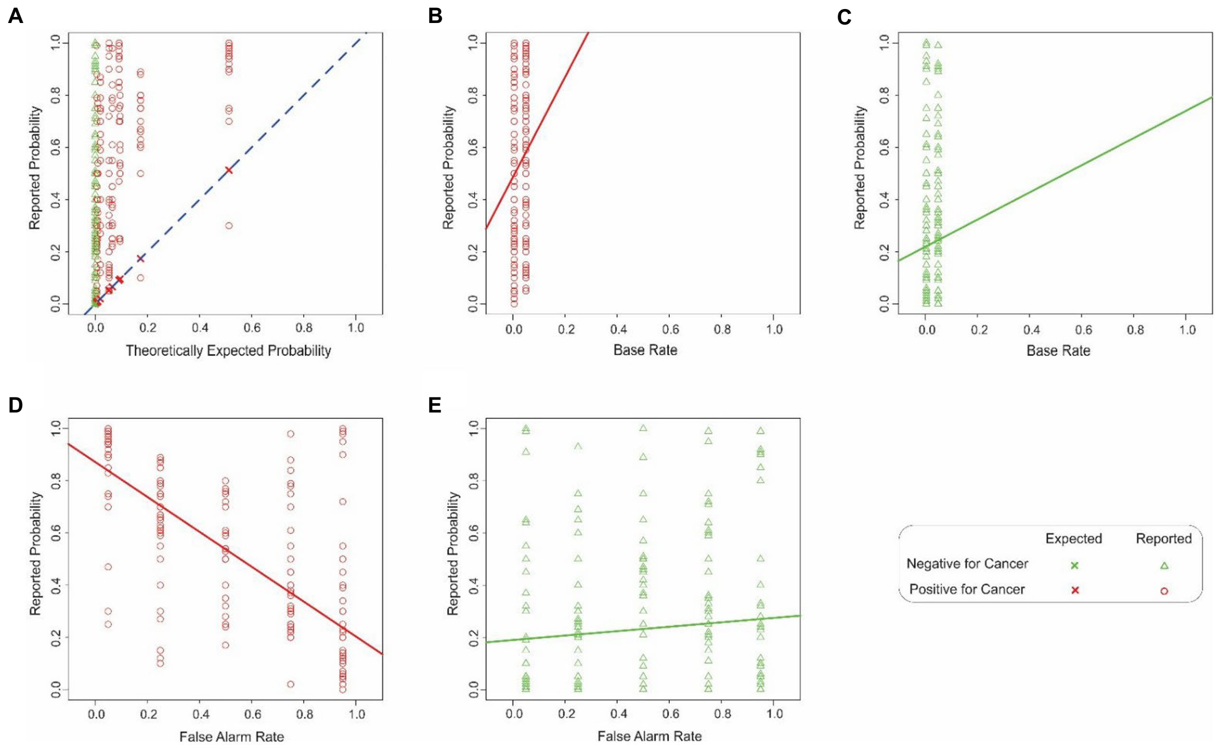


FIGURE 1
Estimation errors in Experiment 1. **(A)** Probability of cancer estimated by the subjects (*y-axis*) as a function of the corresponding theoretically expected probabilities (*x*-axis). Each *red circle* or *green triangle* denotes a single trial in which the hypothetical CAD system decided that the mammogram in question was positive or negative for cancer, respectively (see *legend* at *bottom right*). The '*X*' symbols and the *dashed diagonal* denote hypothetical scenarios where the subjects' estimated probability exactly matched the corresponding expected probability. The color of the plotting symbols (*red* vs. *green*) denote individual trials in which the CAD system determined that the given mammogram was positive or negative for cancer, respectively. The lines denote the best-fitting linear regression line in each case. **(B,C)** The interaction between the base rate and the binary decision of the system. The same plotting conventions as in panel A are used, except that in this panel, the estimated probability (*y-axis*) is plotted against the base rate (*x-axis*). For visual clarity, the data corresponding to the two decisions of the CAD system (mammogram positive or negative for cancer) are shown separately in panel d and e, respectively. In either panel, the *solid* and *dashed lines* denote best-fitting regression line. Panels **(D,E)** similarly show the interaction between the false alarm rate and the binary decision of the system, the estimated probability (*y-axis*) is plotted as a function of whether the CAD system decided that the mammogram was positive or negative for cancer (panel **D** or **E**), respectively. See text for details.

TABLE 1 Summary of regression modeling of the reported probabilities in Experiment 1.

Predictor variable in the initial model‡		Exploratory linear regression model				*lmg* value (% contribution to overall R^2)†*
		Estimated coefficient *β*	Standard error	*t* value	*p* value	
#	Name	A	B	C	D	E
1	Null model (intercept only)	0.20	0.04	5.19	2.99×10^{-7}	(N.A.)
2	Base rate of cancer in the cohort	−3.39	0.99	−3.39	0.69	2%
3	False alarm rate of the CAD system	0.03	0.06	0.57	0.57	19%
4	Binary decision of the system	0.64	0.05	13.75	$<2 \times 10^{-16}$	48%
5	Interaction of base rate & false alarm rate	1.82	1.45	1.25	0.21	0.4%
6	Interaction of base rate & binary decision	1.39	0.94	1.48	0.14	0.5%
7	Interaction of false alarm rate & binary decision	−0.75	0.07	−11.54	$<2 \times 10^{-16}$	31%

‡See Methods for additional details. †The model as a whole accounted for 45.59% of the variance (i.e., $R^2 = 0.4559$). *Model selection procedures retained variables # 2, 3, 4, 6, and 7 in the Final Model (not shown).

$p < 2.2^{-16}$). Together, these results suggest that the subjects were performing the task intuitively, rather than using systematic, logical reasoning.

To help quantify the extent to which the various explanatory (or predictor) variables contributed to the subjects' estimates, we constructed a generalized linear mixed model (GLMM), in which we included all three independent variables we varied in this experiment, along with their pairwise interactions as predictors (see Methods for details). This exploratory model (or 'Initial Model') is summarized in Table 1. We report the results about both the beta (or regression) coefficients β_i (columns A – D in Table 1) and the coefficients of determination R^2 (column E) of this model, because they both provide useful, but mutually distinct, types of information about the underlying data, as briefly outlined below.

The Initial Model is given by the relationship

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_7 x_7 + \varepsilon \qquad (3)$$

where $\hat{y}$ is the model's *estimates* of the values of the response variable (as opposed to the actual observed values y of the response variable); x_1 through x_7 are the seven predictor variables included in this model; β_1 through β_7 are the corresponding weight coefficients of the predictor variables; β_0 is the model offset; and ε is the error, so that $\varepsilon = y - \hat{y}$. That is, the β values are scaling coefficients that collectively specify the offset (in case of β_0) and the slope (in case of β_i) of the regression line that best fits the data. They determine the values of the estimates $\hat{y}$ directly as shown in Eq. 3, and are only indirectly related to actual observed values y. Thus, interpreting β values as representing the contribution of the predictor variables to the observed responses can be misleading to the extent to which $\hat{y}$ differs from y, especially when the observed responses are scattered widely about the regression line (Draper and Smith, 1998; Burnham et al., 2002; Aggarwal and Ranganathan, 2017). On the other hand, to the extent to which $\hat{y}$ is correlated with y, the best coefficients do provide useful information about the contribution of the predictor variables to the observed response. After all, beta coefficients are used for this purpose extensively in psychology, neuroscience, econometrics, etc. (Friston, 2007; Gravetter and Wallnau, 2017; Laha, 2019; Hashimzade and Thornton, 2021). Regression coefficients are also essential for model selection, i.e., for determining which predictor variable/s make a statistically significant contribution to $\hat{y}$, and therefore should be retained in the parsimonious 'Final Model' of the data (Draper and Smith, 1998; Burnham et al., 2002).

On the other hand, for the purposes of measuring the contribution of the various predictor variables to the observed responses, metrics that reflect the *statistical correlation* between x and y are more appropriate (Draper and Smith, 1998; Burnham et al., 2002). For this purpose, we use the well-established *lmg* statistic (column E, Table 1), which denotes the percent contribution of the given predictor variable to the observed responses [see General Methods for details; also see Lindeman et al. (1980)].

An examination of the Initial Model indicated that the base rate of cancer in the patient cohort made a statistically insignificant contribution to the model (row 2). This straightforwardly suggests that the subjects underweighted, i.e., neglected, the base rate in making their decisions.

As noted above, many previous studies have suggested that base rate neglect occurs because subjects not only underweight the base rate but also simultaneously attach too much importance to the 'individuating information', i.e., the binary decision of the system about the individual mammogram in the present case (Kahneman and Tversky, 1973). The contribution of the binary decision factor to the subject's responses was indeed highly significant (row 4).

Note, however, the fact the binary decision *contributed significantly* does not necessarily mean that it *overcontributed*, i.e., that it contributed more than its share to the model. If, for the sake of argument, the subjects attached exactly correct weight to this factor (i.e., neither underweighted nor overweighted it), the contribution of this factor could still be statistically significant. Thus, statistically significant contribution does not necessarily mean overcontribution/overweighting. We will revisit this issue below using additional analyses.

The false alarm rate by itself did not make a statistically significant contribution to the model at the level of 95% confidence in this model (row 3). However, the interaction between the false alarm rate and the binary decision of the system did (row 7). That is, the false alarm rate affected the subjects' reports differentially depending on the binary decision of the system. This interaction is reflected in the fact that the best-fitting regression lines are different in Figures 1B,C. In other words, the subjects' estimates covaried with the false alarm rates when the CAD system decided that the individual mammogram was positive for cancer (Figure 1B), but not when the mammogram was deemed negative for cancer (Figure 1C), a finding confirmed by a 2-way analysis of covariance (ANCOVA; false alarm rate x binary decision; $p < 0.05$ for both factors and their interaction, not shown). It is also worth noting that the estimated coefficient of this interaction factor was negative (Estimated Coefficient $= -0.75$; row 7, column A of Table 1), indicating that the overall effect of this factor was to reduce the estimated probabilities. By contrast, the binary decision had an effect of a comparable magnitude, but of opposite sign (Estimated Coefficient 0.64; row 4, column A). Thus, the overall estimates of the responses reflect a complex interplay of multiple, sometimes counteracting, factors.

The coefficient of determination of the Initial Model, R^2, was 0.4559, indicating that the seven predictor variables in this model collectively accounted for about 46% of the variance in the observed responses (see Footnote to Table 1). This raises the issue of how much each predictor variable contributed to this 45.59%. As noted above, previous studies have variously attributed such estimation errors to neglect or overweighting (i.e., where a given variable contributes less or more than its share) of the various underlying variables. Therefore, it is crucially important to determine the relative contribution of each of the variables in the present case.

To do this, we used the well-established method of the *lmg* index (Lindeman, Merenda and Gold index; Lindeman et al., 1980); *lmg* index (see Methods). The *lmg* index is a principled method for decomposing a given R^2 value into the relative contributions from the various independent variables. It is equivalent to, but distinct from, partial *R*, and offers some advantages over the latter (Lindeman et al., 1980). Under the null hypothesis (i.e., default assumption) that all six variables contributed equally to the overall fit, i.e., that the subjects weighted each variable appropriately, each variable is expected to contribute $1/6 \approx 16.67\%$ to the R^2 value, i.e., explained variance or the model fit (see General Methods). The actual contributions are shown in column E of Table 1. The most important contributor to the subjects' estimates was the individuating information, and it accounted for 48% of the R^2 value (row 4, column E). Similarly, the false alarm-binary decision interaction and the false alarm rate, respectively, accounted for about 31% and 19% of the R^2 value. Thus, the subjects nominally overweighted each of these three variables (also see below). On the other hand, subjects underweighted, or neglected, the remaining three variables (rows 2, 5, and 6, column E).

The above results are based on the Initial Model that included all seven of the original predictor variables. It is well known that including too few or too many predictor variables can lead to modeling artifacts (Draper and Smith, 1998; Fox and Fox, 2016; Hegdé, 2021); therefore, it is desirable to optimally balance model complexity with model fit (Draper and Smith, 1998; Fox and Fox, 2016), i.e., to determine the most parsimonious model that accounts for the most amount of observed data. We used standard model selection procedures to determine such a parsimonious model for this experiment, which we will refer to as the 'Final Model' [see Methods for details; also see Draper and Smith (1998) and Fox and Fox (2016)]. The Final Model retained just five predictor variables: (i) base rate, (ii) false alarm rate, (iii) binary decision, (iv) base rate-binary decision interaction, and (v) the false alarm-binary decision interaction, indicating that only these five factors had a statistically significant effect on the subjects' estimates (rows 2, 3, 4, 6, and 7 in Table 1; also see footnote to Table 1).

The aforementioned *lmg* value analysis did not address whether or not the relative contributions of the various variables were statistically significant. For instance, the fact that base rate is retained in the Final Model as a significant predictor of the outcome is noteworthy, but does it mean that the subjects do not significantly neglect base rate at all, i.e., do they give base rate its due weight in arriving at their estimates?

To help address such issues, we calculated the Relative Contribution Index (RCI) for each of the five predictors in the Final Model (see General Methods for details). The RCI value for a given predictor is essentially its *lmg* value adjusted for the level of randomness in the empirical data. That is, the RCI value of the predictor measured the extent to which the actual *lmg* value for a given predictor compares to the *lmg* value for that predictor expected from random chance (see General Methods for details), where a value of 1.0 indicated that the predictor contributed exactly the expected amount to the outcome, and values >1 and < 1, respectively, indicate correspondingly higher or lower contribution than the contribution expected for that predictor. The RCI values for the five predictors in the Final Model are shown in Figure 2.

The RCI value for the base rate factor was 0.07 (predictor 2 in Figure 2), well below the RCI value expected from random (*solid line* in Figure 2), indicating that the subjects indeed underweighted the base rate substantially. However, this RCI value was still within the

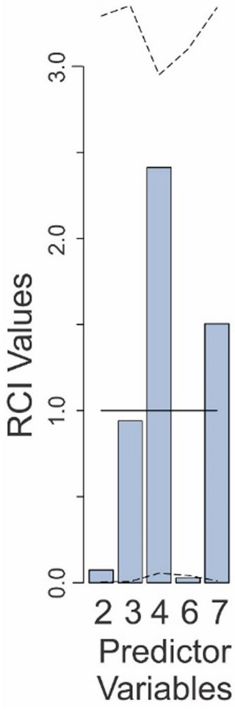


FIGURE 2
Relative contributions of various predictors to the fit of the Final Model in Experiment 1. The predictor variables are those that are retained in the Final Model and are numbered as in Table 1. The *solid line* denotes the expected contributions of the various predictors. The *dashed lines* denote the upper and lower 95% confidence intervals (uncorrected), empirically determined from the data. See text for details.

95% confidence interval (CI: 0.006–3.39; see *dashed lines* in Figure 2), indicating that the underweighting was not statistically significant at 95% confidence level.

On the other hand, the underweighting of the false alarm-binary decision interaction was indeed statistically significant (predictor 6; RCI = 0.027; CI: 0.042–3.1). The false alarm rate contributed slightly less than the expected amount (predictor 3; RCI = 0.94; CI: 0.007–3.35). The binary decision of the system, as well as the false alarm-binary decision interaction both made larger-than-average contributions to the outcome (RCI values of 2.4 and 1.5, respectively), although this was not statistically significant (CIs: 0.06–3.1 and 0.01–3.34, respectively). When results were corrected for multiple comparisons (see Methods), the contribution of none of the variables remained statistically significant (not shown). Collectively, these results show that while subjects substantially underweighted (or neglected) some variables and overweighted some others, while only the binary decision-dependent neglect of the base rate was statistically significant.

Discussion

The above results show that naive subjects significantly overestimate the probability of cancer. They also identify multiple sources of these estimation errors, including the overweighting of some factors such as the binary decision of the CAD system, and underweighting other factors such as the base rate. In this regard, our results confirm and extend the previous studies to the present task.

These results are novel in three main respects. First, our results demonstrate that both underweighting and overweighting contribute to the estimation errors. Second, our results identify two additional contributing factors, namely the base rate-dependent neglect of false alarm rates, and the binary decision-dependent overweighting of the false-alarm rate. Previous studies have reported the neglect of false alarm rates (which the reports referred to as 'miss rate neglect') in the context of legal judgements (Dahlman et al., 2016; Dahlman and Mackor, 2019). But to our knowledge, our study is the first to report the contribution of the above two factors and to report such *conditional* underweighting/overweighting. Finally, we demonstrate that the underweighting or overweighting of *individual* factors is not statistically significant although the collective effect of all the factors together is a significant overestimation of cancer probabilities, as noted above.

Our preliminary studies indicate that highly trained, practicing radiologists also commit similar errors in the same task (Branch et al., 2022). Thus, overestimation of the probabilities was not attributable to the fact that the subjects in the present experiment were untrained professionals.

Experiment 2: Estimating the probability of drunkenness based on breathalyzer evidence

The results of Experiment 1 raise the issue of whether and to what extent they are idiosyncratic to the particular task that the subjects were carrying out. For instance, it may be that subjects tended to overestimate the probability of chance because of the perceived costs of underestimating the cancer risk. To the extent this is true, the pattern of estimation errors would change if the same problem was posed in a different problem context where costs of various types of errors (e.g., false positives and false negatives) were different. We tested this hypothesis in the present experiment by keeping all the parameters exactly the same, but using them to pose a different problem, namely estimating the probability of drunk driving based on the outcome of individual breathalyzer tests.

Methods

This experiment was identical to Experiment 1 except for the task. In this experiment, the subjects were told that the four items of information pertained to a breathalyzer system that was used for testing motorists for drunk driving. Specifically, the four parameters were:

1. The base rate of drunk driving in the given cohort of motorists,
2. The hit rate of a hypothetical breathalyzer system,
3. The false alarm rate of the system, and
4. The binary decision of the system (positive or negative for drunkenness) for a given motorist from the given cohort of motorists. No other data were provided to the subjects. Twelve subjects (eight women; mean age, 19.58 years ±1.44) participated in this experiment.

Results

The reported probabilities in this experiment (Figure 3A) were collectively indistinguishable from the results in Experiment 1 (two-tailed *t*-test, $p > 0.05$; not shown), indicating that changing the task did not result in large-scale changes in the reported probabilities overall. The subjects' reported estimates deviated substantially from the theoretically expected probabilities (Figure 3A). Across all subjects, the maximum and minimum difference between the reported vs. expected percent probabilities were 1.0 and − 0.32, respectively. The average difference was 0.33 ± 0.30. The subjects also significantly overestimated the probabilities (1-tailed paired *t*-test, $t = 23.47$, $df = 459$, $p < 2.2^{-16}$). Also, magnitude of the overestimations was significantly larger when the mammogram was deemed positive for cancer than when it was deemed negative (one-tailed *t*-test, $t = 8.69$, $df = 457.61$, $p < 2.2^{-16}$).

With the exception of the base rate-binary decision interaction, all of the predictors that contributed significantly to the outcome in Experiment 1 also did so in this experiment. The nature of the false alarm-binary decision interaction was similar to that in Experiment 1, so that the subjects took the false alarm rate into account when the breathalyzer system determined that the motorist was drunk, but not when the system decided otherwise (Figures 3B,C; ANCOVA; false alarm rate x binary decision; $p < 0.05$ for both factors and their interaction, not shown). This interaction and the binary decision variable made a

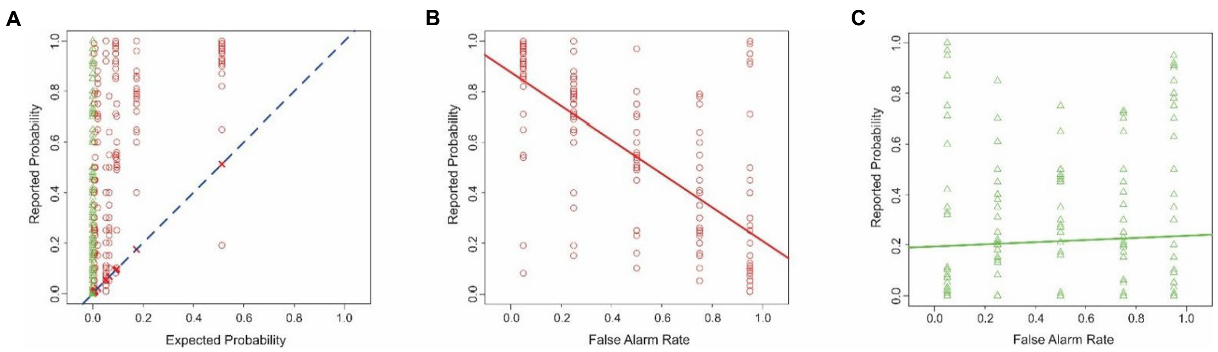


FIGURE 3
Estimation errors in Experiment 2. The data are plotted according to the conventions used in Figure 1. **(A)** Probability of drunk driving estimated by the subjects plotted as a function of the corresponding theoretically expected probability. **(B,C)** Interaction between the false alarm rate and the binary decision of the breathalyzer. See text for details.

TABLE 2 Summary of regression modeling of the reported probabilities in Experiment 2.

Predictor variable in the initial model[‡]		Exploratory linear regression model				*lmg* value (% contribution to overall R^2)[†*]
		Estimated coefficient β	Standard error	*t* value	*p* value	
#	**Name**	A	B	C	D	E
1	Null model (intercept only)	0.15	0.04	3.53	4.6×10^{-4}	(N.A.)
2	Base rate of drunk driving in the cohort	1.57	1.08	1.45	0.15	0.6%
3	False alarm rate of the breathalyzer system	0.05	0.07	0.77	0.44	20%
4	Binary decision of the system	0.72	0.05	13.91	$<2 \times 10^{-16}$	53%
5	Interaction of base rate & false alarm rate	−0.33	1.59	−0.21	0.83	0.01%
6	Interaction of base rate & binary decision	−1.19	1.04	−1.15	0.25	0.3%
7	Interaction of false alarm rate & binary decision	−0.71	0.07	−9.91	$<2 \times 10^{-16}$	26%

[‡]See Methods for additional details. [†]The model as a whole accounted for 45.28% of the variance (i.e., $R^2 = 0.4528$). *Model selection procedures retained variables # 2, 3, 4, and 7 in the Final Model (not shown).

significant contribution to the outcome in the Initial Model (Table 2, rows 4 and 7, column D). These two variables and two additional variables, including the base rate and the false alarm rate, were retained in the Final Model (see footnote to Table 2).

Results of the RCI analysis showed that all four factors retained in the Final Model contributed substantially to the final outcome, and the under/overweighting of none of the contributions was statistically significant, even without correction for multiple comparisons (Figure 4).

Discussion

One notable difference between the results of this experiment from those in Experiment 1 was that the binary decision-base rate interaction was retained in the final mode in Experiment 1, but not in this experiment. Other than that, the results of this experiment were similar to those of Experiment 1. Most notably, our analyses showed no evidence for significant neglect or overweighting of any other variables in the present experiment, either. These results indicate that changing the task had little or no effect on the estimation of probabilities.

Experiment 3: Estimating the probability of an enemy sniper based on evidence from drone reconnaissance system

Methods

This experiment was identical to Experiments 1 and 2, except for the task. In this experiment, the subjects were told that the four items of information pertained to a military drone system that was used to reconnoiter a combat scene for enemy snipers. Specifically, the four parameters were:

1. The prevalence of enemy snipers in the given theater of combat operations,
2. The hit rate of the drone system,

3. The false alarm rate of the system, and
4. The binary decision of the system (positive or negative for the presence of an enemy sniper) for a given combat scene from the given theater of operations. No other data were provided to the subjects. The subjects had to estimate the probability that an enemy sniper was present in the scene of combat. Thirteen subjects (nine women; mean age, 20.23 years ±2.39) participated in this experiment.

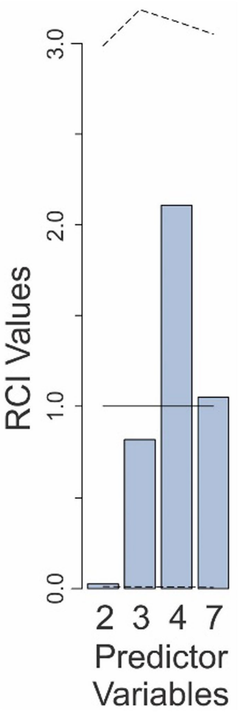


FIGURE 4
Relative contributions of various predictors to the fit of the Final Model in Experiment 2. The predictor variables are those that are retained in the Final Model and are numbered as in Table 2. The *solid line* denotes the expected contributions of the various predictors. The *dashed lines* denote the upper and lower 95% confidence intervals (uncorrected), empirically determined from the data. See text for details.

Results and discussion

The subjects' responses in this experiment (Figure 5A) were indistinguishable from the results in Experiment 1 by a two-tailed *t*-test, $p > 0.05$; not shown.

The subjects' reported estimates deviated substantially from the theoretically expected probabilities (Figure 5A). Across all subjects, the maximum and minimum difference between the reported vs. expected percent probabilities were 0.95 and − 0.46, respectively. The average difference was 0.33 ± 0.29. The subjects significantly overestimated the probabilities (1-tailed paired *t*-test, $t = 23.55$, $df = 419$, $p < 2.2^{-16}$). This systematic bias straightforwardly indicates that the subjects failed to estimate the probabilities accurately. Intriguingly, the subjects' overestimates were significantly larger when the combat scene was deemed positive for enemy sniper than when it was deemed negative (one-tailed *t*-test, $t = 5.94$, $df = 412.77$, $p = 3.08^{-09}$; also see Figures 5B,C).

In this experiment, only three predictor variables were retained in the Final Model: false alarm rate of the drone system, binary decision of the system, and the false alarm-binary decision interaction (see footnote to Table 3). Recall that all three variables were also retained in Experiments 1 and 2, but two additional predictors were retained in those experiments that were not retained in this experiment, raising the possibility that the variables in question were over/underweighted in the present experiment. However, the over/underweighting of none of the variables was statistically significant in this experiment, even without correction for multiple comparisons (Figure 6).

Experiment 4: Estimating the probability of an enemy sniper based on evidence from drone reconnaissance system (version 2)

In Experiments 1–3, only the problem scenario differed across the experiments, but numerical values of the four probabilistic parameters remained the same. This design helped us address the important issue of the extent to which the estimation errors vary or remain the same depending on the problem scenario. The present experiment took the

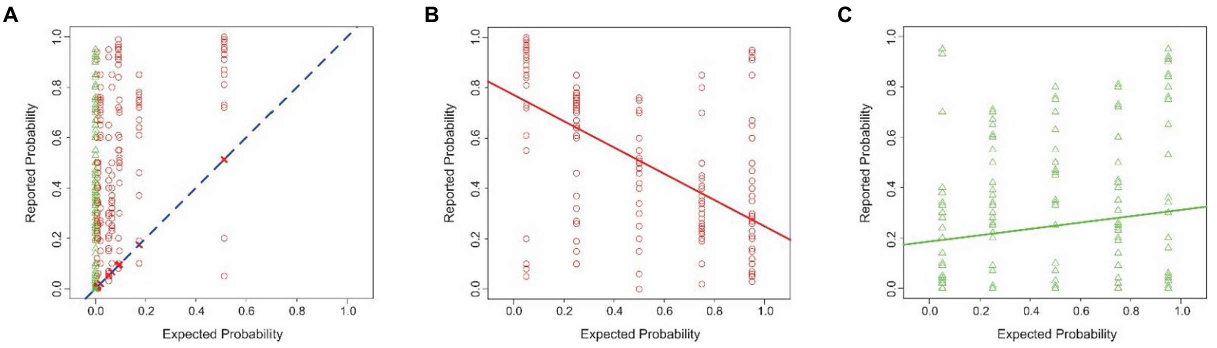


FIGURE 5
Estimation errors in Experiment 3. The data are plotted according to the conventions used in Figure 1. **(A)** Probability of enemy sniper estimated by the subjects is plotted here as a function of the corresponding theoretically expected probability. **(B,C)** Interaction between the false alarm rate and the binary decision of the reconnaissance drone. See text for details.

TABLE 3 Summary of regression modeling of the reported probabilities in Experiment 3.

Predictor variable in the initial model[‡]		Exploratory linear regression model				*lmg* value (% contribution to overall R^2)[†]*
		Estimated coefficient β	Standard error	*t* value	*p* value	
#	Name	A	B	C	D	E
1	Null model (intercept only)	0.16	0.05	3.46	5.9×10^{-4}	(N.A.)
2	Base rate (i.e., prevalence of snipers in the given theater of combat)	0.87	1.19	0.73	0.46	0.8%
3	False alarm rate of the reconnaissance drone system	0.14	0.07	1.88	0.06	13%
4	Binary decision of the system	0.58	0.06	10.31	$<2 \times 10^{-16}$	52%
5	Interaction of base rate & false alarm rate	−0.46	1.75	−0.26	0.79	0.04%
6	Interaction of base rate & binary decision	0.14	1.14	0.13	0.90	0.008%
7	Interaction of false alarm rate & binary decision	−0.65	0.08	−8.23	2.51×10^{-15}	34%

[‡]See Methods for additional details. [†]The model as a whole accounted for 32.55% of the variance (i.e., $R^2 = 0.3255$). *Model selection procedures (not shown) retained variables # 3, 4, and 7 in the Final Model (not shown).

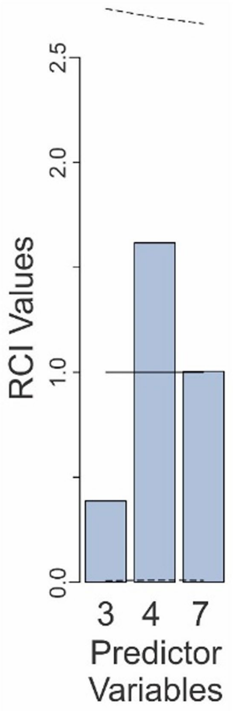


FIGURE 6
Relative contributions of various predictors to the fit of the Final Model in Experiment 3. The predictor variables are those that are retained in the Final Model and are numbered as in Table 3. The *solid line* denotes the expected contributions of the various predictors. The *dashed lines* denote the upper and lower 95% confidence intervals (uncorrected), empirically determined from the data. See text for details.

complementary approach of varying the parameter values while keeping the problem scenario unchanged.

This tweak in the experimental design allowed us to test additional hypotheses about the underlying phenomenon. For instance, subjects in Experiments 1–3 showed a conditional neglect of the false alarm rate, wherein subjects underweighted the false alarm rate differently based on the binary decision of the system. The present experiment was designed to test the hypothesis that the subjects show a similar conditional neglect of the *hit rate.* A second hypothesis is that all other things being equal, subjects attach more weight to the hit rate than to the false alarm rate.

Methods

This experiment was identical to Experiment 3, except in two respects: To help better characterize the effect of varying the false alarm rates, we increased the number of possible hit rates to three (0.05, 0.5, and 0.95), so that the hit rate during any given trial was randomly drawn from these three values. Second, the false alarm rate during any given trial was drawn from the palette of the same three values (i.e., 0.0, 0.05, 0.5, and 0.95). As alluded to above, the problem scenario remained the same as in Experiment 3, so that the subjects estimated the probability that an enemy sniper was present in the scene of combat. Seven subjects (five women and one non-binary person; mean age, 27.71 years ±2.43) participated in this experiment.

Results and discussion

The reported probabilities in this experiment (Figure 7A) were collectively indistinguishable from the results in Experiment 1 (two-tailed *t*-test, $p > 0.05$; not shown), indicating that changing the task did not result in large-scale changes in the reported probabilities overall. The subjects' reported estimates deviated substantially from the theoretically expected probabilities (Figure 7A). Across all subjects, the maximum and minimum difference between the reported vs. expected percent probabilities were 0.96 and − 0.97, respectively. The average difference was 0.19 ± 0.40. The subjects significantly overestimated the probabilities (1-tailed paired *t*-test, $t = 14.81$, $df = 1{,}007$, $p < 2.2^{-16}$). This systematic bias straightforwardly indicates that the subjects failed to estimate the probabilities accurately. However, in contrast to the results obtained in Experiments 1–3, the subjects' overestimates were statistically indistinguishable between the combat scene was deemed positive for enemy sniper than when it was deemed negative (one-tailed *t*-test, $t = –0.59$, $df = 890.22$, $p = 0.72$).

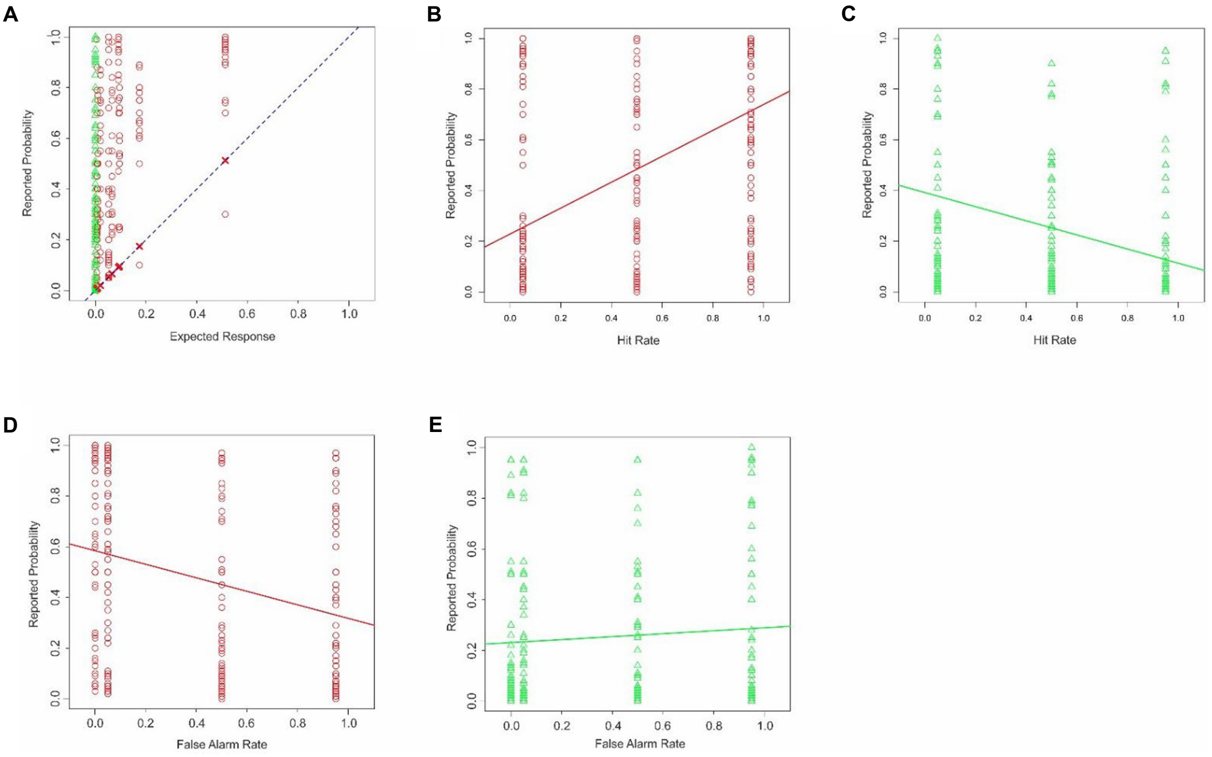


FIGURE 7
Estimation errors in Experiment 4. The data are plotted according to the conventions used in Figure 1. **(A)** Probability of enemy sniper estimated by subjects is plotted in this Figure as a function of the corresponding theoretically expected probability. **(B,C)** Interaction between the hit rate and the binary decision of the reconnaissance drone. **(D,E)** Interaction between the false alarm rate and the binary decision of the drone. See text for details.

As noted above, unlike in Experiments 1–3, the hit rate was varied in this experiment. This manipulation revealed a new interaction, namely the conditional neglect of hit rates, wherein the subjects underweighted the hit rate of the drone system based on the binary decision of the system (Figures 7B,C; ANCOVA; hit rate x binary decision; $p < 0.05$ for both factors and their interaction, not shown). The subjects also showed a conditional neglect of the false alarm rate (Figures 7D,E; ANCOVA; false alarm rate x binary decision; $p < 0.05$ for both factors and their interaction, not shown).

In this experiment, six different predictor variables were retained in the Final Model: base rate, hit rate, false alarm rate, binary decision of the system, and two interaction factors: the hit-rate binary decision interaction and the false alarm rate-binary decision (rows 2–7 in Table 4; also see footnote to Table 4). The two factors involving hit rate retained in this experiment were not available in Experiments 1–3.

RCI analysis (Figure 8) showed that, of the six factors retained in the Final Model in this experiment, the relative contribution of only two—binary decision and the hit rate-binary decision interaction (predictors 5 and 6, respectively)—were significantly outside the uncorrected 95% confidence intervals. However, only the hit rate-binary decision interaction factor survived the correction for multiple comparisons, indicating that the relative contribution of this factor was significantly smaller than expected. That is, judging by the RCI analysis, this factor can be reasonably deemed to be significantly neglected. That is, the subjects' failure to properly weight the individuating information as a function of the hit rate was statistically significant.

While one may be tempted to claim that the hit rate-binary decision interaction factor in this experiment was the only factor in our entire study to be significantly over/underweighted, doing so would be unwise. This is because making this comparison would require correction for this extended multiple comparison, in which this factor does not survive.

General discussion

Generalizability of probability estimation errors

Several aspects of the estimation errors were common to all four experiments in our study. First of all, subjects failed to make accurate judgements in each experiment. Second, the judgement errors were large, and varied widely from the theoretically expected estimations. Finally, the estimation errors represented significant overestimations in all four experiments.

It is also noteworthy that the overall pattern of errors was statistically indistinguishable across the four experiments (one-way ANOVA; $p = 0.98$ for the between-experiment factor; data not shown), even though the tasks and/or underlying probabilistic parameters varied across the experiments. This indicates that the errors were a

TABLE 4 Summary of regression modeling of the reported probabilities in Experiment 4.

Predictor variable in the initial model‡		Exploratory linear regression model				*lmg* value (% Contribution to overall R^2)†*
		Estimated coefficient β	Standard error	*t* value	*p* value	
#	Name	A	B	C	D	E
1	Null model (intercept only)	0.37	0.04	9.74	$<2\times10^{-16}$	(N.A.)
2	Base rate (i.e., prevalence of snipers in the theater of combat)	−0.01	0.92	−0.01	0.99	0.05%
3	Hit rate of the reconnaissance drone system	−0.28	0.05	−5.11	3.90×10^{-07}	4.34%
4	False alarm rate of the system	0.03	0.06	0.59	0.55	3.9%
5	Binary decision of the system	−0.05	0.04	−1.10	0.27	32%
6	Interaction of hit rate & binary decision	0.79	0.05	15.39	$<2\times10^{-16}$	50%
7	Interaction of false alarm rate & binary decision	−0.33	0.05	−6.64	5.14×10^{-11}	9.37%
8	Interaction of hit rate & base rate	−0.30	1.14	−0.26	0.80	0.02%
9	Interaction of false alarm rate & base rate	0.62	1.09	0.57	0.57	0.07%
10	Interaction of base rate & binary decision	0.24	0.84	0.29	0.77	0.02%
11	Interaction of false alarm rate & hit rate	0.02	0.07	0.23	0.82	0.01%

‡See Methods for additional details. †The model as a whole accounted for 32.06% of the variance (i.e., $R^2 = 0.3206$). *Model selection procedures retained variables # 2, 3, 4, 5, 6, and 7 in the Final Model (not shown).

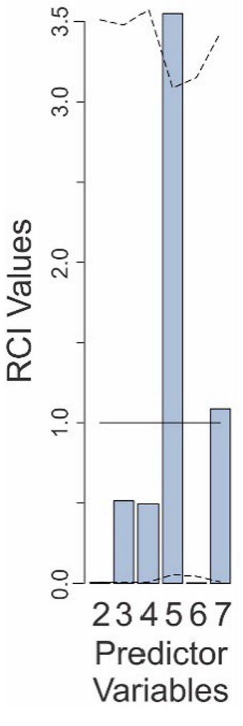


FIGURE 8
Relative contributions of various predictors to the fit of the Final Model in Experiment 4. The predictor variables are those that are retained in the Final Model and are numbered as in Table 4. The *solid line* denotes the expected contributions of the various predictors. The *dashed lines* denote the upper and lower 95% confidence intervals (uncorrected), empirically determined from the data. See text for details.

general feature of the estimation problem used in our study, and generalized across the tasks and the experimental parameters we used. This straightforwardly suggests that the subjects are unlikely to have used grossly different mental strategies for estimating the probability of the outcome.

Factors that contribute significantly to estimation errors

Our analyses identified multiple contributing factors for the errors. Both the similarities and differences among these factors across experiments are noteworthy. On the one hand, factors such as overweighting of the binary decision (i.e., individuating information) and the underweighting (or neglect) of the base rate were major contributing factors to the errors across all four experiments. These findings are consistent with the large body of earlier studies using this task paradigm as well as other task paradigms that have attributed the errors variously to one or both of these factors (Kahneman and Tversky, 1973; Fischhoff and Bar-Hillel, 1984; Thompson and Schumann, 1987; Koehler, 1996; Baratgin and Noveck, 2000; Fantino, 2004; Barbey and Sloman, 2007; Dahlman et al., 2016; Sanborn and Chater, 2016; also see Koehler (1996) and the accompanying commentaries).

On the other hand, some factors made statistically significant contributions to the outcome in some experiments and not others. For instance, the interaction between the base rate and binary decision was evident in Experiments 1 and 4, but not in the other two experiments. Further studies are needed to address the issue of why exactly the relative contributions of factors differed across tasks.

Our study also identified several additional contributing factors that, to our knowledge, have not been previously reported. The most notable among these are the factor interactions. We identified many statistically significant interactions across the experiments (Tables 1–4). Of particular note is the interaction between false alarm rates and binary decisions, whereby the subjects attach different weight to the false alarm rates depending on the binary decision (and vice versa). Intriguingly, this interaction was statistically significant in all four of the experiments. To our knowledge, such interaction (or, 'conditional') effects have not been reported before, although previous studies have reported a neglect of the false alarm rates (sometimes referred to as the "miss rate neglect") in the context of legal decision-making (Thompson and Schumann, 1987; Dahlman et al., 2016).

Causes of the errors are significant as a group, not individually

Our results show that the aforementioned factors, as a group, do account for a significant amount of the subjects' estimates of the probabilities. Depending on the experiment, the independent variables collectively account for about 30 to 45% of the variance, depending on the experiment. Of course, this is unsurprising, because in any study, the independent variables would be expected to account for the response variable/s, to the extent that the former have any bearing on the latter.

What is surprising about our results is the fact that, by a principled set of criteria, none of the contributing factors by itself significantly accounts for the outcome (see below for caveats). As noted earlier, many previous studies have attributed these errors variously to the neglect of base rates, overweighting of the evidence for the individual event, or both [for an overview, see Koehler (1996) and the accompanying commentaries]. The collective effect of these studies has been substantial, in that the estimation errors in question have come to be widely known as the base rate neglect, base rate fallacy, or base rate bias (Kahneman et al., 1982; Gigerenzer and Hoffrage, 1995; Fantino, 2004). Some previous studies have attributed these errors in other contexts, such as legal decision-making, to the so-called fallacy of the transposed conditional or the prosecutor's fallacy, where the subjects conflate $p(A|B)$ for $p(B|A)$ (Thompson and Schumann, 1987), or to the neglect of false alarm rates, sometimes referred to as the miss rate neglect (Dahlman et al., 2016).

While these studies provide empirical evidence that subjects do underweight (or conflate, in case of the prosecutor's fallacy) the relevant variables, they do not show that these factors by themselves fully account for the errors. In fairness to such studies, few of them expressly claim that factors such as base rate neglect fully account for the errors. However, factors such as base rate neglect have somehow come to be thought of as sufficient explanations for the underlying errors.

Our study successfully reproduces the estimation errors, and demonstrates that such claims are misleading at best, because they obscure the complexities of the underlying phenomena. On the one hand, our results unambiguously show that subjects make large, systematic errors, which straightforwardly means that the subjects fail to correctly weight the various underlying factors to one degree or another. This in turn raises the question of what level of underweighting constitutes neglect. For instance, if the subject underweights the base rate factor by, say, an average of 10%, can this legitimately be deemed base rate neglect? Previous studies have generally avoided this issue. This study takes the position that underweighting can be deemed neglect if it is statistically significant, i.e., if the weight is significantly lower than that expected from random chance. Similarly, a given factor can be considered overweighted if it is significantly larger than that expected from random chance. These clearly are principled criteria, but by no means the only possible ones (see below).

Some important caveats

In addition to the various methodological caveats noted in context throughout this report, a few caveats are especially worth highlighting here: First, as alluded to above, our study focused narrowly on the question of whether and to what extent the observed biases can be accounted for by the overweighting or neglect of *individual* factors, as implied by the earlier studies. For this reason, our study remained advisedly agnostic about a variety of important, vigorously debated questions in the field. Chief among these are issues such as (i) how the subjects arrive at their estimates (Kahneman et al., 1982; Koehler, 1996), (ii) approaches to reducing the estimation errors and efficacy of these errors (Hoffrage and Gigerenzer, 1998; Uhlmann et al., 2007; Raab and Gigerenzer, 2015), (iii) the methodological and conceptual validity and usefulness of formulating and studying the probability estimations within the Bayesian framework (Koehler, 1996; Baratgin and Noveck, 2000; Fantino, 2004; Barbey and Sloman, 2007; Sanborn and Chater, 2016), and (iv) whether and to what extent our findings generalize to other task paradigms of probability estimation (e.g., Gigerenzer, 1996; Koehler, 1996), or when tested using a larger number of disparate problem scenarios. Further studies are needed to address each of these questions.

In addition to the various methodological caveats noted in context throughout this report, two caveats are especially worth highlighting here: First, as its name indicates, GLMM assumes a *linear* relationship between the predictor variables and the response variable. While our GLMMs did indeed satisfy the underlying assumptions (data not shown), this does not by itself prove that the actual underlying relationship is linear. Indeed, it remains possible that there exists an unknown non-linear relationship that accounts for the observed data even better.

Concluding remarks

A main significance of our study is that it calls into question the validity of attributing the probability estimation errors to individual factors. But in a larger sense, the significance of our study is that it proposes a set of reasonable criteria and methods for evaluating the potential causes of probability estimation errors.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Institutional Review Board (IRB) of Augusta University, Augusta, GA, United States. The patients/participants provided their written informed consent to participate in this study.

Author contributions

FB and JH jointly designed the study, collected and analyzed the data, and wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by grant # W911NF-15-1-0311 from the Army Research Office (ARO) to JH.

Acknowledgments

Erin Park, Khadeja Cohen, and Matthew Williams provided technical assistance during some sessions of data collection. Havilah Ravula provided helpful assistance during manuscript preparation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1132168/full#supplementary-material

References

Aggarwal, R., and Ranganathan, P. (2017). Common pitfalls in statistical analysis: linear regression analysis. *Perspect. Clin. Res.* 8, 100–102. doi: 10.4103/2229-3485.203040

Baratgin, J., and Noveck, I. A. (2000). Not only base rates are neglected in the engineer-lawyer problem: an investigation of reasoners' underutilization of complementarity. *Mem. Cogn.* 28, 79–91. doi: 10.3758/BF03211578

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241–254. doi: 10.1017/S0140525X07001653

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologia.* 44, 211–233. doi: 10.1016/0001-6918(80)90046-3

Bar-Hillel, M. (1991). Commentary on Wolford, Taylor, and Beck: the conjunction fallacy? *Mem. Cogn.* 19, 412–414. doi: 10.3758/BF03197146

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Statistical Society B.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Berridge, D., and Crouchley, R. (2011). *Multivariate generalized linear mixed models using R.* Boca Raton, FL: CRC Press. *xxiii*, 280.

Bono, R., Alarcon, R., and Blanca, M. J. (2021). Report quality of generalized linear mixed models in psychology: a systematic review. *Front. Psychol.* 12:666182. doi: 10.3389/fpsyg.2021.666182

Branch, F., Williams, K. M., Santana, I. N., and Hegdé, J. (2022). How well do practicing radiologists interpret the results of CAD technology? A quantitative characterization. *Cognitive Research: Principles and Implications* 7:52. doi: 10.1186/s41235-022-00375-9

Burnham, K. P., Anderson, D. R., and Burnham, K. P. (2002). *Model selection and multimodel inference: A practical information-theoretic approach. 2nd* New York: Springer. *xxvi*, 488.

Cooper, H. M.American Psychological Association. (2018). *Reporting quantitative research in psychology: How to meet APA style journal article reporting standards. Second* Washington, DC: American Psychological Association. *vii*, 217.

Dahlman, C., and Mackor, A. R. (2019). Coherence and probability in legal evidence. *Law, Probability and Risk.* 18, 275–294. doi: 10.1093/lpr/mgz016

Dahlman, C., Zenker, F., and Sarwar, F. (2016). Miss rate neglect in legal evidence. *Law, Probability and Risk.* 15, 239–250. doi: 10.1093/lpr/mgw007

Dean, C. B., and Nielsen, J. D. (2007). Generalized linear mixed models: a review and some extensions. *Lifetime Data Anal.* 13, 497–512. doi: 10.1007/s10985-007-9065-x

Draper, N. R., and Smith, H. (1998). *Applied regression analysis. 3rd* New York: Wiley. *xvii*, 706.

Eddy, D. M. (1982). "Probabilistic reasoning in clinical medicine: problems and opportunities" in *Judgment under uncertainty: Heuristics and biases.* eds. D. Kahneman and A. PaulTversky (New York, NY: Cambridge University Press), 249–267.

Eddy, D. M. (2005). Evidence-based medicine: a unified approach. *Health Aff (Millwood).* 24, 9–17. doi: 10.1377/hlthaff.24.1.9

Fantino, E. (2004). Behavior-analytic approaches to decision making. *Behav. Process.* 66, 279–288. doi: 10.1016/j.beproc.2004.03.009

Fischhoff, B., and Bar-Hillel, M. (1984). Diagnosticity and the base-rate effect. *Mem. Cogn.* 12, 402–410. doi: 10.3758/BF03198301

Fox, J., and Fox, J. (2016). *Applied regression analysis and generalized linear models. Third Edition* Los Angeles: SAGE. *xxiv*, 791.

Fox, J., and Weisberg, S. *An R companion to applied regression. Third* Los Angeles: SAGE. (2019). *xxx*, 577p.

Friston, K. J. (2007). *Statistical parametric mapping: The analysis of funtional brain images. 1st.* Amsterdam; Boston: Elsevier/Academic Press. *vii*, 647.

Gigerenzer, G. (1996). The psychology of good judgment: frequency formats and simple algorithms. *Med. Decis. Mak.* 16, 273–280. doi: 10.1177/0272989X9601600312

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gravetter, F. J., and Wallnau, L. B. (2017). *Statistics for the behavioral sciences. 10th* Australia; United States: Cengage Learning. *xix*, 732p.

Grömping, U. (2006). Relative importance for linear regression in R: the package relaimpo. *J. Stat. Softw.* 17, 1–27. doi: 10.18637/jss.v017.i01

Hashimzade, N., and Thornton, M. A. (2021). *Handbook of research methods and applications in empirical microeconomics.* Cheltenham, UK; Northampton, MA: Edward Elgar publishing. *xii*, 650.

Hegdé, J. (2021). "Overfitting" in *Encyclopedia of research design.* ed. N. J. Salkind (Thousand Oaks, CA: SAGE Publications), 981–983.

Hoffrage, U., and Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Acad. Med.* 73, 538–540. doi: 10.1097/00001888-199805000-00024

Hoffrage, U., Krauss, S., Martignon, L., and Gigerenzer, G. (2015). Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Front. Psychol.* 6:1473. doi: 10.3389/fpsyg.2015.01473

Jaccard, J., and Turrisi, R. (2003). *Interaction effects in multiple regression. 2nd* Thousand Oaks, Calif.: Sage Publications. *vii*, 92.

Kahneman, D., Slovic, P., and Tversky, A. *Judgment under uncertainty: Heuristics and biases.* Cambridge; New York: Cambridge University Press. (1982). *xiii*, 555.

Kahneman, D., and Tversky, A. (1973). On the psychology of prediction. *Psychol. Rev.* 80, 237–251. doi: 10.1037/h0034747

Kahneman, D., and Tversky, A. (1982). "Evidential impact of base rates" in *Judgment under uncertainty: Heuristics and biases.* eds. D. Kahneman and A. PaulTversky (New York, NY: Cambridge University Press), 153–160.

Kalinowski, P., Fidler, F., and Cumming, G. (2008). Overcoming the inverse probability fallacy: a comparison of two teaching interventions. *Methodology* 4, 152–158. doi: 10.1027/1614-2241.4.4.152

Koehler, J. J. (1996). The base rate fallacy reconsidered: descriptive, normative, and methodological challenges. *Behav. Brain Sci.* 19, 1–17. doi: 10.1017/S0140525X00041157

Laha, A. K. (2019). *Advances in Analytics and Applications*. Singapore: Springer.

Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Glenview, Ill.: Scott, Foresman.

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Front Psychol.* 5:1144. doi: 10.3389/fpsyg.2014.01144

Manly, B. F. J. (2007). *Randomization, bootstrap, and Monte Carlo methods in biology*. *3rd* Boca Raton, FL: Chapman & Hall/ CRC. 455.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Elsevier.

R_Core_Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raab, M., and Gigerenzer, G. (2015). The power of simplicity: a fast-and-frugal heuristics approach to performance science. *Front. Psychol.* 6:1672. doi: 10.3389/fpsyg.2015.01672

Raacke, J. D. (2005). *Improving use of statistical information by jurors by reducing confusion of the inverse*. Doctoral Thesis. Manhattan, Kansas: Kansas State University.

Ranganathan, P., and Aggarwal, R. (2018). Common pitfalls in statistical analysis: understanding the properties of diagnostic tests - part 1. *Perspect. Clin. Res.* 9, 40–43. doi: 10.4103/picr.PICR_170_17

Sanborn, A. N., and Chater, N. (2016). Bayesian brains without probabilities. *Trends Cogn. Sci.* 20, 883–893. doi: 10.1016/j.tics.2016.10.003

Sevilla, J., and Hegdé, J. (2017). "Deep" visual patterns are informative to practicing radiologists in mammograms in diagnostic tasks. *J. Vis.* 17:90. doi: 10.1167/17.10.90

Spellman, B. A. (1996). The implicit use of base rates in experiential and ecologically valid tasks. *Behavioral and Brain Sciences* 19:38. doi: 10.1017/S0140525X00041406

Thompson, W. C., and Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials - the Prosecutor's fallacy and the defense Attorney's fallacy. *Law Hum. Behav.* 11, 167–187. doi: 10.1007/BF01044641

Uhlmann, E. L., Victoria, L., and Pizarro, D. (2007). The motivated use and neglect of base rates. *Behav. Brain Sci.* 30, 284–285. doi: 10.1017/S0140525X07001938

Venables, W. N. R., and Ripley, B. D. *Modern applied statistics with S*. New York, NY: Springer. (2003).

Villejoubert, G., and David, R. (2002). The inverse fallacy: an account of deviations from Bayes's theorem and the additivity principle. *Mem. Cogn.* 30, 171–178. doi: 10.3758/BF03195278

frontiers | Frontiers in Psychology

TYPE Systematic Review
PUBLISHED 03 April 2023
DOI 10.3389/fpsyg.2023.1145246

OPEN ACCESS

EDITED BY
Dario Monzani,
University of Palermo,
Italy

REVIEWED BY
Rick Thomas,
Georgia Institute of Technology,
United States
Richard S. John,
University of Southern California,
United States

*CORRESPONDENCE
Vincent Berthet
✉ vincent.berthet@univ-lorraine.fr

SPECIALTY SECTION
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

RECEIVED 15 January 2023
ACCEPTED 02 March 2023
PUBLISHED 03 April 2023

CITATION
Berthet V and de Gardelle V (2023) The heuristics-and-biases inventory: An open-source tool to explore individual differences in rationality.
Front. Psychol. 14:1145246.
doi: 10.3389/fpsyg.2023.1145246

The heuristics-and-biases inventory: An open-source tool to explore individual differences in rationality

Vincent Berthet[1,2]* and Vincent de Gardelle[2,3]

[1]Department of Psychology, Université de Lorraine, 2LPN, Nancy, France, [2]Centre d'Économie de la Sorbonne, CNRS UMR 8174, Paris, France, [3]CNRS and Paris School of Economics, Paris, France

Over the last two decades, there has been a growing interest in the study of individual differences in how people's judgments and decisions deviate from normative standards. We conducted a systematic review of heuristics-and-biases tasks for which individual differences and their reliability were measured, which resulted in 41 biases measured over 108 studies, and suggested that reliable measures are still needed for some biases described in the literature. To encourage and facilitate future studies on heuristics and biases, we centralized the task materials in an online resource: The Heuristics-and-Biases Inventory (HBI; https://sites.google.com/view/hbiproject). We discuss how this inventory might help research progress on major issues such as the structure of rationality (single vs. multiple factors) and how biases relate to cognitive ability, personality, and real-world outcomes. We also consider how future research should improve and expand the HBI.

KEYWORDS

individual differences, heuristics and biases, rationality, decision-making, measures

1. Introduction

The heuristics-and-biases (HB) research program, introduced by Tversky and Kahneman in the early 1970s (Kahneman and Tversky, 1972; Tversky and Kahneman, 1973, 1974), is a descriptive approach to decision-making that consists of invoking heuristics (mental shortcuts) to explain systematic deviations from rational choice behavior. For instance, people may misestimate a numerical value because of an overreliance on information that comes to mind and insufficient adjustment (anchoring-and-adjustment heuristic; Tversky and Kahneman, 1974). Another well-known example of a cognitive bias is the framing effect, by which individuals respond differently to a choice problem when the possible outcomes are framed as gains or as losses.

Since its inception, research on HB has produced a large literature on errors in judgment and decision-making (Gilovich et al., 2002) and triggered much discussion. Important questions include, among others, whether deviations from rationality can be reduced to randomness in choice (Stanovich and West, 2000), and whether HB effects are universal or instead vary across situations (e.g., due to the ecological or non-ecological nature of the task; Gigerenzer, 1996, 2008) or across individuals (Stanovich and West, 1998; Baron, 2008). Indeed, not all HB effects are present to the same extent in all individuals. Some biases are more prevalent than others: loss aversion might be found in a large majority of individuals (Gächter et al., 2022), whereas framing effects might not (Li and Liu, 2008). Some individuals might be more susceptible than others. In the case of attribution bias, for instance, that is the observation that individuals are more prone to credit themselves for positive than for negative events, a large meta-analysis conducted by Mezulis et al. (2004) demonstrated significant variations across countries, across genders, as well as associations with clinical symptoms.

In addition, to go beyond establishing a list of biases, efforts have been made to describe how different biases relate to each other. In this line of work, some studies have argued for a common decision-making competence underlying several HB tasks (Bruine de Bruin et al., 2007), akin to the *g*-factor (Carroll, 1993), whereas other studies covering a more heterogeneous set of tasks have provided support for a more complex, multidimensional structure (Klaczynski, 2001; Weaver and Stewart, 2012; Aczel et al., 2015; Teovanović et al., 2015; Ceschi et al., 2019; Berthet et al., 2022; Erceg et al., 2022; Rieger et al., 2022; Burgoyne et al., 2023). This research illustrates how the cognitive structure underlying heuristics and biases in decision-making can be investigated using individual differences.

Individual differences, however, have not been the main focus of earlier research on HB effects. The first reason is that the goal of this research was to demonstrate the existence of HB effects in the first place, on average, across participants. A second reason was the methodological choice to do so using between-subjects designs. This choice was notably motivated by the assumption that between-subjects designs favor spontaneous, intuitive answers in individuals, which are precisely the phenomenon of interest in HB research. As Kahneman (2000, p. 682) puts it: "much of life resembles a between-subjects experiment." By contrast, within-subject designs would be more transparent to participants, emphasizing the comparison between the conditions of interest, which might trigger the engagement of a slower, more deliberative system, and the override of intuitive answers, thereby reducing HB effects (Kahneman and Tversky, 2000; Kahneman and Frederick, 2005).

Critically though, the assumption that within-subject designs would produce smaller HB effects has not always been supported in empirical studies (Piñon and Gambara, 2005; Aczel et al., 2018; Gächter et al., 2022). In addition, regarding transparency, participants may remain unable to identify the research hypothesis in within-subject designs (Lambdin and Shaffer, 2009). In addition, they offer better statistical power than between-subject designs, and they eliminate confounds related to potential differences between participants in the different experimental conditions. Thus, within-subject designs seem appropriate tools to examine HB effects. As they allow for measuring biases at the individual level, these designs are particularly suited for individual differences. However, the measurement of individual differences in HB raises a practical and methodological issue: Finding such measures can be difficult and time-consuming while we still do not know much about their reliability.

The goal of the present study is to address these issues. First, we identify the currently available tasks that measure HB effects at the individual level. To do so, we conduct a systematic survey of empirical studies measuring one or more cognitive biases in a within-subject manner, focusing on studies in which the reliability of the measure used to quantify the bias is documented. Indeed, tasks optimized for large average effects might turn out to be less reliable at the individual level, producing a tradeoff between effect size and reliability (Hedge et al., 2018). We find that when reliability is documented, it is usually good. However, there are also a good number of HB effects for which no within-subject design has been tested or for which reliability is not known.

Second, we introduce an open online resource for the scientific community: the Heuristics-and-Biases Inventory (HBI[1]). This platform aims primarily at providing in a single location the experimental material for quantifying HB effects at the single subject level. The platform is meant to include new tasks as they are developed. Our hope is that this contribution will foster research on individual differences regarding cognitive heuristics and biases.

2. Methods

We conducted a systematic review in accordance with the PRISMA guidelines (Page et al., 2021).[2]

2.1. Search strategy

The following databases were searched for peer-reviewed empirical articles in June 2022: Web of Science, PsycINFO, and Pubmed. Our search strategy was based on the conjunction of two criteria: (1) the presence of "individual differences" in the title or in the abstract and (2) the presence of the terms "heuristics and biases" OR "cognitive bias" OR "cognitive biases" OR "behavioral biases" OR "rationality" OR "Decision-Making Competence" in the title or the abstract. All entries were imported in Zotero to remove duplicates, after which titles and abstracts were screened independently by two coders, according to predefined eligibility criteria.

Noteworthy, this search strategy had two implications. First, we likely missed relevant papers as we did not enter every single HB as a keyword, thereby limiting the comprehensiveness of our inventory. Second, our search strategy would not necessarily filter out studies that addressed psychological biases other than those pertaining to the heuristics-and-biases tradition (judgment and decision-making) such as health anxiety-related biases (e.g., interpretive bias and negativity bias) and the cognitive bias modification paradigm which aims at reducing them (e.g., Hallion and Ruscio, 2011).

2.2. Inclusion and exclusion criteria for studies

Included studies had to (1) be published in peer-reviewed scientific journals, (2) be written in English, and (3) be conducted on human participants. Reviews, conceptual or theoretical articles, book chapters, conference proceedings, dissertations, and editorial materials were excluded. We also excluded as follows: (1) Studies that addressed biases not pertaining to the HB tradition (e.g., health anxiety-related biases and implicit biases) for reasons

[1] https://sites.google.com/view/hbiproject/

[2] The original ADMC included a seventh task (path independence) which has been removed due to low reliability and validity.

previously mentioned, (2) studies in which self-report (questionnaires) rather than behavioral measures were used, (3) studies that merely applied the Adult Decision-Making Competence (ADMC), (4) studies in which a between-subject design was used. In addition, we chose not to include in the inventory two biases related to risk aversion (ambiguity aversion and zero-risk bias), which refers to a preference rather than a rationality failure (refer to the Discussion section).

2.3. Data collection and analysis

Relevant data was extracted by VB. The following information was collected for each study: author names, year of publication, title and journal where the study was published, study design, number of participants, inclusion/exclusion criteria, the HB task(s) used, whether the task(s) included single or multiple items, and the estimated reliability when reported. Discrepancies that emerged after full-text screening were resolved through a consensus meeting.

3. Results

Figure 1 displays the PRISMA flowchart with full detail of this process. The complete search resulted in 1429 articles, leaving 1,091 articles once duplicates were removed. After title and abstract screening, 109 articles met the inclusion criteria and were eligible for full-text assessment. One study was subsequently excluded because the author published the same data in another article. A total of 108 studies met eligibility criteria and were included in the review.

3.1. Study characteristics

Overall, the 108 studies included a total of 58,808 participants. Slightly more than half of the studies investigated a single HB ($n = 56$), while the rest addressed multiple HB ($n = 51$). Regarding the number of items, studies used one or several single-item tasks ($n = 29$), one or several multi-item tasks ($n = 64$), or both single and multi-item tasks ($n = 14$). Critically, out of the 78 studies that used multi-item tasks in the present survey, only 14 reported estimates of score reliability.

3.2. An inventory of tasks measuring heuristics and biases

Table 1 provides a reduced presentation of the outcome of this systematic survey. We identified 41 heuristics and biases for which there are tasks to measure individual differences. For each bias, we indicate the original paper introducing a typical task to measure the bias, the description of the task, the number of items, and estimated reliability when reported. A full version of the table is available in the supplementary material and on the HBI website, which also indicates, for each bias, the scoring rule and references of studies that merely used the task but did not report reliability. Note that there can be different measures available for some biases (e.g., anchoring) and that the same task can be associated with different

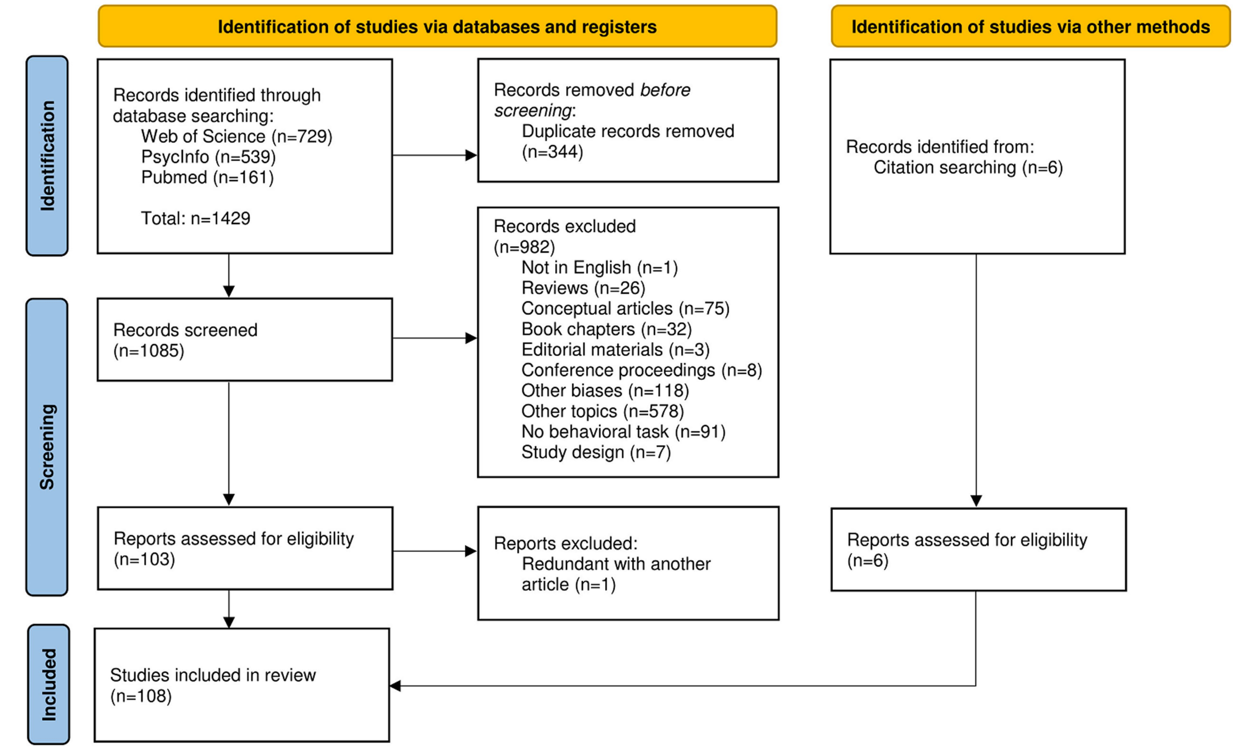

Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Included
Records identified through database searching:
Web of Science (n=729)
PsycInfo (n=539)
Pubmed (n=161)
Total: n=1429
Records removed *before screening*:
Duplicate records removed (n=344)
Records identified from:
Citation searching (n=6)
Records screened (n=1085)
Records excluded (n=982)
Not in English (n=1)
Reviews (n=26)
Conceptual articles (n=75)
Book chapters (n=32)
Editorial materials (n=3)
Conference proceedings (n=8)
Other biases (n=118)
Other topics (n=578)
No behavioral task (n=91)
Study design (n=7)
Reports assessed for eligibility (n=103)
Reports excluded:
Redundant with another article (n=1)
Reports assessed for eligibility (n=6)
Studies included in review (n=108)

FIGURE 1
PRISMA flowchart of study selection.

TABLE 1 Inventory and reliability of tasks measuring individual differences in heuristics and biases.

Task	Source	Items	Reliability
Anchoring heuristic: Tendency to adjust judgments toward the first piece of information.	Teovanović et al. (2015)	24	0.77
	Berthet (2021)	12	0.68
	Stanovich et al. (2016)	8	0.48
	Berthet et al. (2022)	8 pairs	0.67, 0.75, r_{tt} = 0.63
Attribution bias (including self-attribution bias): Tendency to refer to internal rather than external factors when explaining a person's behavior.	None		
Availability heuristic: Tendency to judge events' likelihood or frequency based on ease of recall.	Berthet et al. (2022)	4 pairs	0.67, 0.81, r_{tt} = 0.48
	Erceg et al. (2022)	4	0.77
Base-rate neglect (statistical): Tendency to ignore base rates in favor of individuating information.	Burič and Šrol (2020)	8	0.77, 0.78
	Šrol and De Neys (2021)	8	0.82
	Burgoyne et al. (2023)	11	0.46
	Šrol (2022)	4	0.70
	Erceg et al. (2022)	4	0.93, 0.95
	Berthet (2021)	4	0.70
Base-rate neglect (causal): Tendency to ignore causally relevant base rates in favor of individuating information.	Teovanović et al. (2015)	10	0.71
	Erceg et al. (2022)	3	0.55, 0.48
Belief bias in syllogistic reasoning: Tendency to evaluate deductive arguments based on the believability of the conclusion rather than its logical validity.	Teovanović et al. (2015)	8	0.76
	Berthet (2021)	4	−0.15
	Stanovich et al. (2016)	16	0.65
	Erceg et al. (2022)	8	0.79, 0.82
	Burič and Šrol (2020)	8	0.67, 0.78
	Šrol and De Neys (2021)	8	0.80
Better-than-average effect: Tendency to perceive one's abilities as superior to the average.	Rieger et al. (2022)	3	0.60
Bias blind spot: Tendency to see themselves as less biased than other people.	Scopelliti et al. (2015)	14	0.86
Confirmation bias (four-card selection task): Tendency to confirm rather than infirm the hypothesis (logical rule) at hand.	Burgoyne et al. (2023)	10	0.66
	Erceg et al. (2022)	4	0.86, 0.80
	Berthet et al. (2022)	4	0.84
Confirmation bias (2–4-6 task): Tendency to confirm rather than infirm the hypothesis (numerical rule) at hand.	Berthet et al. (2022)	3	0.75
Confirmation bias (interviewee's personality task): Tendency to confirm rather than infirm the hypothesis (personality trait) at hand.	Berthet (2021)	4	0.68
	Berthet et al. (2022)	4	0.83, 0.88, r_{tt} = 0.75
	Berthet et al. (2022)	4	0.64
Confirmation bias (financial decision-making) Tendency for people to disregard the counterevidence regarding their financial investments.	Rieger et al. (2022)	5	0.66
Conjunction fallacy: Tendency to judge that a conjunction of two possible events is more likely than one or both of the conjuncts.	Burgoyne et al. (2023)	7	0.69
	Šrol and De Neys (2021)	8	0.78
	Šrol (2022)	4	0.63
Conservatism: Tendency to overweight prior experience relative to new information.	None		
Covariation detection: Tendency for people to ignore essential comparative (control group) information.	None		
Debt account aversion: Tendency for consumers saddled with multiple debts to be motivated to reduce their total number of outstanding loans, rather than their total debt across loans.	None		
Denominator neglect/ratio bias: Tendency to pay too much attention to numerators and inadequate attention to denominators.	Stanovich et al. (2016)	12	0.88

(Continued)

TABLE 1 (Continued)

Task	Source	Items	Reliability
Framing (risk and attribute): Tendency to be affected by how information is structured.	Bruine de Bruin et al. (2007)	14 pairs	0.62, $r_{tt} = 0.58$
	Parker and Fischhoff (2005)	5 pairs	0.30
	Stanovich et al. (2016)	11 pairs	0.66
	Berthet (2021)	8 pairs	0.74
	Berthet et al. (2022)	8 pairs	0.76, 0.85, $r_{tt} = 0.45$
	Erceg et al. (2022)	8 pairs	0.35, 0.17, 0.24
Fungibility of money: Tendency for people to ignore the fact that all money is the same.	None		
Gambler's fallacy: Tendency to believe that the probability for an outcome after a series of outcomes is not the same as the probability for a single outcome.	Erceg et al. (2022)	4	0.76
	Šrol (2022)	4	0.52
Hindsight bias: Tendency to make different judgments (e.g., judging the probability of an outcome) between hindsight and foresight conditions.	Teovanović et al. (2015)	14	0.66
	Berthet (2021)	10	0.62
House money effect: Tendency for people to make decisions dependent on the prior gain or loss; includes greater tendency to gamble with recently won money.	None		
Illusion of Control: Tendency to overestimate their ability to control events.	None		
Insensitivity to sample size: Tendency to neglect sample size in inferential judgments.	None		
Irrational diversification: Tendency for people to favor a portfolio based on the perceived risk rather than the actual risk of the portfolio (based on real variance or probability).	None		
Loss Aversion: Tendency to prefer avoiding losses to acquiring equivalent gains.	None		
Mental accounting: Tendency to assign different mental values to the same sum of money.	None		
Money illusion: Tendency for people to think of money in nominal, rather than real, terms.	None		
Myside bias: Tendency to evaluate evidence, generate evidence, and test hypotheses in a manner biased toward their own prior opinions and attitudes.	None		
Omission bias: Tendency to avoid actions that carry some risk but prevent a larger risk.	None		
Outcome bias: Tendency to evaluate the quality of a decision based on its outcome.	Teovanović et al. (2015)	10 pairs	0.83
	Berthet (2021)	16	0.85[b]
	Berthet et al. (2022)	16	0.89[b], 0.91[b], $r_{tt} = 0.78$
	Erceg et al. (2022)	4 pairs	0.65, 0.68
Overconfidence: Tendency to overestimate their abilities.	Bruine de Bruin et al. (2007)	34	0.77, $r_{tt} = 0.47$
	Parker and Fischhoff (2005)	42	0.79
	Teovanović et al. (2015)	21	0.94
	Stanovich et al. (2016)	36	0.55
	Berthet (2021)	11	0.81[b]
	Berthet et al. (2022)	11	0.73[b], 0.59[b], $r_{tt} = 0.54$
	Hansson et al. (2008)	40, 40	0.84, 80
	Glaser et al. (2013)	15	0.83
Probability matching: Tendency to match choice proportions to outcome proportions in a binary prediction task.	Fletcher et al. (2011)	2	0.68
Probability neglect bias: Tendency for people to disregard the small probability of an outcome when facing a situation that arouses strong emotions.	None		
Proportion dominance: Preference for proportionally higher gains, such that the same absolute quantity is valued more as the reference group decreases (e.g., saving 10/10 lives is preferred to saving 10/100 lives).	None		

(Continued)

TABLE 1 (Continued)

Task	Source	Items	Reliability
Regression to the mean: Tendency to neglect that extremely high or extremely low observations tend to become more moderate (i.e., closer to the mean) over time.	None		
Regret aversion: Tendency to make decisions in order to avoid feeling regret in future.	Rieger et al. (2022)	3	0.60
Representativeness heuristic: Tendency to assess similarity of objects and organize them based around the category prototype.	Yoon et al. (2021)	26, 26, 26	0.90, 0.86, 0.88
	Morsanyi et al. (2009)	3	0.51
Status quo bias (or default bias): Tendency to choose the default option.	None		
Sunk cost fallacy: Tendency to continue an endeavor once an investment in money, effort, or time has been made.	Bruine de Bruin et al. (2007)	10	0.54, $r_{tt} = 0.61$
	Parker and Fischhoff (2005)	2	0.03
	Berthet (2021)	5	0.35
	Berthet et al. (2022)	10	0.38
	Teovanović et al. (2015)	8	0.76
	Erceg et al. (2022)	4	0.56, 0.39
Temporal discounting: Tendency to prefer smaller immediate over larger delayed reward.	Stanovich et al. (2016)	26	0.97

Reliability is measured by Cronbach's alpha, Spearman-Brown corrected split-half reliability (^{b}), or test–retest correlation (r_{tt}). Some studies reported several estimates of score reliability, which are all included in the table. "None" means that all studies that aimed to measure the bias used single-item tasks or that multi-item tasks were used, but the authors reported no estimates of reliability. Refer to the HBI website (https://sites.google.com/view/hbiproject/), for the full table.

scoring procedures (e.g., framing). The items for each task are available on the HBI website.

It turns out that the list includes the main biases studied in HB research. In fact, 18 out of the 41 HB are among the biases that violate normative models listed by Baron (2008). Our review points out, however, that there has been no attempt to measure individual differences for several significant biases, such as planning fallacy and prominence effect.

3.3. Reliability

Reliability (internal consistency) can be only estimated when multi-item tasks are used. Although this was the case for 23 of the HB tasks identified here, only 14 of the reviewed studies reported estimates of internal consistency, and two studies assessed test–retest reliability (Bruine de Bruin et al., 2007; Berthet et al., 2022). For instance, for the status quo bias, or for the insensitivity to sample size, the reliabilities of the measures are unknown. In addition, 11 HB have been measured only with single-item tasks so far (ambiguity aversion, attribution bias, conservatism, denominator neglect, illusion of control, loss aversion, mental accounting, myside bias, omission bias, proportion dominance, and regression to the mean).

Based on the available estimates of internal consistency (excluding estimates of test–retest reliability which are too infrequent), the reliability of HB scores is most often above the generally accepted standard of 0.70 (Nunnally and Bernstein, 1994). This finding is noteworthy and confirms that despite the "reliability paradox" described by Hedge et al. (2018), tasks that were primarily designed to produce robust between-subject experimental effects can be turned into reliable measures of individual differences (note, however, that our estimate might be inflated by publication bias). Some exceptions are the framing effects and sunk cost fallacy, for which low reliabilities have been repeatedly found.

4. Discussion

The aim of the present study was to provide a systematic review of individual difference measures used in heuristics-and-biases research. Based on 108 studies, we listed 41 biases for which at least one behavioral task allows one to calculate individual scores. While it is apparent that some of the tasks belong to a particular category (e.g., availability heuristic, conjunction fallacy, gambler's fallacy, probability matching, and base-rate neglect all assess biases in probability), we did not organize the tasks according to a particular theoretical taxonomy (e.g., Baron, 2008; Stanovich et al., 2008). Indeed, a key aim of the HBI is to help researchers build a robust empirical classification of HB by allowing them to include a large number of tasks in the study design and, therefore, to test the validity of the existing theoretical taxonomies (Refer to the following text).

Noteworthy, our review raised the issue of the reliability of such scores. Indeed, a significant number of HB have been measured only with single-item tasks, which does not allow checking reliability. When multi-item tasks are used, the reliability of scores is not systematically reported. In addition, low-reliability estimates have been repeatedly found for some biases (e.g., framing and sunk cost fallacy). However, based on the available estimates of internal consistency, the reliability of HB scores turns out to be most often above the generally accepted standard of 0.70. We encourage researchers to (1) use multi-item tasks and systematically report score reliability, (2) avoid calculating composite scores derived from single-item HB tasks as such scores are unreliable (West et al., 2008; Toplak et al., 2011; Aczel et al., 2015).

In the following subsections, we discuss the limits of our systematic review, how the HBI relates to existing taxonomies, and how it could be used to further address the impact of cognitive biases on real-life behavior.

4.1. Limitations of the systematic review

There are several limitations of this systematic review worth considering. First, the comprehensiveness of our inventory is limited by our search strategy. In order to cover all papers that addressed individual differences in HB, one should enter every single heuristic and bias as a keyword, which we did not do for practical reasons. Note, however, that our review was not meant to be exhaustive but rather to lay the foundation for listing HB tasks that are suited for the measurement of individual differences. As a collaborative and evolutive repository, the HBI may become more exhaustive over time.

The second limit relates to the selection of biases. As mentioned in the Methods section, we excluded psychological biases that do not fall within the category of heuristics and biases, defined as rationality failures. In particular, health anxiety-related biases such as interpretive bias (the tendency to inappropriately analyze ambiguous stimuli) and negativity bias (the tendency to pay more attention or give more weight to negative experiences over neutral or positive experiences) are typically not considered in the classification of biases in the heuristics-and-biases approach (Baron, 2008). Similarly, we did not include in our inventory two biases related to risk aversion (ambiguity aversion and zero-risk bias), which refers to a preference rather than a rationality failure (an individual is considered risk averse if she prefers a certain or risky option to a riskier option with equal or higher expected value while an individual who prefers a risky option to a certain or less risky option with higher expected value will be considered risk-seeking; Fox et al., 2015). However, one could argue that the exclusion of such biases is somewhat arbitrary as there is no objective criterion to qualify a bias under the heuristics-and-biases approach. Based on how the HBI is used by researchers, we will consider the possibility of expanding the scope of the inventory to include other types of biases.

4.2. HBI and existing inventories

We discuss here how the HBI compares with two related tools, the ADMC and the Comprehensive Assessment of Rational Thinking (CART; Stanovich et al., 2016). The ADMC is a set of six behavioral tasks measuring different aspects of decision-making (resistance to framing, recognizing social norms, overconfidence, applying decision rules, consistency in risk perception, and resistance to sunk costs) (Parker and Fischhoff, 2005; Bruine de Bruin et al., 2007)[3]. Three of the ADMC tasks can be identified as HB tasks (resistance to framing, overconfidence, and resistance to sunk costs). The full-form CART includes 20 subtests, some of them measuring HB (e.g., gambler's fallacy, four-card selection task, and anchoring). Noteworthy, the CART and the HBI have different aims. The CART is an instrument that aims to provide an overall measure of rational thinking (the same way IQ tests measure intelligence): A given number of points is attributed to each subtest, and an overall rational thinking score (Rationality Quotient) is calculated (the full-form CART takes about 3 h to complete). Indeed, each subtest is thought to reflect a single subconstruct within the concept of rationality. Accordingly, the CART subtests are not thought to be used separately. On the other hand, the HBI follows a more basic and practical aim: Providing researchers with an open, collaborative, and evolutive inventory of HB tasks, each of which can be used separately.

4.3. HBI and future research

We argue that the HBI has the potential to help researchers in their investigation of several issues. The first one is the structure of rationality. Similar to other topics in psychology (e.g., intelligence, personality, executive functions, and risk preference), early studies on HB that followed an individual differences approach aimed to explore whether single or multiple factors accounted for the correlations between performance on various tasks. While some studies have suggested the existence of a single rationality factor (Stanovich and West, 1998; Bruine de Bruin et al., 2007; Erceg et al., 2022), several factor analytic studies supported multiple-factor solutions, which relate more or less to existing taxonomies of HB (e.g., Klaczynski, 2001; Weaver and Stewart, 2012; Aczel et al., 2015; Teovanović et al., 2015; Ceschi et al., 2019; Berthet, 2021; Rieger et al., 2022).

Irrespective of their results, virtually all studies that explored the structure of rationality suffered from two limitations. First, scores for some HB tasks (even multi-item ones) failed to reach satisfactory levels of reliability (e.g., Ceschi et al., 2019; Erceg et al., 2022), thereby questioning the robustness of the factorial solution. Second, the sample of HB tasks submitted to factor analysis was limited (mainly because of practical limits such as total testing duration) and not being representative of all biases listed in the literature. That limitation is important as one could reasonably expect that a higher number of tasks would result in a higher number of factors extracted. Indeed, Berthet et al. (2022) showed that there was no longer evidence of a general decision-making competence when adding four HB tasks to the six ADMC tasks while ensuring satisfactory levels of score reliability. By providing researchers with more HB tasks producing reliable scores, the HBI will further shed light on the structure of rationality. Indeed, performing factor analysis on more exhaustive samples of tasks might eventually lead to more robust empirical taxonomies of biases (Ceschi et al., 2019).

Second, the HBI will allow researchers to further address how heuristics and biases correlate with cognitive ability (Stanovich and West, 2008; Oechssler et al., 2009; Stanovich, 2012; Teovanović et al., 2015; Erceg et al., 2022; Burgoyne et al., 2023), personality traits (Soane and Chmiel, 2005; McElroy and Dowd, 2007; Weller et al., 2018), and real-life behavior (Toplak et al., 2017). Regarding the latter, Bruine de Bruin et al. (2007) reported that the ADMC components predicted significant and unique (after controlling for cognitive ability) variance on the Decision Outcome Inventory (DOI), a self-report questionnaire measuring the tendency to avoid negative real-life decision outcomes (e.g., rented a movie and returned it without having watched it at all) (refer to also Parker et al., 2015). However, Erceg et al. (2022) found no evidence that performance on HB tasks predicts various self-reported real-life decision outcomes (DOI, job and career satisfaction, peer-rated decision-making quality). In particular, personality traits

[3] Studies included in the systematic review are referenced in the online material only: https://osf.io/5xg92/.

(conscientiousness and emotional stability) were the most predictive of DOI scores (Berthet et al., 2022, found similar—unpublished—results). It is worth noting, however, that these studies included relatively few HB so how they relate to real-life behavior remains an issue to be further addressed.

5. Conclusion

As highlighted by Gertner et al. (2016, p. 3), "the study of bias within an individual difference framework is still largely in its infancy." The present article aims to introduce the HBI, an exhaustive inventory of behavioral tasks that allow for a reliable measurement of individual differences in heuristics and biases. The aim of the HBI is to foster individual differences research in heuristics and biases by improving the visibility and accessibility of the relevant measures. As a collaborative and evolutive repository of all available measures, the success of the HBI project depends on the scientific community. Indeed, we invite researchers to support the HBI by reporting any use of the tasks (published or unpublished) and submit their own—new or alternative—measures of heuristics and biases. This open and collaborative approach will allow us to share results and continually expand the inventory.

Large-scale studies will allow to establish norm data from the general population and specific groups (e.g., documenting effects of gender and age) for each bias. Thus, our hope is that the HBI can help the research on individual differences in heuristics and biases to progress from infancy to adulthood.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

VB: conceptualization, methodology, and writing. VG: conceptualization and writing. All authors contributed to the article and approved the submitted version.

Acknowledgments

The author is grateful to Predrag Teovanović for feedback on an earlier draft of this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1145246/full#supplementary-material

References

Aczel, B., Bago, B., Szollosi, A., Foldes, A., and Lukacs, B. (2015). Measuring individual differences in decision biases: methodological considerations. *Front. Psychol.* 6:1770. doi: 10.3389/fpsyg.2015.01770

Aczel, B., Szollosi, A., and Bago, B. (2018). The effect of transparency on framing effects in within-subject designs. *J. Behav. Decis. Mak.* 31, 25–39. doi: 10.1002/bdm.2036

Baron, J. (2008). *Thinking and deciding (4th)*. New York: Cambridge University Press.

Berthet, V. (2021). The measurement of individual differences in cognitive biases: a review and improvement. *Front. Psychol.* 12:630177. doi: 10.3389/fpsyg.2021.630177

Berthet, V., Autissier, D., and de Gardelle, V. (2022). Individual differences in decision-making: a test of a one-factor model of rationality. *Personal. Individ. Differ.* 189:111485. doi: 10.1016/j.paid.2021.111485

Berthet, V., Teovanović, P., and de Gardelle, V. (2022). *Confirmation bias in hypothesis testing: A unitary phenomenon?*

Bruine de Bruin, W., Parker, A. M., and Fischhoff, B. (2007). Individual differences in adult decision-making competence. *J. Pers. Soc. Psychol.* 92, 938–956. doi: 10.1037/0022-3514.92.5.938

Burič, R., and Šrol, J. (2020). Individual differences in logical intuitions on reasoning problems presented under two-response paradigm. Journal of Cognitive Psychology 32, 460–477.

Burgoyne, A. P., Mashburn, C. A., Tsukahara, J. S., Hambrick, D. Z., and Engle, R. W. (2023). Understanding the relationship between rationality and intelligence: a latent-variable approach. *Think. Reason.* 23, 1–42. doi: 10.1080/13546783.2021.2008003

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press

Ceschi, A., Costantini, A., Sartori, R., Weller, J., and Di Fabio, A. (2019). Dimensions of decision-making: an evidence-based classification of heuristics and biases. *Personal. Individ. Differ.* 146, 188–200. doi: 10.1016/j.paid.2018.07.033

Erceg, N., Galić, Z., and Bubić, A. (2022). Normative responding on cognitive bias tasks: some evidence for a weak rationality factor that is mostly explained by numeracy and actively open-minded thinking. *Intelligence* 90:101619. doi: 10.1016/j.intell.2021.101619

Fletcher, J. M., Marks, A. D. G., and Hine, D. W. (2011). Working memory capacity and cognitive styles in decision-making *Personality and Individual Differences*. 50, 1136–1141.

Fox, C. R., Erner, C., and Walters, D. (2015). "Decision under risk: from the field to the lab and back" in *Handbook of judgment and decision making*. eds. G. Keren and G. Wu (New York: Wiley), 43–88.

Gächter, S., Johnson, E. J., and Herrmann, A. (2022). Individual-level loss aversion in riskless and risky choices. *Theor. Decis.* 92, 599–624. doi: 10.1007/s11238-021-09839-8

Gertner, A., Zaromb, F., Schneider, R., Roberts, R. D., and Matthews, G. (2016). *The assessment of biases in cognition: Development and evaluation of an assessment instrument for the measurement of cognitive bias (MITRE technical report MTR160163)*. McLean, VA: The MITRE Corporation.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: a reply to Kahneman and Tversky. *Psychol. Rev.* 103, 592–596. doi: 10.1037/0033-295X.103.3.592

Gigerenzer, G. (2008). Why heuristics work. *Perspect. Psychol. Sci.* 3, 20–29. doi: 10.1111/j.1745-6916.2008.00058.x

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

Glaser,, M., Langer, T., and Weber, M. (2013). True overconfidence in interval estimates: Evidence based on a new measure of miscalibration. *Journal of Behavioral Decision Making* 26, 405–407.

Hallion, L. S., and Ruscio, A. M. (2011). A meta-analysis of the effect of cognitive bias modification on anxiety and depression. *Psychol. Bull.* 137, 940–958. doi: 10.1037/a0024355

Hansson, P., Rönnlund, M., Juslin, P., and Nilsson, L. G. (2018). Adult age differences in the realism of confidence judgments: overconfidence, format dependence, and cognitive predictors. Psychology and aging 23, 531–544.

Hedge, C., Powell, G., and Sumner, P. (2018). The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* 50, 1166–1186. doi: 10.3758/s13428-017-0935-1

Kahneman, D. (2000). A psychological point of view: violations of rational rules as a diagnostic of mental processes. *Behav. Brain Sci.* 23, 681–683. doi: 10.1017/S0140525X00403432

Kahneman, D., and Frederick, S. (2005). "A model of heuristic judgment" in *The Cambridge handbook of thinking and reasoning*. eds. K. J. Holyoak and R. Morrison (Cambridge: Cambridge University Press), 267–293.

Kahneman, D., and Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cogn. Psychol.* 3, 430–454. doi: 10.1016/0010-0285(72)90016-3

Kahneman, D., and Tversky, A. (Eds.). (2000). *Choices, values and frames*. New York: Cambridge University Press

Klaczynski, P. A. (2001). Analytic and heuristic processing influences on adolescent reasoning and decision-making. *Child Dev.* 72, 844–861. doi: 10.1111/1467-8624.00319

Lambdin, C., and Shaffer, V. A. (2009). Are within-subjects designs transparent? *Judgm. Decis. Mak.* 4, 544–566. doi: 10.1017/S1930297500001133

Li, S., and Liu, C.-J. (2008). Individual differences in a switch from risk-averse preferences for gains to risk-seeking preferences for losses: can personality variables predict the risk preferences? *J. Risk Res.* 11, 673–686. doi: 10.1080/13669870802086497

McElroy, T., and Dowd, K. (2007). Susceptibility to anchoring effects: how openness-to-experience influences responses to anchoring cues. *Judgm. Decis. Mak.* 2, 48–53. doi: 10.1017/S1930297500000279

Mezulis, A. H., Abramson, L. Y., Hyde, J. S., and 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. *Psychol. Bull.* 130, 711–747. doi: 10.1037/0033-2909.130.5.711

Morsanyi, K., Primi, C., Chiesi, F., and Handley, S. (2009). The effects and side-effects of statistics education: Psychology students' (mis-)conceptions of probability. *Contemporary Educational Psychology,* 34, 210–220.

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric theory* (*3rd*). New York: McGraw-Hill.

Oechssler, J., Roider, A., and Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *J. Econ. Behav. Organ.* 72, 147–152. doi: 10.1016/j.jebo.2009.04.018

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372:n71 doi: 10.1136/bmj.n71

Parker, A. M., de Bruin, W. B., and Fischhoff, B. (2015). Negative decision outcomes are more common among people with lower decision-making competence: an item-level analysis of the decision outcome inventory (DOI). *Front. Psychol.* 6:363. doi: 10.3389/fpsyg.2015.00363

Parker, A. M., and Fischhoff, B. (2005). Decision-making competence: external validation through an individual-differences approach. *J. Behav. Decis. Mak.* 18, 1–27. doi: 10.1002/bdm.481

Piñon, A., and Gambara, H. (2005). A meta-analytic review of framing effect: risky, attribute and goal framing. *Psicothema* 17, 325–331. doi: 10.1136/bmj.n71

Rieger, M. O., Wang, M., Huang, P.-K., and Hsu, Y.-L. (2022). Survey evidence on core factors of behavioral biases. *J. Behav. Exp. Econ.* 100:101912. doi: 10.1016/j.socec.2022.101912

Scopelliti, I., Morewedge, C. K., McCormick, E., Min, H., Lebrecht, S., and Kassam, K. S. (2015). Bias blind spot: Structure, measurement, and consequences. *Management Science.* 61:2468–2486.

Soane, E., and Chmiel, N. (2005). Are risk preferences consistent? The influence of decision domain and personality. *Personal. Individ. Differ.* 38, 1781–1791. doi: 10.1016/j.paid.2004.10.005

Šrol, J. (2022). Individual differences in epistemically suspect beliefs: The role of analytic thinking and susceptibility to cognitive biases *Thinking and reasoning*. 28, 125–162.

Šrol, J., and De Neys, W. (2021). Predicting individual differences in conflict detection and bias susceptibility during reasoning. *Thinking and Reasoning.* 27, 38–68.

Stanovich, K. E. (2012). "On the distinction between rationality and intelligence: implications for understanding individual differences in reasoning" in *The Oxford handbook of thinking and reasoning*. eds. K. J. Holyoak and R. G. Morrison (New York: Oxford University Press), 433–455.

Stanovich, K. E., Toplak, M. E., and West, R. F. (2008). "The development of rational thought: a taxonomy of heuristics and biases" in *Advances in child development and behavior*. ed. R. V. Kail (San Diego, CA: Elsevier Academic Press), 251–285.

Stanovich, K. E., and West, R. F. (1998). Individual differences in rational thought. *J. Exp. Psychol. Gen.* 127, 161–188. doi: 10.1037/0096-3445.127.2.161

Stanovich, K. E., and West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *Behav. Brain Sci.* 23, 645–665. doi: 10.1017/S0140525X00003435

Stanovich, K. E., and West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *J. Pers. Soc. Psychol.* 94, 672–695. doi: 10.1037/0022-3514.94.4.672

Stanovich, K. E., West, R. F., and Toplak, M. E. (2016). *The rationality quotient: Toward a test of rational thinking*. Cambridge, MA: MIT Press

Teovanović, P., Knežević, G., and Stankov, L. (2015). Individual differences in cognitive biases: evidence against one-factor theory of rationality. *Intelligence* 50, 75–86. doi: 10.1016/j.intell.2015.02.008

Toplak, M. E., West, R. F., and Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Mem. Cogn.* 39, 1275–1289. doi: 10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., and Stanovich, K. E. (2017). Real-world correlates of performance on heuristics and biases tasks in a community sample. *J. Behav. Decis. Mak.* 30, 541–554. doi: 10.1002/bdm.1973

Tversky, A., and Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cogn. Psychol.* 5, 207–232. doi: 10.1016/0010-0285(73)90033-9

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131. doi: 10.1126/science.185.4157.1124

Weaver, E. A., and Stewart, T. R. (2012). Dimensions of judgment: factor analysis of individual differences. *J. Behav. Decis. Mak.* 25, 402–413. doi: 10.1002/bdm.748

Weller, J., Ceschi, A., Hirsch, L., Sartori, R., and Costantini, A. (2018). Accounting for individual differences in decision-making competence: personality and gender differences. *Front. Psychol.* 9:2258. doi: 10.3389/fpsyg.2018.02258

West, R. F., Toplak, M. E., and Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: associations with cognitive ability and thinking dispositions. *J. Educ. Psychol.* 100, 930–941. doi: 10.1037/a0012842

Yoon, H. F., Scopelliti, I., and Morewedge, C. K. (2021). Decision making can be improved through observational learning. *Behavior and Human Decision Processes.* 162, 155–188.

TYPE Hypothesis and Theory
PUBLISHED 02 May 2023
DOI 10.3389/fpsyg.2023.1189704

OPEN ACCESS

EDITED BY
Sergio Da Silva,
Federal University of Santa Catarina, Brazil

REVIEWED BY
Bin Liu,
Hainan University, China
Murilo Silva,
University of California, Riverside, United States

*CORRESPONDENCE
Jyrki Suomala
✉ jyrki.suomala@laurea.fi

RECEIVED 19 March 2023
ACCEPTED 05 April 2023
PUBLISHED 02 May 2023

CITATION
Suomala J and Kauttonen J (2023)
Computational meaningfulness as the source of beneficial cognitive biases.
Front. Psychol. 14:1189704.
doi: 10.3389/fpsyg.2023.1189704

Computational meaningfulness as the source of beneficial cognitive biases

Jyrki Suomala[1]* and Janne Kauttonen[2]

[1]Department of NeuroLab, Laurea University of Applied Sciences, Vantaa, Finland, [2]Competences, RDI and Digitalization, Haaga-Helia University of Applied Sciences, Helsinki, Finland

The human brain has evolved to solve the problems it encounters in multiple environments. In solving these challenges, it forms mental simulations about multidimensional information about the world. These processes produce context-dependent behaviors. The brain as overparameterized modeling organ is an evolutionary solution for producing behavior in a complex world. One of the most essential characteristics of living creatures is that they compute the values of information they receive from external and internal contexts. As a result of this computation, the creature can behave in optimal ways in each environment. Whereas most other living creatures compute almost exclusively biological values (e.g., how to get food), the human as a cultural creature computes meaningfulness from the perspective of one's activity. The computational meaningfulness means the process of the human brain, with the help of which an individual tries to make the respective situation comprehensible to herself to know how to behave optimally. This paper challenges the bias-centric approach of behavioral economics by exploring different possibilities opened up by computational meaningfulness with insight into wider perspectives. We concentrate on *confirmation bias* and *framing effect* as behavioral economics examples of cognitive biases. We conclude that from the computational meaningfulness perspective of the brain, the use of these biases are indispensable property of an optimally designed computational system of what the human brain is like. From this perspective, cognitive biases can be rational under some conditions. Whereas the bias-centric approach relies on small-scale interpretable models which include only a few explanatory variables, the computational meaningfulness perspective emphasizes the behavioral models, which allow multiple variables in these models. People are used to working in multidimensional and varying environments. The human brain is at its best in such an environment and scientific study should increasingly take place in such situations simulating the real environment. By using naturalistic stimuli (e.g., videos and VR) we can create more realistic, life-like contexts for research purposes and analyze resulting data using machine learning algorithms. In this manner, we can better explain, understand and predict human behavior and choice in different contexts.

KEYWORDS

confirmation bias, framing effect, computational meaningfulness, naturalistic stimuli, fMRI, machine learning

Introduction

When making judgments or decisions, it is said that people often rely on simplified information processing strategies called heuristics, which may lead to systematic errors called cognitive biases (Berthet, 2021). Cognitive biases are considered human behaviors that violate normative standards of rationality from perspectives of classic logic and mathematics, described for example by the Expected Utility Theory (EUT; Von Neumann and Morgenstern, 2007). According to Gigerenzer (2018), the irrationality argument has become the backbone of behavioral economics. In this paper, we challenge such bias-centric approach to behavioral economics by exploring different possibilities by opening a wider perspective through the analysis of the phenomenon of computational meaningfulness.

It is a generally accepted idea that rationality is reasoning according to certain rules. Aristotle developed the logical syllogism and enthymeme as norms of human rationality. Logical syllogism links together a set of known premises to reach deductive conclusions, whereas enthymeme is suitable when a human has only limited knowledge about premises (Clayton, 2021). Furthermore, Descartes regarded the ability to use language during reasoning process as the hallmarks of rationality (Oaksford and Chater, 1994). However, most contemporary researchers emphasize, that rational rules should be described by rules of logic and mathematics. This idea of a rational decision-maker applying classical logic and mathematics is perhaps best described by EUT (Von Neumann and Morgenstern, 2007).

According to EUT, a rational decision-maker is a utility maximizer and s/he chooses the best option from those available (Kőszegi, 2010). Furthermore, EUT makes strong assumptions about rational decision-makers. First, they have stable and accurate representations of preferences and people respond to the options available to them independent of context and unaffected by other alternatives or temporal order (Suomala, 2020). Finally, a rational decision-maker behaves consistently and has all the necessary information to make a rational decision (Von Neumann and Morgenstern, 2007).

However, EUT produces predictions that are quite different from human behavior. It came under attack from researchers Tversky and Kahneman (1974) and Kahneman and Tversky (1979), who showed that humans cannot make rational decisions in the way that EUT and other normative theories had shown (Mckenzie, 2005). This BIAS-centric approach to BEHavioral Economics (BIASBEHA) has found a large number of cognitive biases and fallacies related to human choice (Tversky and Kahneman, 1974; Shafir and LeBoeuf, 2002; Ariely, 2009; Thaler, 2016). What the BIASBEHA has clearly shown is that the assumptions of the rationality of human behavior according to EUT do not have the power to explain, describe and predict human behavior in natural contexts. BIASBEHA has shown that people's decision-making is predictably irrational because they use simple heuristics, which lead to systematic errors, or biases relative to EUT (Leonard, 2008; Ariely, 2009; Thaler, 2016).

When BIASBEHA has shown that a human's decision-making does not follow the traditional principles of rationality, it falls into two serious fallacies. First, it does not take into account the complexity and flexibility of the human brain and real-life behavior with uncertainty. Behavioral research has traditionally been based on simplified models in which a certain behavioral phenomenon is explained by two or a few parameters. For example, Plato divided the mind into reason and emotion, and Descartes into the soul and body. Similarly, Kahneman (2011) follows Stanovich and West (2000), dividing thinking into system 1 (fast belief system) and system 2 (slow conscious and critical system). Although such simple divisions are fruitful metaphors for thinking, they are not capable of grasping the multidimensionality and flexibility of human thought. Second, it has mostly stripped the decision-maker of essential information—like prior beliefs—from its experimental setups. To move forward in the behavioral sciences, we should study people in those environments where they can use different sources of information in their behavior. We do not argue that BIASBEHA-approach has not any value in behavioral science. Of course, this tradition has increased our understanding of human behavior in different contexts. However, traditional experimental setups in psychology and other behavioral science are often too simple to capture the multidimensional human behavior and decision-making that takes place in different real-life contexts. We suggest that new neuroscientific and machine learning methods give new opportunities to provide an opportunity to bridge the gap between experimental research and real-life behavior (Jolly and Chang, 2019).

In this case, what is essential in a person's behavior and decision-making is computational meaningfulness (Suomala, 2020; Suomala and Kauttonen, 2022), with which a person makes decisions in complex situations of everyday life. The computational meaningfulness approach assumes, that the brain/mind operates in different contexts by inquiring directly from the structure of the real world by optimizing multidimensional—with millions of parameters—information relating to the contexts. Previously, both the satisficing (Simon, 1955) and the bounded rational model (Gabaix et al., 2006) emphasize the study of human behavior in realistic and meaningful contexts. However, the model of computational meaningfulness takes into account the enormous parameter space of the brain, which is missing from the mentioned models.

According to the contextual approach to human behavior and decision-making, the task of the human brain/mind is to interpret the continuous complex information it encounters in a meaningful way in terms of one's subjective goals and activities. There are thousands of potentially informative demographics-, dispositional-, personal-, genetic-, and neurobiological variables that correlate and affect human behavior. This process is inevitably very multidimensional and complex. Therefore, behavioral science needs tools to describe, explain and predict human behavior through models, which include hundreds or maybe thousands of parameters (variables; Yarkoni and Westfall, 2017; Jolly and Chang, 2019; Hasson et al., 2020). In addition, we describe the functioning of the human brain as a typical example of a biological computer processing huge information flows. The human brain's basic processes are inductions and approximations and cognitive biases are a by-product of a process where the brain processes huge amounts of information utilizing induction and approximation. These are essential features of an optimally designed computing system, like the human brain.

With the recent development in machine learning and neuroscientific methodology as well as the increasing availability of large-scale datasets recording human behavior, we have good tools to understand better human behavior in real-life contexts (Yarkoni and Westfall, 2017). Therefore, from the computational meaningfulness perspective of the brain/mind, the use of cognitive biases may not be foolish at all and can be rational under some conditions (Gershman, 2021).

The article is organized as follows. We begin by describing typical assumptions of the BIASBEHA tradition. In addition, we describe more specifically cognitive heuristics relating to confirmation bias and framing effect. In conclusion of these, we highlight the problems relating to this tradition. Then, we describe a contextual approach with the recent development in machine learning and neuroscientific methodology. We end with our conclusions and suggestions on how to move forward BIASBEHA tradition.

The heuristics and biases approach

The main aim of BIASBEHA was to study people's beliefs about uncertainty and the extent to which they were compatible with the normative rules of EUT and other traditional logical calculus. This research program has been quite successful with thousands of scientific articles, Nobel laureates Daniel Kahneman and Richard Thaler in economics, and practical applications [e.g., Behavioral Insight Team in the United Kingdom government; popular non-fiction books: (Thaler and Sunstein, 2009; Kahneman, 2011)]. Moreover, new cognitive biases are constantly being discovered (Baron, 2008; Berthet, 2022), which give a rather pessimistic picture of human rationality. It is impossible to cover all these thinking biases in one article, so we will choose only two quite common and much-studied cognitive biases. These are the confirmation bias and framing effect. Below we describe typical example studies of both of them and the different interpretations made of them from the perspective of human rationality.

Confirmation bias as an example of irrational human reasoning

The behavioral literature on how people should form and test hypotheses has borrowed heavily from the logic of scientific discovery. People tend to seek and interpret evidence in a way that supports their beliefs and opinions and reject information that contradicts them. This tendency has been regarded as confirmation bias (Nickerson, 1998; Austerweil and Griffiths, 2008; Gershman, 2021). The proclivity toward confirmation bias is considered one manifestation of people's inability to think rationally (Wason, 1960, 1968; Popper, 2014). For example, Popper (2014) argued that science progresses through falsification, i.e., disconfirmation. A descriptive example of this is the discovery of helicobacter pylori.

In June 1979—on his 42nd birthday—Robin Warren saw something surprising with the new electron microscopy he had just adopted. A sample taken from the stomach of a patient with gastritis appeared to contain new types of curved bacteria. Although according to the bacteriology of that time, bacteria cannot live in the stomach because of its acidity, Robin Warren believed his eyes almost immediately (Warren, 2005). He was ready to disconfirm (i.e., falsify) the current theory of gastritis and started to find human and material resources, to make experiments to prove his observation correct (Thagard, 1998).

Despite strong opposition from his colleagues, he worked purposefully and decisively. Eventually, he was able to reform bacteriology with his colleague Barry Marshall related to the fight against diseases caused by helicobacter pylori in the stomach, and in 2005 they received the Nobel Prize in Medicine for their work (Warren, 2005).

Without a doubt, inventing something new is perhaps the highest degree of human mental ability and the clearest manifestation of human rationality. The cognitive-historical studies have shown that often scientific-, technological-and business breakthrough starts from unexpected perceptions (Suomala et al., 2006; Thagard, 2009). Warren's case is a good example of this. The discovery of helicobacter pylori and demonstration of its effect in the development of gastritis and gastric ulcer is also a textbook example of the power of falsification in scientific discovery. The theory of bacteriology at the time was contradicted by Warren's observation. Similarly, Galileo disconfirmed his time's common theory that Moon has not any mountains. He made observations of mountains on the Moon with his new telescope and disconfirmed previous wrong theories. As Popper argued, science advances by falsification of current theories and hypotheses rather than by continually supporting theories (Popper, 2014). Typical for Warren's and Marshall's as well as Galileo's case was that other scientists were against them and came up with several explanations with which they tried to save the old theories.

However, most ordinary people—like many scientists—do not apply disconfirmation as an inference strategy. Rather, they try to find support for their current knowledge and beliefs. The tendency to use confirmation means people's proclivity to embrace information that supports their current beliefs and rejects information that contradicts them (Austerweil and Griffiths, 2008).

Illustrative examples of confirmation bias are attitude experiments about the death penalty (Lord et al., 1979) and the right to bear arms (Kahan et al., 2017). In the death penalty study, its supporters and opponents were asked to familiarize themselves with two fictional empirical studies. Individuals who supported capital punishment subsequently strengthened their belief in the effectiveness of the death penalty after reading the two studies, whereas individuals who opposed capital punishment subsequently strengthened their beliefs in its ineffectiveness. The conclusion of the effect of the data evaluations is that opinion shifts of the participants increase attitude polarization (Lord et al., 1979; Gershman, 2021). The same body of evidence confirms people's individual beliefs in opposite directions indicating humans' tendency to confirmation bias.

While the content of the study of Lord et al. (1979) above was a complex and emotional social issue, does the effect of confirmation bias decrease, when the content is not so emotionally charged content? The attitude study about the right to bear arms (Kahan et al., 2017) tackled this question. In the study, the participants were presented with a difficult problem that required numeracy—a measure of the ability to make use of quantitative information. As expected, participants highest in numeracy did to a great extent better than less numerate ones when the data were presented as results from a study of a new skin-rash treatment. However, when the content of the inference changed from fact-based to emotionally charged content, the situation changed. Now, the participants evaluated the results from the fictional study of a gun-control ban. Now subjects' responses became less accurate and politically polarized. Such polarization did not abate among subjects highest in numeracy, rather, people who were good at numeracy used their talent to strengthen their own beliefs similarly to people with lower numeracy.

The rule learning task of Wason (1960) and selection task of Wason (1968) are the most cited examples relating to confirmation bias. Human reasoning in these tasks has been considered an apt exemplification of human irrationality. In the rule learning task, participants need to generate triples of numbers to figure out what the experimenter has in mind. This task is a more demanding version of the generally known object recognition task with 20 questions (Navarro and Perfors, 2011). The allowable queries in both queries are in the general form "Does x satisfy the rule?," where x is an object in 20 question game and a number in Wason's rule learning game (Navarro and Perfors, 2011). Wason gave the triple "2-4-6" as an example of the rule. Then the participants were asked to construct a rule that applies to a series of triples of numbers to test their assumptions about the rule the experimenter had in mind. For every three numbers the subjects will be coming up with, the experimenter will tell them whether it satisfies the rule or not, until the subject comes up with the right rule (Wason, 1960).

Most participants first formed a hypothesis about the rule: a sequence of even numbers. Then they tested this rule by proposing more sequences of numbers typically "4-8-10," "6-8-12," and "20-22-24." The feedbacks to all these sequences were positive. The participants produced a few more tries until they felt sure they have already discovered the rule. Most participants did not discover the rule, which was simply "increasing numbers." Wason (1960) showed that most of the participants avoided falsifying their hypotheses and instead sought to find confirmation for their hypotheses.

In the selection task (Wason, 1968), participants are presented with four cards (A, K, 2, and 7), each with a number on one side and a letter on the other, and a rule "If a card has a vowel on one side, then it has even number on the other side." Thus, the rule has a general form "if p, then q." Participants have to select those cards that they must turn over to infer whether the rule is true or false. Following the argument of Popper (2014) about falsification (disconfirmation), the correct choice is to turn over the vowel card (A) and the odd card (7) because finding an odd number behind the vowel or a vowel behind the odd number would reveal the hypothesis to be false. In other words, according to Popperian rationally, the correct answer follows a falsificationist (i.e., disconfirmation) strategy. It appeared that only 4% of subjects used the disconfirmation strategy. By contrast, the vast majority of participants used the confirmation strategy by either only turning over the vowel card (A; 33%) or turning over the vowel (A) and even cards (2; 46%). In other words, people seem to be following a confirmation test strategy, turning over cards that confirm the rule.

The studies described above regarding confirmation bias have been taken as strong evidence that humans are fundamentally irrational in their reasoning. This shows up as an irrational belief updating of individuals (Kunda, 1990; Gershman, 2021) and a strong tendency to strong logical errors in individuals reasoning (Wason, 1960, 1968; Johnson-Laird and Wason, 1970; Kahneman, 2011; Thaler, 2016). These experiments show that participants violated Popper's normative rule, according to which a rational actor pays attention to things that contradict the reasoner's presuppositions. Instead, participants tested their hypotheses in a way that would lead them to be confirmed. We as humans gather information in a manner that leads us to believe or to strengthen our subjective presuppositions regardless of their correctness.

Confirmation bias as an example of the adaptability of human reasoning

However, wider interpretations of the phenomena of confirmation bias have been presented (Oaksford and Chater, 1994; Mckenzie, 2005; Navarro and Perfors, 2011; Gershman, 2021). In addition, many philosophers of science have rejected falsificationism as unfaithful to the history of science and to be anyway unworkable (Lakatos, 1970; Kuhn, 1996; Churchland, 2002). These new interpretations emphasize that confirmation bias can be rational under some conditions (Gershman, 2021). We present some of them below.

According to this broader view, when a person acts in a certain situation, the person tries to grasp those environmental cues that increase his/her understanding of this situation. Especially, the interpretation of an event is an inferential process and during this process, an individual tries to increase knowledge and decrease uncertainty. In this case, the confirmation approach can be the most effective strategy.

Whereas Warren's and Marshall's discovery of helicobacter pylori is a good example of Popper's understanding of scientific discovery (Popper, 2014); science progresses by falsification. However, there are also contrasting examples in the history of science. When astronomers discovered Uranus in 1781 and noticed that it was deviating from its predicted orbit, they did not try to disconfirm the prevailing Newtonian theory of gravitation (Clayton, 2021; Gershman, 2021). Thus, they behaved in similar ways as participants in Wason's experiments. They persistently sought a Newtonian-compatible explanation for Uranus' unusual trajectory and Le Verrier and Adams in 1845 independently completed calculations showing that the unusual trajectory of Uranus could be entirely explained by the gravity of a previously unobserved planetary body (See Gershman, 2021). Eventually, a year later Johann Gottfried Galle found through telescopic observation Neptune in the night sky almost exactly where Le Verrier and Adams predicted it had to be. These astronomers succeeded in two ways: they discovered a new planet, and they rescued the Newtonian theory from disconfirmation (Gershman, 2019).

Moreover, contemporary research has argued that belief polarization might arise from different auxiliary hypotheses about the data-generating process (Jaynes, 2003; Jern et al., 2014; Cook and Lewandowsky, 2016; Gershman, 2019). The mental simulations of people's brains do not include perfect natural, mental, and cultural events. As Gershman (2021) argues, resistance to disconfirmation can arise from the rational belief updating process, provided that an individual's intuitive theories include a strong prior belief in the central hypothesis, coupled with an inductive bias (Suomala and Kauttonen, 2022) to posit auxiliary hypotheses that place a high probability on observed anomalies. Jern et al. (2014) explained the findings of Lord et al. (1979) by using a rational Bayesian framework. When subjects in the experiment do not trust the results of the research, then reading a report about the ineffectiveness of capital punishment may strengthen their belief. These beliefs in research bias could include doubt about the validity of the experimenter, data source of stimuli, and other auxiliary arguments against the evidence presented during experiments as a whole (Corner et al., 2010). Similarly, Cook and Lewandowsky (2016) demonstrated that belief polarization and contrary updating are consistent with a normative rational approach using the Bayesian framework. Thus, various auxiliary hypotheses are almost always in play when a human makes inferences. When one's beliefs about auxiliary hypotheses will change, then the interpretation of observations will also change (Gershman, 2021). Next, we will look at the new interpretations of the results of Wason's tasks.

Several researchers consider that the structure of Wason's tasks is such that it favors the confirmation strategy in reasoning. Klayman and Ha (1987) found that confirmation bias can be understood as resulting from a basic hypothesis-testing heuristic, which they call the Positive Test Strategy (PTS). According to PST, people tend to look at instances where the target property is assumed to be present. Klayman and Ha (1987) emphasized that most task environments are probabilistic and then it is not necessarily the case that falsification provides more information than verification. What is the best strategy depends on the characteristics of the specific problem at hand.

For example, the true rule in the rule learning task, which the experimented has in mind ("increasing numbers") is more general than the tentative plausible hypotheses in participants' minds ("increasing intervals of two"; typically "4-8-10″, "6-8-12″, "20-22-24″). In this case, people tend to test those cases that have the best chance of verifying current beliefs rather than those that have the best chance of falsifying them (Klayman and Ha, 1987). Furthermore, PTS is more likely when testing cases people expect will not work to lead to disconfirmation when people are trying to predict a minority phenomenon (Klayman and Ha, 1987; Mckenzie, 2005). These two conditions are commonly met in real-world reasoning situations and the confirmation strategy appears to be the rational strategy during reasoning.

Furthermore, Oaksford and Chater (1994) argue that turning the A and the 2 cards (confirmation) in Wason's card selection task is the most informative for determining if the rule is true or not. The confirmation strategy epitomizes general findings that rare events are more informative than common events (Klayman and Ha, 1987; Mckenzie, 2005). Thus people infer that the rule includes rare items—as vowels in English are—then the PTS shows the rational approach to the task contrary to Wason's interpretations and many other researchers' interpretations (Wason, 1960, 1968; Johnson-Laird and Wason, 1970; Kahneman, 2011; Thaler, 2016).

A descriptive example of the human ability for adaptable reasoning is manifested in a version of the game "Battleship" (Hendrickson et al., 2016). The game took place on a 20 by 20 grid partially covered by 5 ships (gray rectangles). The task of participants in this game is to discover the correct arrangement of the ships in the grid. They could ask where the ships were located (confirmation strategy) or where they were not located (disconfirmation strategy). Participants were told that their goal was to position the ships in their correct positions. The correct positions were randomly selected from a large set of possible configurations (Hendrickson et al., 2016). Participants were randomly assigned to one experimental condition in which the size of the ships was manipulated such that the portion of the grid covered by the ships ranged from 10% to 90%. In small ship conditions, there were many more legal candidate hypotheses than in large ship conditions since there were many more possibilities in which no ships overlapped in small ship conditions (Hendrickson et al., 2016).

The research demonstrated that there is a clear relationship between hypothesis size (i.e., legal potential position) and the degree to which people prefer confirmation strategy. In the 10% condition the average preference for confirmation strategy (i.e., questions, where the ships are located) was 86%, whereas, in the 90% condition, it was only 36%. Consistent with optimal information-acquisition strategy, when the size of ships increased (i.e., legal potential positions decreased), the confirmation request declined. The study showed that the request for positive evidence (confirmation) declined as the size of hypotheses (literally the size of ships) increased, consistent with the optimal information-acquisition strategy.

When the findings of confirmation biases have been regarded as a manifestation of irrational human behavior, contemporary research—as we described above—has shown that this traditional approach is too narrow. Preference for confirmation reflects the structure of how people represent the world (Gershman, 2021). The ability to adapt, to act actively and flexibly in different environments is an indication of human rationality, although can sometimes lead to preposterous beliefs. Now we concentrate on other cognitive biases presented in heuristics and bias tradition, namely the framing effect.

Framing effect as an example of irrational human reasoning

The framing effect occurs when people's choices systematically depend more on how the information of objects or outcomes is described than the substance of the pertinent information (Mckenzie, 2005; Leong et al., 2017). It is considered cognitive bias because an individual's choice from a set of options is influenced more by how the information is worded than by the information itself.

In attribute framing tasks one frame is usually positive and one negative (Levin et al., 1998). Ground beef is evaluated as better tasting and less greasy among participants when it is described in a positive frame (75% lean) rather than in a negative frame (25% fat; Levin and Gaeth, 1988). Similarly, when a basketball player's performance is described in terms of performance of shots "made" (positive frame) rather than "missed" (negative frame), participants rate the player as better in terms of abilities in positive than negative condition (Müller-Trede et al., 2015).

Furthermore, the attribute framing effect is found in contexts of plea bargaining (Bibas, 2004) and among economists (Gächter et al., 2009). The analysis of plea-bargaining literature has brought up the effect of framing on the criminal justice system (Bibas, 2004). The effect of framing appears to be a crucial component in the process, although skillful lawyering may ameliorate its effect. Similarly, the framing effect of conference payment for the participants of a scientific conference for behavioral economics has been studied (Gächter et al., 2009). The results showed that while the junior experimental economics was influenced by the framing effect, the more senior economists were not (Gächter et al., 2009). In a similar vein, people who are knowledgeable about an attribute's distribution (i.e., what is the typical number of free throws scored per season by an athlete playing basketball in the NBA) exhibited a reduced framing effect in the basketball framing scenario. However, the framing effect was unaltered among the same people in the medical framing scenario, of which they had no prior knowledge (Leong et al., 2017).

It is worth noticing that the information framed above examples is not the outcome of a risky choice but an attribute or characteristic of the goods. However, the best-known examples of framing effects involve choosing between a risky and a riskless option that is described in terms of either gains or losses (Kahneman and Tversky, 1979, 1984; Tversky and Kahneman, 1981). When the options are framed as risk-level, gains, and losses, the reference point has an important role. Moreover, people are more willing to take risks when the information is framed negatively but seek to avoid risks when the information is framed positively (Tversky and Kahneman, 1981).

According to Prospect Theory (Kahneman and Tversky, 1979), a decision maker transforms objective values of offers to subjective values at the present of the reference point according to the S-shaped value function. In this case, a human feels the loss relatively stronger than the gain about a reference point. At first, the Prospect Theory has described human choice in contexts, where a decision maker's status quo at the time of each choice dictates the subjective reference point (Kahneman, 2003). In these situations, a decision maker perceives any negative departure from her status quo as a loss, while perceiving any positive departure from the same status quo as a gain (Tversky and Kahneman, 1981; Louie and De Martino, 2014). Later, there is growing evidence that people evaluate the outcomes in light of the expectations or their subjective goals which act as a reference point, similar to the status quo as a reference point (Camerer et al., 1997; Heath et al., 1999; Koszegi and Rabin, 2006; Abeler et al., 2011; Suomala et al., 2017). Therefore, the prospect theory is crucial to understanding the framing effect. It describes how people evaluate their losses and acquire insight asymmetrically.

This phenomenon is aptly described in the famous Asian disease-study (Tversky and Kahneman, 1981). In the study, the participants were asked to choose between two options for treatment for 600 people, who suffer from a dangerous imagined Asian disease. The first treatment was likely to result in the deaths of 400 people, whereas the second treatment had a 66% possibility of everyone dying and a 33% possibility of no one dying. These two treatments were then described to the participants of the experiment with either a negative framing (describing how many would die) or a positive framing (relating how many would live). The result of the study (Tversky and Kahneman, 1981) showed that 72% of participants chose the first option for treatment when it was framed positively, i.e., as saving 200 lives. However, only 22% of participants chose the same option when it was framed negatively, i.e., resulting in the deaths of 400 people. Similarly, when survival rates of a surgery or other medical procedure are emphasized, people are more likely to approve of the procedure than when the mortality rates of the procedure are emphasized (Levin et al., 1998).

Despite there being some evidence that the framing effect was attenuated for those participants knowledgeable about the context (Gächter et al., 2009; Leong et al., 2017), it is widely considered to provide clear-cut evidence of irrationality and systematic violations of the axioms of rationality in decision-making in the same way as the confirmation bias (Kahneman and Tversky, 1979; Kahneman, 2011). Framing effect violates especially the description invariance-principle (Von Neumann and Morgenstern, 2007) essential normative principle in EUT (Mckenzie, 2005). However, recent studies—as we described below—have shown that this is not necessarily the case.

Framing effects as an example of the adaptability of human reasoning

Recent studies related to human behavior have shown, that humans and other mammals are sensitive to the context as a whole (Gallistel and Matzel, 2013; Müller-Trede et al., 2015). The context as a whole has often a stronger effect on behavior than single objects or objects' attributes. Even when participants process information about artificial objects (i.e., stimuli) in decontextualized experiments, participants have a proclivity to form rich and versatile mental simulations, which include not only the stimuli but also the likely context and its latent causes in which these stimuli typically occur (Gershman et al., 2015; McKenzie et al., 2018; Cushman and Gershman, 2019). In these experimental as well as in real-life contexts, an individual infers based on her/his prior experience and expectation relating to a context as a whole (Baum, 2004; Gershman and Niv, 2013; Suomala, 2020; Suomala and Kauttonen, 2022). For example, when the above-described task includes the wording "the ground beef is 75% lean," a participant likely tries to understand this wording from the point of view of either the experimenter or the butcher (Leong et al., 2017). Then this context leaks information about the experimenter's and the butcher's intentions, and these informative signals are different in different options, despite options being logically equivalent (McKenzie and Nelson, 2003; Suomala, 2020).

Each real-life context contains an almost infinite number of configurations in terms of human interpretation ability. The human resolves this problem of abundant information flows by utilizing prior experiences (i.e., memories) and contextual information. When a researcher constructs the experiment, the narrative, and single words form the information context for participants. Mckenzie (2005) and Sher and McKenzie (2006) argues that the frame chosen by the researcher and its linguistic expression constitute the information content for the test subjects with reference points chosen by the researcher. In these cases, logically equivalent frames can signal relevant information beyond the chosen frame's literal content. For example, McKenzie and Nelson (2003) found that the "speaker" participants were more likely to express a cup with liquid at the halfway mark as "half empty" rather than "half full" when the cup had initially been full and was therefore empty. Then "Listener" participants, in turn, "absorbed" the information signaled by the speaker's choice of frame and were more likely to infer that a cup was originally full when it was described as "half empty." In other words, listeners' inferred reference points matched the actual reference points that guide speakers' frame selection. McKenzie and Nelson (2003) conclude that logically equivalent frames can often implicitly convey different information and participants are sensitive to this different information. Then logically equivalent frames can convey choice-relevant information and participants in the experiments exploit this information effectively (McKenzie and Nelson, 2003; Sher and McKenzie, 2006).

Human behavior from sensory observation to mental simulation constructions is guided by the principle of meaningfulness (Suomala, 2020; Suomala and Kauttonen, 2022; Gershman, 2023). This sense-making process emphasizes certain features of the context at the expense of other features. The human brain integrates incoming extrinsic information with prior intrinsic information to form rich, context-dependent models of situations as they unfold over time (Yeshurun et al., 2021). How individuals can weigh different elements when constructing the important elements of the context? An illuminating example is the study (Sher and McKenzie, 2014), which provided experiments, where the participants were asked to evaluate a suitable salary for coders and buy CDs.

In the salary experiment, participants saw three things about two applicants. Both had graduated from the University of San Diego with majors in programming. The average grade of Applicant A was 3.8 (max 4.0) and Applicant B was 3.1. In addition, A had programmed 10 programs in the YT programming language, while B had programmed 70 programs in the same language. The essential point here is that knowledge relating to the University of San Diego and grade were familiar to the participants, whereas the YT programming language was unknown to them. The participant groups, which evaluate individual applicants, based their evaluation on the known attributes. In this case, A applicant got a better salary suggestion than B applicant. This is understandable because the A applicant was better in grade than the B applicant. These individual evaluation groups ignored the effect of programming experience because they likely did not understand its meaning. However, the third group of participants evaluated both A and B applicants' salaries at the same time. In this case, participants suggested better salaries for B applicants. Despite the YT programming language being unknown among participants in this group, they were likely sensitive the relatively large difference (10 programs vs. 70 programs) between applicants.

Similarly, in CD study, participants showed their willingness to pay for different CD boxes. When individual CD-box was presented, unknown attributes were ignored by participants. However, when different versions of CD-boxes were presented at the same time, participants were capable to evaluate different versions and they also interpret unknown attributes of each other to make suitable price estimates (Sher and McKenzie, 2014). Thus, people are very sensitive to both implicit and explicit contextual clues, when trying to make sense of the context.

It is possible to assume, that the researchers planning an experiment form specific frames and reference points, and these original choices affect test subjects' inference processes about these frames. For example, the medical tasks described above illustrate that describing the treatment in terms of percent survival signals that the treatment is relatively successful, whereas describing it in terms of percent mortality signals that the treatment is relatively unsuccessful. This speaker–listener interpretation help explain also people's behavior in other framing contexts, which we described above.

The speaker-listener framework is reminiscent of Gricean notion of conversational implicature (Grice, 1975; Corner et al., 2010). According to conversational implicature, information is not contained in the literal content of an utterance but can be implied from the context in which it is given (Grice, 1975). Corner et al. (2010) emphasized that participants may infer more about the experiment than is contained in the literal content of the instructions and participants might have different ideas about what key task parameters are—such as the diagnosticity of evidence in belief revision experiments.

Similarly, people try to interpret the content of information based on plausibility (Jaynes, 2003). For example, in the case of Asian disease (Tversky and Kahneman, 1981) described above, it is very difficult to imagine that such a treatment would exist in real life. Recent research (Cohen et al., 2017) on the ability to reason in medical cases showed, that people's inference is rational in the traditional sense when the probabilities were believable. Similar logically consistent reasoning has been observed in syllogistic reasoning, where beliefs about the plausibility of statements based on everyday experience influence truth judgments (Revlin et al., 1980). Jaynes (2003) emphasizes that people's inference is neither deductive nor inductive, but it is plausible reasoning. It has strong convincing power, and a human decides this way all the time (Suomala and Kauttonen, 2022). Thus, people's reasoning process is not necessarily purely syntactic or computational. Rather, it is sensitive to meaningful properties of the combination formed by observation and prior experience. When the occurrence of objects and their frames and their relationships are meaningful from an individual perspective, her/his reasoning process appears to be rational (Gershman, 2021).

Above we have described examples of heuristic and biased approaches to the confirmation bias and the framing effect. Results in these studies appear to show that people do not reason according to the principles of classical rationality. In both confirmation effect—and framing effect experiments people's performance appears biased when compared with the standards of logic, probability theory, and EUT. However, contemporary critical studies showed that the human mind is more flexible, context-sensitive, and capable to interpret environmental features based on an individual's prior experiences. These studies considered misleading the purely negative view of human performance implied by the BIASBEHA approach.

Despite the current new critical approach to heuristics and biases, tradition has taken important steps in contextualizing human behavior, we must go further. As most of the empirical studies of human behavior—also these critical studies—suffer from the flatland fallacy (Jolly and Chang, 2019).

Term Flatland fallacy refers to Edwin Abbott's famous Novella Flatland: a Romance of Many Dimensions (Abbott, 2019), in which the creatures (Flatlanders) with limited perceptual capacities (i.e., seeing in only two dimensions) come to reason in a limited way. They ignored the complexity of the world and believed that their perceptions are veridical. Jolly and Chang (2019) argued that much like Flatlanders, humans exhibit strong biases in their reasoning about a complex and high-dimensional world due to finite limitations on their cognitive capacities. They claim that most psychological researchers are like Flatlanders and try to understand human behavior with impoverished models of human behavior. We agree and suggest that most of the results of BIASBEHA-tradition are a result of not taking the multidimensionality of human behavior into account. To overcome this fallacy, we should study human behavior under as natural conditions as possible. In the following chapters, we describe this approach more specifically.

Computational meaningfulness as the core of the human rationality

To move forward in the behavioral sciences, it is central to understand the behavior of people in real-life contexts. Our mind is not a photocopier. Rather it is a biological computer that extracts meaningful patterns from contexts to know how to behave adaptively in each context (Suomala, 2020). In this chapter, we describe factors that, according to our understanding, help behavioral scientists to conduct better research that takes into account human operating naturalistic environments. At first, we need a theoretical model of human behavior. Such a model should include the following factors (Hofstadter, 1979; Gallistel, 2009):

1. It realistically describes the signals that humans process, and how those signals are processed to yield action.
2. It realistically identifies meaningful actions.
3. Research results increase our understanding of human behavior in natural environments.

We claim that BIASBEHA approach does not include the three factors listed above. Next, we describe the foundation for a new behavior model based on the criteria described above.

The signals that humans process

Living creatures, from single-celled organisms to humans, always function in a certain context (Suomala, 2020). For a human, these contexts are usually cultural environments, the meanings of which a growing child learns to understand. When behaving in a certain context, a person computes information from the context to serve her activities. We call this process of transformation and utilization computation. Computation means the process by which the human brain transforms the contextual information and combines these with mental simulations previously adopted by the individual in order to behave in optimal ways (Tegmark, 2017; Suomala and Kauttonen, 2022).

This means that a person always develops, learns, and acts in a certain cultural context. This is aptly illustrated by the study (DeCasper and Spence, 1986) that showed that a child learned to prefer the fairy tale "The Cat in the Hat" during the fetal period, which one's mother read regularly at the end of the waiting period. Thus, children's preferences begin to be biased toward certain cultural things—in this case specific fairy tales—that are present in their environments. In other words, a child begins to embrace important cultural entities and to behave in this specific cultural context adaptable. Whereas the early learning of a child is likely limited to reasoning about objects and agents in their immediate vicinity, the wider cultural artifacts, values, and habits develop later with interactions of the child and other people and official institutions. During this process, the most crucial aspect of the human mind is the motivation to share culturally meaningful aspects with others (Tomasello et al., 2005; Tomasello, 2014; Suomala and Kauttonen, 2022). So, the contexts include not only the physical objects but above all the cultural entities. These contexts offer a person potential behavioral opportunities, which we call cultural affordances. A person learns and acquires knowledge and skills and may develop into an expert in some field. Growing into an expert is situational in nature.

Humans process signals from their contexts, which include constellations of cultural affordances. Described in this way, the concept of cultural affordances is related to Gibson's concept of affordance (Gibson, 1979) and Hasson's direct-fit approach (Hasson et al., 2020). The human brain constructs continuous experiences about the world to behave in optimal ways in a specific context. The real-life contexts in our society are complex, dynamic and uncertain, containing typically "countless" numbers of objects, the path of objects, people, and their interactions.

Thus, the world—physical and cultural—around us includes an almost infinite amount of information from a human point of view. The human resolves this problem of abundant information flows by using prior experiences (i.e., memories) and contextual information. In other words, from the point of view of humans, the world contains much more potential information than one can convert into knowledge according to her/his purposes.

The human brain computes the meaningful constellations about the contexts. It can extract meaningful patterns from complex and information-rich environments because the human brain has evolved specifically to function in complex and uncertain contexts. Despite the absolute number of neurons in the human brain remaining unknown, the approximation is that it has about 85 billion neurons (Azevedo et al., 2009) and it is each cubic millimeter contains roughly 50,000 neurons. Because these neurons may support approximately 6,000 adjustable synapses with other cells, this structure yields about 300 million parameters in each cubic millimeter of the cortex and over 100 trillion adjustable synapses across the entire brain (Azevedo et al., 2009; Hasson et al., 2020). Thus the human brain is overparameterized organ and it can produce flexible, adaptive behavior in a complex world (Hasson et al., 2020).

Even though the brain is efficient, an individual is only able to compute a small part of the information in the context with it. Let us imagine a six-year-old child buying penny candies with 10 different candies. The child is allowed to choose 10 candies. Mathematically, and following the rules of EUT, 10 different candy combinations in this context can form 92,378 different options. If it took 15 s to collect one bag, it would take a child a good 384 h, or a good 16 days, to try all these candy combinations if she did nothing else during that time. However, in real life, she can choose candies in a few minutes. We all make this kind of decision daily and despite the department store including over 100,000 items, we rarely spend more than an hour there. We do not behave according to EUT (Bossaerts and Murawski, 2017).

In a conclusion, people process only part of potential signals in a context. People are developed and learned to see easily things that our culture hands us ready-made as cultural affordances in different contexts (Hofstadter, 2001; Zadbood et al., 2021). If these meaningful constellations are lacking—like in typical BIASBEHA experiments—people still try to interpret minor context clues to make them understandable to themselves. This leads to false conclusions about behaviors that do not align with those made in real life.

People learn most frequently encountered cultural constellations over a lifetime. The learned constellations are stored in long-term memory as multidimensional and dynamic experiences. We call these stored memories as mental simulations because these memories are more vivid and dynamic movies than static object-like properties (Barsalou, 2009). Through these learned constellations, the past is intertwined with a person's present and future (Gallistel, 2017). Mental simulations of the contexts in the brain are dynamics networks where context-related information is stored in nodes. The links are synapses that carry messages from nodes to other nodes.

The meaningful actions as the human represent it

Like the contexts surrounding the individual, the mental simulations relating to the contexts stored in the individual's brain are also "countless." An individual has constructed them of experienced contexts during her/his lifetime. These context-based simulations are strongly domain-specific and intuitive. These mental simulations support an individual to produce flexible and meaningful behavior in a complex world.

The meaning of a context and meaningful actions are formed by the weights of individual nodes and their links to other elements of the context (and between contexts) in the brain (Hofstadter, 2001; Yeshurun et al., 2021). This forms a graph where context and actions are interconnected, not independent from each other. In other words, the elements of mental simulations, which need more memory resources, are more meaningful for a subject than elements that need just a few resources.

The objects, other people, cultural artifacts, and conventions and their interactions happen in specific contexts, and humans learn to behave in these contexts gradually. We are not born with an understanding of entities and their roles in specific contexts. This understanding must be learned from experience. As a child grows up, one's starts to perceive constellations of events. Then a growing child begins to construct fragments from life's streams as constellations as high-level wholes (Hofstadter, 2001). These learned complex constellations are constructed based on the principle of computational meaningfulness. This principle means, that the human brain can produce a set of constraints concerning the distinction between different constellations (a bunch of stimuli) of cultural affordances. Thus, computational meaningfulness is the result of human's ability to differentiate constellations from one another on a given set of observations. To do that, humans need the mental resources to choose the most meaningful features of the environment to behave in optimal ways in this environment (Ratneshwar et al., 1987; Suomala, 2020; Suomala and Kauttonen, 2022). In this way, a person learns to extract important aspects of the experienced context (Gallistel and Matzel, 2013).

Since each real-life situation contains an almost infinite number of possible configurations in terms of human interpretation ability, the human ability to assign meanings to certain constellations at the expense of others can be considered rational behavior (Hofstadter, 1979, 2001; Ratneshwar et al., 1987; Suomala, 2020).

Above we described the properties of contexts, the human brain, and mental simulations. When an individual acts in the context, s/he tries to find meaningful constellations about the current context and tries to figure out, how these constellations support her/his personal goals. How do these comprehensive processes and the human ability to find meaningful constellations in different contexts manifest human rationality?

Computational meaningfulness means the process of the human brain, with the help of which an individual tries to make the respective situation comprehensible to herself to know how to behave optimally in a specific context. Then rationality means four things. First, it means that the brain makes different contexts understandable by inquiring directly from the structure of the real world by recognizing the relative importance of different elements in these contexts by optimizing multidimensional—with millions of parameters—information relating to these contexts (Hofstadter, 1979; Hasson et al., 2020). Second, it means that a human can respond to contexts very flexibly and can make sense of ambiguous or contradictory messages (Hofstadter, 1979; Geary, 2005; Gershman, 2021). Third, it means that an individual can set complex goals and finally, it means that an individual can achieve these goals (Geary, 2005; Tegmark, 2017). In summary, computational meaningfulness embodies the human capacity for rationality.

Research results of behavioral studies should increase our understanding of human behavior in natural environments

When we take understanding human behavior in natural environments as a criterion to build a theory of behavior, it means that we are better able to describe, explain and predict human behavior (Gallistel, 2009, 2020; Yarkoni and Westfall, 2017; Jolly and Chang, 2019).

To better understand human behavior, as researchers we should leverage as natural stimuli and problems as possible in our experiments to capture realistic behavior. Despite the naturalness of stimuli in experiments lying along a spectrum, there can be described by three factors (Hamilton and Huth, 2020). First, a stimulus should represent a situation that a participant might reasonably be exposed to outside of an experimental setting. Second, the stimulus should appear in the same context as it would in real life. Third, the participants' motivation and feeling to solve problems or make decisions should be as similar as possible in the experiments as in real life. These properties are reminiscent of previous requirements that psychologists should focus on the structure of natural environments that the mind relies on to perform inferences and to guide behavior (Brunswik, 1955; Simon, 1955; Todd and Gigerenzer, 2007; Holleman et al., 2020). We argue that these three factors are absent from typical BIASBEHA studies.

However, most current ecological studies have shown that we can bridge the gap between theoretically simple traditional psychological experimental setups and real-life human behavior. We describe these studies as follows. Generally, the effect of a stimulus or other message on people has been studied from the point of view of the recipient of the message. However, the expression of the original context by the person who conveys the message is also important for how the recipient understands the message. Whether it is a single message or an entire experiment setup, it oozes latent meaning that the receiver instinctively interprets (McKenzie and Nelson, 2003).

Examples of studies that use natural stimuli in their experiments

The need for ecologically valid models has been also realized in the field of neuroscience (Nastase et al., 2020). As stated by Nastase (2021, 46): "We're left with a veritable zoo of piecemeal models that are difficult to synthesize and, considered individually, account for a disappointing amount of variance under natural conditions." Below we describe studies, which have used naturalistic and multidimensional stimuli in their experiments. Natural stimuli are videos, real advertisements, real health messages, stories, and immersive VR and AR technologies (Mobbs et al., 2021). Two groups of students participated in the Buzz study (Falk et al., 2013). A group of message communicators watched and evaluated new entertainment program concepts in the fMRI scanner intended for television. Immediately after the fMRI scan each message communicator presented the concepts outside of the scanner during video-interview. Then another group of students, who were message recipients, watched these videos.

Finally, message recipients were asked how willing they were to recommend the concept proposals they saw to their friends. The study showed that successful ideas were associated with neural responses initially measured by fMRI in the mentalizing system and the reward system of message communicators when they first heard, before spreading them during video-interview. Similarly, message communicators more able to spread their preferences to others produced greater mentalizing system activity during initial encoding. Thus, people are very sensitive to the semantics of the messages and can interpret the intention of the sender (in this case message communicators), not only the literal meanings of these messages. It is also valuable that the results of the fMRI-experiment generalize beyond the experimental situation to the natural video interview and its viewing, as well as the personal preference caused by viewing.

Similarly, Falk et al. (2011, 2012) examined how smokers' neurophysiological responses to antismoking advertisements predict subsequent smoking behavior. They found that the brain activation patterns in the valuation network of participants, when they were exposed to an anti-smoke message in the fMRI-scanner, more accurately predicted participants' proclivity to quit smoking 1 month after the initial fMRI than traditional behavioral measurements. Even more noteworthy is that the activity in the same region of the mean brain activation patterns in the valuation network of participants predicted population-level behavior in response to health messages and provided information that was not conveyed by participants' self-reports (Falk et al., 2012). Therefore, neural activity in the brain's valuation network predicted the population response, whereas the self-report judgments did not. Thus, the participants' neural patterns activation during fMRI-experiments "leaks" information about their valuation and desires, which have predictive power to real-like contexts.

In the same way, the research group of Genevsky and Knutson (2015); Genevsky et al. (2017) sought to find brain networks in laboratory samples to forecasted real microloans (Genevsky and Knutson, 2015) and crowdfund success (Genevsky et al., 2017) on the Internet. They found that the sample's average activity in the part of the brain's valuation network forecasted loan appeal and crowdfund success on the Internet. Findings demonstrate that a subset of the neural predictors in the valuation network of individual choice can generalize to forecast the market-level behavior of consumers.

Naturalistic stimuli as the path toward novel findings in neurosciences

Heretofore we have argued that we humans are sensitive to meanings and semantics of the messages in contexts (Grice, 1975; Corner et al., 2010), not so much their literal content from a purely logical perspective, as the BIASBEHA-approach assumes. One of the pioneer researchers who used naturalistic context as stimuli is Uri Hasson. He has not so much looked for ways to predict people's behavior outside of experimental situations, but rather he has tried to find a general common ground, especially for human communication and generally for human experiences. For example, in his seminal brain study (Hasson et al., 2004), the participants lay in a brain scanner and watched the Western film *The Good, the Bad, and the Ugly*. When the brain activations of all the participants measured by fMRI were looked at as a whole, the researchers found that the brains of the individuals activated in a very similar way to the important points of that classic Western movie. It was about the similar activation profile of individuals' brains, i.e., synchronization in certain movie scenes. Especially emotionally powerful moments in the film synchronize the brains of the participants. Such emotional moments were stages that contained excitement, surprise, and joy. In addition, emotional activation also increased at points where the theme changed to another. Other researchers have found that scenes featuring people or animals generally and the other person's eyes and face especially are especially powerful emotion stimulants and synchronize people's brains in similar ways (Sharot and Garrett, 2016).

Hasson and colleagues have studied the basis of the human communication system and narrative processing in the brain (Lerner et al., 2011; Silbert et al., 2014; Yeshurun et al., 2021). The human communication system is an effective storyteller and it does record an individual's memories, ideas, and dreams and transmits them to the brains of other people's communication systems. Similarly, like watching a Western film, also when listening to a meaningful story, the participant's brain showed similar activation patterns (i.e., synchronization) during the story listening. This occurred even when the same story was presented in Russian to subjects who were native speakers of Russia (Honey et al., 2012). Synchronization in higher-order brain regions, such as frontal, temporal, and parietal lobes, occurs regardless of the specific format of the narrative, e.g., textual or visual (Tikka et al., 2018). In other words, the meaning of the story (semantic structure) activates the human brain in similar ways even though the story is presented in a different syntax. More broadly, it is about a human's capability to compute holistic meanings in their surroundings (=computational meaningfulness) and this process operates mostly based on meanings. However, BIASBEHA-approach operates almost exclusively at the level of stimulus forms and syntaxes.

Furthermore, Hasson and colleagues have found that the Default Mode Network (DMN) in the brain has an essential role on the individual level when an individual integrates extrinsic and intrinsic information and when s/he tries to establish shared meaning, communication tools, shared narratives, and social networks (Kauttonen et al., 2018; Yeshurun et al., 2021). DMN is usually considered an "intrinsic" region, specializing in internally oriented mental processes such as daydreaming, reminiscing, future planning, and creativity (Raichle et al., 2001; Heinonen et al., 2016). DMN with other brain networks together forms the comprehension system, which allows the formation of the meaning of the narrative on individual levels and allows it to couple across the speaker's and listener's minds during the production and comprehension of the same narrative. Nevertheless, this common ground for understanding breaks easily, when a certain part of the story is not understandable to the listener or if some part of the element does not belong in the story (Lerner et al., 2014; Yeshurun et al., 2017b). Elements that disturb the understanding of the story include, for example, scrambled sentences, nonsense sounds, and speaking sentences too quickly (Lerner et al., 2014). Even one unclear word can make it difficult to interpret the whole story (Zadbood et al., 2021).

Moreover, certain types of cultural products, such as stories, films, pieces of music, and speeches by well-known persons, cause the meaningful areas of people's brains to activate in a very similar way (Schmälzle et al., 2015; Sharot and Garrett, 2016; Tikka et al., 2018; Zadbood et al., 2021). However, differences in people's beliefs can substantially impact their interpretation of a series of events. When researchers manipulated participants' beliefs in an fMRI study, this led two groups of participants to interpret the same narrative in different ways. They found that responses in the communication network of the brain tended to be similar among people who shared the same interpretation, but different from those of people with an opposing interpretation (Yeshurun et al., 2017b). This study showed that brain responses to the same narrative context tend to cluster together among people who share the same views. Similarly, small changes in the word of a story can lead to dramatically different interpretations of narratives among people despite the grammatical structure being similar across stories (Yeshurun et al., 2017a).

Confirmation bias and framing effect as artifacts of impoverished experimental conditions

The brain studies described above give indications that human behavior is guided by the principle of meaningfulness. This sense-making process gives weight to certain features of the context at the expense of other features. The human brain combines incoming sensory information with prior intrinsic information—i.e. mental simulations in memory—to form rich, context-dependent models of contexts as they unfold over time (Yeshurun et al., 2021). The task of people's brains is not to copy the physical world as accurately as possible via the senses but to support and participate in useful behaviors (Purves et al., 2015; Suomala, 2020; Suomala and Kauttonen, 2022).

Most previous studies of BIADBEHA literature assume discrete trials with no reference to participants' real-life contexts. In addition, the experiments often are organized in ways, in which a subject chooses between only two options. In addition, these options are usually unfamiliar to participants and they cannot learn the meanings of these options. Therefore, the results according to the heuristics and biases framework relating to confirmation bias and framing effects give a too pessimistic picture of human behavior. When we take as a starting point the human ability to survive and adapt to countless life contexts, experiences of meaning and complexity enter the explanatory pattern. Some of the reason for this impoverished experimental tradition is a consequence of the fact that in the past it has been very difficult to study people in meaningful experimental settings. Today, the situation is different and as we described above, researchers can create real-like experiments, in which human participants could feel these situations are meaningful.

Previous examples showed, how it is possible to bring the multidimensionality of real contexts to brain studies and collect brain data in these situations in real time while the subject construct representations of contexts or solves various tasks in these experiments. In everyday life, a multitude of cognitive functions and the brain networks that subserve them are seamlessly and dynamically integrated (Snow and Culham, 2021). Rather than trying to isolate stimulus or task features, the idea of data-driven analysis strategies is that features that co-occur in the real world are likely jointly represented in brain organizational principles. When studying the fluctuations of human brain activations with fMRI—as previously described studies above—a huge amount of data is obtained from each subject. While the results based on this big data is sometimes difficult to interpret (i.e., difficult to explain the phenomenon behind the data), the benefits of enormous data from people's brain are, that it can generalize to real-life situations and the ability to predict people's choices in real-life situations (Knutson and Genevsky, 2018; Doré et al., 2019).

The term big data often refers to amounts of datasets that are enormous orders of magnitude larger than the datasets that behavioral scientists work with. In this case, data sets are sized terabytes or even petabytes in size (Yarkoni and Westfall, 2017). Similarly, the applications of big data have increased about people's behavior. The possibility to access mobile and online data, coupled with a collect of enormous archival datasets from social networks and other websites, means that studies based on sample sizes of tens of thousands of participants (Schulz et al., 2019) to even sample sizes of millions of participants (Yarkoni and Westfall, 2017) is today possible. In addition to the fact that big data can be used to predict people's future behavior (Knutson and Genevsky, 2018; Doré et al., 2019), its great advantage is that they provide a natural guard against overfitting (Yarkoni and Westfall, 2017; Hasson et al., 2020). The larger the data, the more representative it is of the population's real behavior it is drawn from and it becomes increasingly difficult for a statistical model to capitalize on patterns that occur in the training data but not in the broader population (Yarkoni and Westfall, 2017). An essential challenge for this situation is how to analyze such enormous amounts of data. The development of machine learning algorithms gives tools to solve this challenge (Suomala and Kauttonen, 2022).

Machine learning algorithms for analyzing multidimensional data relating to human behavior

How do the above complexity and multidimensionality affect designing and executing behavioral experiments? To describe, explain and predict human behavior better than before, it is useful to collect big datasets and analyze these data with data-driven methods and machine-learning algorithms. In recent years, machine learning has been able to solve difficult problems in many disciplines (Suomala and Kauttonen, 2022). Indeed, cognitive neuroscience is finally at a crossroads where we have enough data to start understanding brain-behavior associations (Zhou and Zuo, 2023). Together with increasing computational power and data set availability have led to breakthroughs in machine learning and artificial intelligence. Illustrative of this development is DeepMind's program AlphaFold, which can predict the shape of almost all proteins based on their amino-acid sequences (Callaway, 2020). This problem has been biology's grandest challenge for decades. Similar progress has been found in the context of geology (Beroza et al., 2021).

Machine learning algorithms allow researchers to fit large sets of parameters including both linear and non-linear functions and a goal state. When a large amount of data is given to these algorithms, they can find approximated functions that best explain the final result. In this way, for example, the amino acid chains associated with each protein pattern have been found. Machine learning is useful in understanding complex phenomena—like human behavior—in the following ways (Glaser et al., 2019; Suomala and Kauttonen, 2022). It helps to build better predictive models, identify predictive variables by applying regularization and finding causal relationships, benchmark linear and non-linear models, and serve as a model of the brain/mind to compare against algorithms. Due to the complexity of behavioral and neurophysiological datasets that can be both non-linear and recurrent, it is beneficial to apply machine learning methods that can extract meaningful dynamics and structures (Glaser et al., 2019).

The classical statistical modeling—which BIASBEHA uses almost exclusively—relies on inference rather than predictive power, and is insufficient when trying to find working principles of neurophysiology and behavior of humans (Yarkoni and Westfall, 2017; Jolly and Chang, 2019; Hasson et al., 2020). In a recent study by Schrimpf et al. (2021), researchers demonstrated that specific language models based on deep neural networks and transformer architecture could predict human neural and behavioral responses to linguistic input with perfect predictivity relative to the noise ceiling. The researcher suggests that "testing model ability to predict neural and behavioral measurements, dissecting the best-performing models to understand which components are critical for high brain predictivity, developing better models leveraging this knowledge, and collecting new data to challenge and constrain the future generations of neutrally plausible models of language processing" (Schrimpf et al., 2021). We argue that a similar approach should be pursued to other behavior as well beyond language. With enough data, artificial neural networks can handle the messy complexities of the natural world, including nonlinearities, redundancies, and interactions, as does the brain itself (Snow and Culham, 2021).

To make the discussion of impoverished experiments, irrational decisions, multidimensionality, and usefulness of machine learning techniques more concrete, let us consider an illustrative example of a hypothetical behavioral experiment. Imagine that an investigator wants to find out how the need and cost affect a decision to buy a certain product. The investigator asks 400 people how much they need this product (variable X) and whether they would buy the product at a specific price (variable Y). For simplicity, let us assume that these two variables are on an arbitrary scale between 0 (minimum value) and 1 (maximal value). The result is depicted in Figure 1A. The decision boundary appears clean and can be fitted well using a linear logistic regression model with 2 parameters. Using a typical 80–20 train-test data split (i.e., 80% for model training and 20% for testing), the error rate is 3.4%. Now, imagine another scenario where the same survey is performed by a brick-and-mortar shopkeeper, and the responders are expected to come by physically and buy the product. Now the physical distance between the shop and the customer (variable Z) will be a new variable. As depicted in Figure 1B, the decision boundary now appears as a non-linear function of the three variables. If this new data is plotted on X-Y plane, omitting Z, data appear noisy and some decisions irrational; even with a very high need for the product (close to 1) and very low product price (close to 0), some buying decisions are still negative and wise-versa. If we try to fit a model to this lower-dimensional data, results are poor as neither linear nor non-linear models work well. This is demonstrated in Figure 1C using linear (3 parameters) and quadratic (5 parameters) logistic regression models, and a neural network classifier model (3 hidden layers, 88 parameters). The models resulted in testing error rates 18.9%, 14.9%, and 14.9%. However, when all variables are included in the model, a good approximation of the original decision boundary can be found using a neural network model (98 parameters, error rate 0%) as shown in Figure 1D.

With the above example, we highlighted three aspects: context-dependent decision making, the difference between controlled (laboratory) experiments vs. messy complexities of real-life behavior, and the usefulness of machine learning and data-driven analysis favoring predictive power over model simplicity. In real-life scenarios, human decisions are affected by factors that are difficult to anticipate and emulate in impoverished, highly-controlled experimental settings. What may appear as irrational decisions in the second situation, are in reality rational when considering the constraints of real life, which in this case was the effort needed to buy the product. This highlights the importance of the multidimensional nature of ecological decision-making. Of course, our example is an oversimplification as a researcher cannot collect a dataset with all possible variables that could affect human behavior. However, this difficulty is not an excuse to omit ecological data collection completely.

As a summary, we may conclude that tightly-controlled (laboratory) experiments are useful for testing hypotheses about the contributions of components, e.g., which variables should be included in a model, ecological experiments are useful for testing whether those hypotheses generalize to natural settings, and for generating new hypotheses that consider the complexities of the organism in its environment (Nastase et al., 2020; Snow and Culham, 2021). Hypotheses should be formulated with ecological considerations in mind and rather than constraining data collection, data should be collected in representative contexts for the ecological behaviors that you want to study (Nastase et al., 2020).

Summary and conclusion

The article describes typical BIASBEHA studies relating to confirmation bias and framing effects. Whereas these studies have shown that human reasoning differs decisively from the EUT's concept of rationality, we presented a more realistic view of human rationality. We share the view of Gigerenzer (2018) to omit the ideas of irrationality and bias-centric view in behavioral economics, however, we need to take steps further toward life-like experimental settings and predictive modeling.

According to our approach, human is rational, because they can compute meaningful constellations and produce mental simulations of these, i.e., behave according to the principle of computational meaningfulness. Then rationality means firstly, that the human brain makes different contexts understandable by recognizing the relative importance of different elements in these contexts by optimizing multidimensional information relating to these contexts (Hofstadter, 1979; Hasson et al., 2020). Secondly, it means that a human can respond to contexts very flexibly and can make sense of ambiguous or contradictory messages. Third, it means that an individual can set complex goals and finally, it means that an individual can achieve these goals.

To understand human behavior and its multidimensionality, we need to study human behavior in real-life contexts. We presented some fMRI-studies, which have successfully shown, how using multidimensional data collected from real-like situations (by using videos, stories, real advertisements, and real health messages) can help our understanding and help to predict human behavior in real-life contexts. By using multidimensional stimuli and machine learning methodology we can go toward a better theory of human behavior. This means moving away from overly simplified, few-parameter models that generalize poorly with actual behavior and between subjects, and explaining behavior with a bias when decisions are meaningful from an individual's point of view. One practical way to do this is to take advantage of immersive VR and AR technologies that allow building experiments closer to ecological conditions while also allowing experimental control.

Formalizing behavioral theories using neuroscientific and computational models provides a way to overcome the Flatland fallacy through the consideration of high-dimensional explanations of behavioral phenomena. Jolly and Chang (2019, p. 442) argue: "We

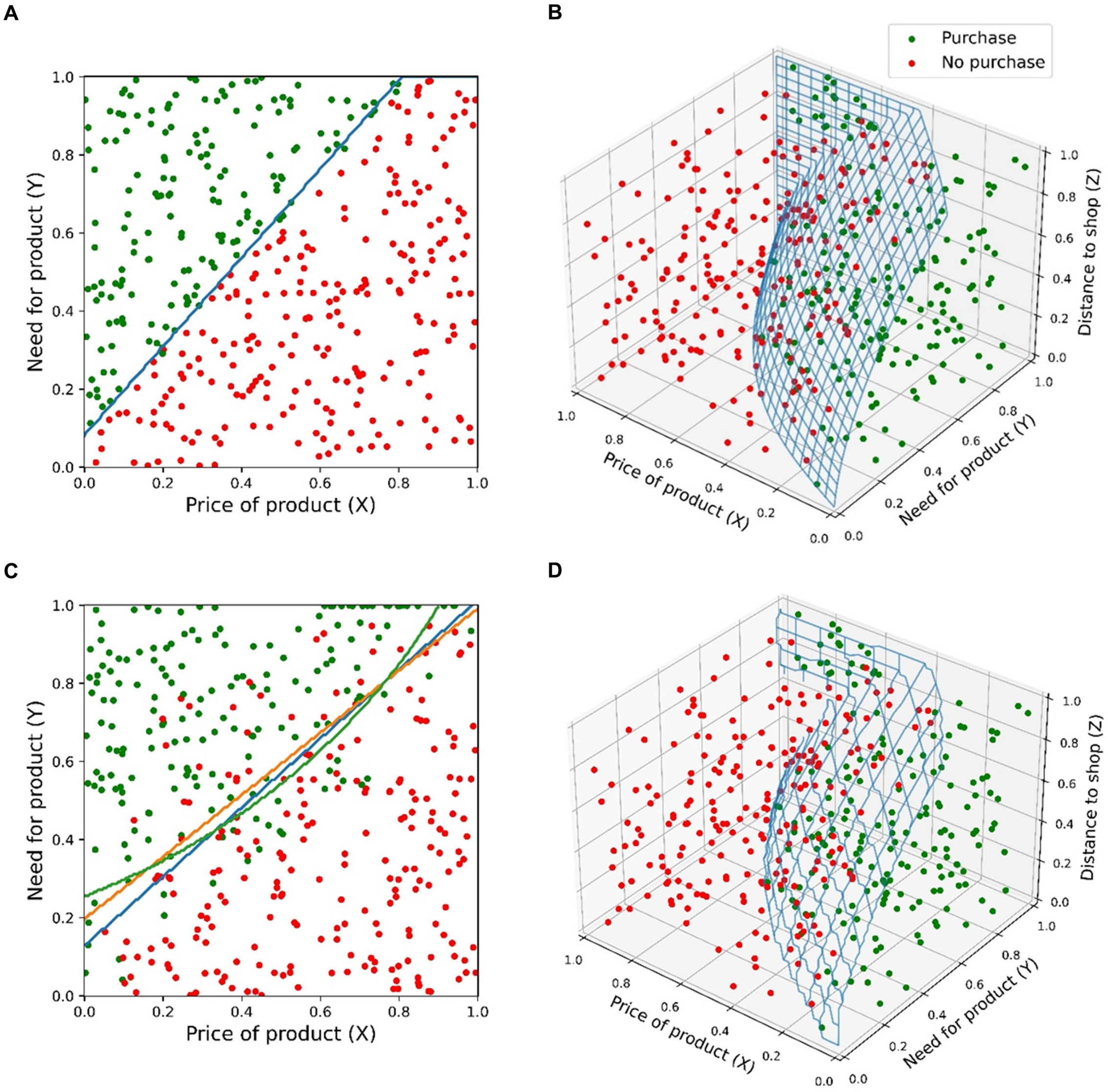


FIGURE 1
Hypothetical illustration of a decision to buy a certain product surveyed from 400 respondents. **(A)** Survey results in a laboratory setting depend on only two parameters: Price (X) and need (Y) for the product. Decision boundary fitted using a linear logistic regression model with red and green points corresponding to negative and positive decisions to buy. **(B)** A repeat of the experiment outside the laboratory with a third variable (Z) as a customer distance to the shop. The decision boundary is a complex, non-linear function. **(C)** Three models fitted to data with only two parameters included; models are linear (orange), quadratic (green), and neural network (blue). **(D)** Neural network model fitted to the full data with all three variables.

believe the use of computational models will likewise better enable researchers to capture this complexity within psychological theories." We agree and this article aims to sketch the theory of human behavior based on the principle of computational meaningfulness.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abbott, E. A. (2019). *Flatland: A romance of many dimensions*. London: Bibliotech Press.

Abeler, J., Falk, A., Goette, L., and Huffman, D. (2011). Reference points and effort provision. *Am. Econ. Rev.* 101, 470–492. doi: 10.1257/aer.101.2.470

Ariely, D. (2009). *Predictably irrational: The hidden forces that shape our decisions. 3rd* Edn. New York: Harper Collins Publ.

Austerweil, J. L., and Griffiths, T. L. (2008). A rational analysis of confirmation with deterministic hypotheses. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (Vol. 30).

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., et al. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.* 513, 532–541. doi: 10.1002/cne.21974

Baron, J. (2008). *Thinking and deciding. 4th* Edn. London: Cambridge University Press.

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosoph Transac R Soc B Biol Sci* 364, 1281–1289. doi: 10.1098/rstb.2008.0319

Baum, E. B. (2004). *What is thought?* Cambridge, Mass: MIT Press.

Beroza, G. C., Segou, M., and Mostafa Mousavi, S. (2021). Machine learning and earthquake forecasting—next steps. *Nat. Commun.* 12:4761. doi: 10.1038/s41467-021-24952-6

Berthet, V. (2021). The measurement of individual differences in cognitive biases: a review and improvement. *Front. Psychol.* 12:630177. doi: 10.3389/fpsyg.2021.630177

Berthet, V. (2022). The impact of cognitive biases on professionals' decision-making: a review of four occupational areas. *Front. Psychol.* 12:802439. doi: 10.3389/fpsyg.2021.802439

Bibas, S. (2004). Plea bargaining outside the shadow of trial. *Harv. Law Rev.* 117, 2463–2547. doi: 10.2307/4093404

Bossaerts, P., and Murawski, C. (2017). Computational complexity and human decision-making. *Trends Cogn. Sci.* 21, 917–929. doi: 10.1016/j.tics.2017.09.005

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychol. Rev.* 62, 193–217. doi: 10.1037/h0047470

Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* 588, 203–204. doi: 10.1038/d41586-020-03348-4

Camerer, C., Babcock, L., Loewenstein, G., and Thaler, R. (1997). Labor supply of new York City cabdrivers: one day at a time. *Q. J. Econ.* 112, 407–441. doi: 10.1162/003355397555244

Churchland, P. S. (2002). *Brain-wise: Studies in neurophilosophy*. Cambridge, Mass: MIT Press.

Clayton, A. (2021). *Bernoulli's fallacy: Statistical illogic and the crisis of modern science.* New York: Columbia University Press.

Cohen, A. L., Sidlowski, S., and Staub, A. (2017). Beliefs and Bayesian reasoning. *Psychon. Bull. Rev.* 24, 972–978. doi: 10.3758/s13423-016-1161-z

Cook, J., and Lewandowsky, S. (2016). Rational irrationality: modeling climate change belief polarization using Bayesian networks. *Top. Cogn. Sci.* 8, 160–179. doi: 10.1111/tops.12186

Corner, A., Harris, A., and Hahn, U. (2010). Conservatism in belief revision and participant skepticism. In *Proceedings of the annual meeting of the 32th annual conference of the cognitive science society*, (Vol. 32).

Cushman, F., and Gershman, S. (2019). Editors' introduction: computational approaches to social cognition. *Top. Cogn. Sci.* 11, 281–298. doi: 10.1111/tops.12424

DeCasper, A. J., and Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behav. Dev.* 9, 133–150. doi: 10.1016/0163-6383(86)90025-1

Doré, B. P., Scholz, C., Baek, E. C., Garcia, J. O., O'Donnell, M. B., Bassett, D. S., et al. (2019). Brain activity tracks population information sharing by capturing consensus judgments of value. *Cereb. Cortex* 29, 3102–3110. doi: 10.1093/cercor/bhy176

Falk, E. B., Berkman, E. T., and Lieberman, M. D. (2012). From neural responses to population behavior: neural focus group predicts population-level media effects. *Psychol. Sci.* 23, 439–445. doi: 10.1177/0956797611434964

Falk, E. B., Berkman, E. T., Whalen, D., and Lieberman, M. D. (2011). Neural activity during health messaging predicts reductions in smoking above and beyond self-report. *Health Psychol.* 30, 177–185. doi: 10.1037/a0022259

Falk, E. B., Morelli, S. A., Welborn, B. L., Dambacher, K., and Lieberman, M. D. (2013). Creating buzz: the neural correlates of effective message propagation. *Psychol. Sci.* 24, 1234–1242. doi: 10.1177/0956797612474670

Gabaix, X., Laibson, D., Moloche, G., and Weinberg, S. (2006). Costly information acquisition: experimental analysis of a Boundedly rational model. *Am. Econ. Rev.* 96, 1043–1068. doi: 10.1257/aer.96.4.1043

Gächter, S., Orzen, H., Renner, E., and Starmer, C. (2009). Are experimental economists prone to framing effects? A natural field experiment. *J. Econ. Behav. Organ.* 70, 443–446. doi: 10.1016/j.jebo.2007.11.003

Gallistel, C. R. (2009). "The neural mechanisms that underlie decision making" in *Neuroeconomics*. eds. P. W. Glimcher, C. F. Camerer, E. Fehr and R. A. Poldrack (Oxford: Elsevier), 417–424.

Gallistel, C. R. (2017). The coding question. *Trends Cogn. Sci.* 21, 498–508. doi: 10.1016/j.tics.2017.04.012

Gallistel, C. R. (2020). Where meanings arise and how: building on Shannon's foundations. *Mind Lang.* 35, 390–401. doi: 10.1111/mila.12289

Gallistel, C. R., and Matzel, L. D. (2013). The neuroscience of learning: beyond the Hebbian synapse. *Annu. Rev. Psychol.* 64, 169–200. doi: 10.1146/annurev-psych-113011-143807

Geary, D. C. (2005). *The origin of mind: Evolution of brain, cognition, and general intelligence. 1st* Edn. Washington, DC: American Psychological Association.

Genevsky, A., and Knutson, B. (2015). Neural affective mechanisms predict market-level microlending. *Psychol. Sci.* 26, 1411–1422. doi: 10.1177/0956797615588467

Genevsky, A., Yoon, C., and Knutson, B. (2017). When brain beats behavior: Neuroforecasting crowdfunding outcomes. *J. Neurosci.* 37, 8625–8634. doi: 10.1523/JNEUROSCI.1633-16.2017

Gershman, S. J. (2019). How to never be wrong. *Psychon. Bull. Rev.* 26, 13–28. doi: 10.3758/s13423-018-1488-8

Gershman, S. J. (2021). *What makes us smart: The computational logic of human cognition*. Princeton: Princeton University Press.

Gershman, S. J. (2023). The molecular memory code and synaptic plasticity: a synthesis. *Biosystems* 224:104825. doi: 10.1016/j.biosystems.2022.104825

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. (2015). Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. *Science* 349, 273–278. doi: 10.1126/science.aac6076

Gershman, S. J., and Niv, Y. (2013). Perceptual estimation obeys Occam's razor. *Front. Psychol.* 4, 1–11. doi: 10.3389/fpsyg.2013.00623

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Gigerenzer, G. (2018). The bias bias in behavioral economics. *Rev Behav Econ* 5, 303–336. doi: 10.1561/105.00000092

Glaser, J. I., Benjamin, A. S., Farhoodi, R., and Kording, K. P. (2019). The roles of supervised machine learning in systems neuroscience. *Prog. Neurobiol.* 175, 126–137. doi: 10.1016/j.pneurobio.2019.01.008

Grice, H. P. (1975). "Logic and conversation" in *Syntax and semantic, 3: Speech acts*. eds. P. Cole and J. L. Morgan (New York: Academic Press), 41–58.

Hamilton, L. S., and Huth, A. G. (2020). The revolution will not be controlled: natural stimuli in speech neuroscience. *Lang Cogn Neurosci* 35, 573–582. doi: 10.1080/23273798.2018.1499946

Hasson, U., Nastase, S. A., and Goldstein, A. (2020). Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron* 105, 416–434. doi: 10.1016/j.neuron.2019.12.002

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science* 303, 1634–1640. doi: 10.1126/science.1089506

Heath, C., Larrick, R. P., and Wu, G. (1999). Goals as reference points. *Cogn. Psychol.* 38, 79–109. doi: 10.1006/cogp.1998.0708

Heinonen, J., Numminen, J., Hlushchuk, Y., Antell, H., Taatila, V., and Suomala, J. (2016). Default mode and executive networks areas: association with the serial order in divergent thinking. *PLoS One* 11:e0162234. doi: 10.1371/journal.pone.0162234

Hendrickson, A. T., Navarro, D. J., and Perfors, A. (2016). Sensitivity to hypothesis size during information search. *Decision* 3, 62–80. doi: 10.1037/dec0000039

Hofstadter, D. R. (1979). *Gödel*, Escher, Bach: An eternal golden braid. Basic Books.

Hofstadter, D. R. (2001). "Epilogue: analogy as the Core of cognition" in *The analogical mind. Perspectives from cognitive science*. eds. D. Gentner, K. J. Holyoak and B. N. Kokonov (New York: The MIT Press), 499–538.

Holleman, G. A., Hooge, I. T. C., Kemner, C., and Hessels, R. S. (2020). The 'real-world approach' and its problems: a critique of the term ecological validity. *Front. Psychol.* 11:721. doi: 10.3389/fpsyg.2020.00721

Honey, C. J., Thompson, C. R., Lerner, Y., and Hasson, U. (2012). Not lost in translation: neural responses shared across languages. *J. Neurosci.* 32, 15277–15283. doi: 10.1523/JNEUROSCI.1800-12.2012

Jaynes, E. T. (2003). *Probability theory: The logic of science*. New York: Cambridge University Press.

Jern, A., Chang, K. K., and Kemp, C. (2014). Belief polarization is not always irrational. *Psychol. Rev.* 121, 206–224. doi: 10.1037/a0035941

Johnson-Laird, P. N., and Wason, P. C. (1970). A theoretical analysis of insight into a reasoning task. *Cogn. Psychol.* 1, 134–148. doi: 10.1016/0010-0285(70)90009-5

Jolly, E., and Chang, L. J. (2019). The flatland fallacy: moving beyond low-dimensional thinking. *Top. Cogn. Sci.* 11, 433–454. doi: 10.1111/tops.12404

Kahan, D. M., Peters, E., Dawson, E. C., and Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behav Public Policy* 1, 54–86. doi: 10.1017/bpp.2016.2

Kahneman, D. (2003). Maps of bounded rationality: psychology for behavioral economics. *Am. Econ. Rev.* 93, 1449–1475. doi: 10.1257/000282803322655392

Kahneman, D. (2011). *Thinking, fast and slow*. *1st* Edn. New York: Farrar, Straus and Giroux.

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47:263. doi: 10.2307/1914185

Kahneman, D., and Tversky, A. (1984). Choices, values, and frames. *Am. Psychol.* 39, 341–350. doi: 10.1037/0003-066X.39.4.341

Kauttonen, J., Hlushchuk, Y., Jääskeläinen, I. P., and Tikka, P. (2018). Brain mechanisms underlying cue-based memorizing during free viewing of movie memento. *Neuro Image* 172, 313–325. doi: 10.1016/j.neuroimage.2018.01.068

Klayman, J., and Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychol. Rev.* 94, 211–228. doi: 10.1037/0033-295X.94.2.211

Knutson, B., and Genevsky, A. (2018). Neuroforecasting aggregate choice. *Curr. Dir. Psychol. Sci.* 27, 110–115. doi: 10.1177/0963721417737877

Kőszegi, B. (2010). Utility from anticipation and personal equilibrium. *Econ. Theory* 44, 415–444. doi: 10.1007/s00199-009-0465-x

Koszegi, B., and Rabin, M. (2006). A model of reference-dependent preferences. *Q. J. Econ.* 121, 1133–1165. doi: 10.1093/qje/121.4.1133

Kuhn, T. S. (1996). *The structure of scientific revolutions (3rd ed)*. Chigago: University of Chicago Press.

Kunda, Z. (1990). The case for motivated reasoning. *Psychol. Bull.* 108, 480–498. doi: 10.1037/0033-2909.108.3.480

Lakatos, I. (1970). "Falsification and the methodology of scientific research Programmes" in *Critisism and the growth of knowledge*. ed. S. G. Harding (London: Cambridge University Press), 91–195.

Leonard, T. C. (2008). Richard H. Thaler, Cass R. Sunstein, nudge: Improving decisions about health, wealth, and happiness. *Const Polit Econ* 19:293. doi: 10.1007/s10602-008-9056-2

Leong, L. M., McKenzie, C. R. M., Sher, S., and Müller-Trede, J. (2017). The role of inference in attribute framing effects: inference in attribute framing effects. *J. Behav. Decis. Mak.* 30, 1147–1156. doi: 10.1002/bdm.2030

Lerner, Y., Honey, C. J., Katkov, M., and Hasson, U. (2014). Temporal scaling of neural responses to compressed and dilated natural speech. *J. Neurophysiol.* 111, 2433–2444. doi: 10.1152/jn.00497.2013

Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* 31, 2906–2915. doi: 10.1523/JNEUROSCI.3684-10.2011

Levin, I. P., and Gaeth, G. J. (1988). How consumers are affected by the framing of attribute information before and after consuming the product. *J. Consum. Res.* 15:374. doi: 10.1086/209174

Levin, I. P., Schneider, S. L., and Gaeth, G. J. (1998). All frames are not created equal: a typology and critical analysis of framing effects. *Organ. Behav. Hum. Decis. Process.* 76, 149–188. doi: 10.1006/obhd.1998.2804

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. *J. Pers. Soc. Psychol.* 37, 2098–2109. doi: 10.1037/0022-3514.37.11.2098

Louie, K., and De Martino, B. (2014). "The neurobiology of context-dependent valuation and choice" in *Neuroeconomics* (Oxford: Elsevier), 455–476.

Mckenzie, C. R. M. (2005). "Judgment and decision making" in *Handbook of cognition*. eds. K. Lamberts and R. Goldstone (London: SAGE Publications Ltd.), 322–339.

McKenzie, C. R. M., and Nelson, J. D. (2003). What a speaker's choice of frame reveals: reference points, frame selection, and framing effects. *Psychon. Bull. Rev.* 10, 596–602. doi: 10.3758/BF03196520

McKenzie, C. M. R., Sher, S., Leong, L. M., and Müller-Trede, J. (2018). Constructed preferences, rationality, and choice architecture. *Rev Behav Econ* 5, 337–370. doi: 10.1561/105.00000091

Mobbs, D., Wise, T., Suthana, N., Guzmán, N., Kriegeskorte, N., and Leibo, J. Z. (2021). Promises and challenges of human computational ethology. *Neuron* 109, 2224–2238. doi: 10.1016/j.neuron.2021.05.021

Müller-Trede, J., Sher, S., and McKenzie, C. R. M. (2015). Transitivity in context: a rational analysis of intransitive choice and context-sensitive preference. *Decision* 2, 280–305. doi: 10.1037/dec0000037

Nastase, S. (2021). Toward a more ecological cognitive neuroscience. *Brunswik Soc Newsletter* 36, 1–23.

Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *Neuro Image* 222:117254. doi: 10.1016/j.neuroimage.2020.117254

Navarro, D. J., and Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychol. Rev.* 118, 120–134. doi: 10.1037/a0021110

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2, 175–220. doi: 10.1037/1089-2680.2.2.175

Oaksford, M., and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychol. Rev.* 101, 608–631. doi: 10.1037/0033-295X.101.4.608

Popper, K. R. (2014). *The logic of scientific discovery*. London, New York: Routledge.

Purves, D., Morgenstern, Y., and Wojtach, W. T. (2015). Perception and reality: why a wholly empirical paradigm is needed to understand vision. *Front. Syst. Neurosci.* 9, 1–10. doi: 10.3389/fnsys.2015.00156

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proc. Natl. Acad. Sci.* 98, 676–682. doi: 10.1073/pnas.98.2.676

Ratneshwar, S., Shocker, A. D., and Stewart, D. W. (1987). Toward understanding the attraction effect: the implications of product stimulus meaningfulness and familiarity. *J. Consum. Res.* 13:520. doi: 10.1086/209085

Revlin, R., Leirer, V., Yopp, H., and Yopp, R. (1980). The belief-bias effect in formal reasoning: the influence of knowledge on logic. *Mem. Cognit.* 8, 584–592. doi: 10.3758/BF03213778

Schmälzle, R., Häcker, F. E. K., Honey, C. J., and Hasson, U. (2015). Engaged listeners: shared neural processing of powerful political speeches. *Soc. Cogn. Affect. Neurosci.* 10, 1137–1143. doi: 10.1093/scan/nsu168

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., et al. (2021). The neural architecture of language: integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci.* 118:e2105646118. doi: 10.1073/pnas.2105646118

Schulz, E., Bhui, R., Love, B. C., Brier, B., Todd, M. T., and Gershman, S. J. (2019). Structured, uncertainty-driven exploration in real-world consumer choice. *Proc. Natl. Acad. Sci.* 116, 13903–13908. doi: 10.1073/pnas.1821028116

Shafir, E., and LeBoeuf, R. A. (2002). Rationality. *Annu. Rev. Psychol.* 53, 491–517. doi: 10.1146/annurev.psych.53.100901.135213

Sharot, T., and Garrett, N. (2016). Forming beliefs: why valence matters. *Trends Cogn. Sci.* 20, 25–33. doi: 10.1016/j.tics.2015.11.002

Sher, S., and McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition* 101, 467–494. doi: 10.1016/j.cognition.2005.11.001

Sher, S., and McKenzie, C. R. M. (2014). Options as information: rational reversals of evaluation and preference. *J. Exp. Psychol. Gen.* 143, 1127–1143. doi: 10.1037/a0035128

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., and Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proc. Natl. Acad. Sci.* 111, E4687–E4696. doi: 10.1073/pnas.1323812111

Simon, H. A. (1955). A behavioral model of rational choice. *Q. J. Econ.* 69:99. doi: 10.2307/1884852

Snow, J. C., and Culham, J. C. (2021). The treachery of images: how realism influences brain and behavior. *Trends Cogn. Sci.* 25, 506–519. doi: 10.1016/j.tics.2021.02.008

Stanovich, K. E., and West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *Behav. Brain Sci.* 23, 645–665. doi: 10.1017/S0140525X00003435

Suomala, J. (2020). The consumer contextual decision-making model. *Front. Psychol.* 11:570430. doi: 10.3389/fpsyg.2020.570430

Suomala, J., Hlushchuk, Y., Kauttonen, J., Heinonen, J., Palokangas, L., and Numminen, J. (2017). Distributed brain networks reflect salary offer in accordance with the prospect theory's value function. *J. Neurosci. Psychol. Econ.* 10, 167–180. doi: 10.1037/npe0000083

Suomala, J., and Kauttonen, J. (2022). Human's intuitive mental models as a source of realistic artificial intelligence and engineering. *Front. Psychol.* 13:873289. doi: 10.3389/fpsyg.2022.873289

Suomala, J., Taatila, V., Siltala, R., and Keskinen, S. (2006). Chance discovery as a first step to economic innovation. In: *Proceedings of the 28th annual conference of the cognitive science society* (pp. 2204–2209).

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. *1st* Edn. New York: Alfred A. Knopf.

Thagard, P. (1998). Ulcers and bacteria I: discovery and acceptance. *Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci* 29, 107–136. doi: 10.1016/S1369-8486(98)00006-5

Thagard, P. (2009). Why cognitive science needs philosophy and vice versa. *Top. Cogn. Sci.* 1, 237–254. doi: 10.1111/j.1756-8765.2009.01016.x

Thaler, R. H. (2016). *Misbehaving: The making of behavioural economics*. New York: W. W. Norton & Company.

Thaler, R. H., and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth and happiness* Penguin Books.

Tikka, P., Kauttonen, J., and Hlushchuk, Y. (2018). Narrative comprehension beyond language: common brain networks activated by a movie and its script. *PLoS One* 13:e0200134. doi: 10.1371/journal.pone.0200134

Todd, P. M., and Gigerenzer, G. (2007). Environments that make us smart: ecological rationality. *Curr. Dir. Psychol. Sci.* 16, 167–171. doi: 10.1111/j.1467-8721.2007.00497.x

Tomasello, M. (2014). The ultra-social animal. *Europ. J. Soc. Psychol.* 44, 187–194. doi: 10.1002/ejsp.2015

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behav. Brain Sci.* 28, 675–691. doi: 10.1017/S0140525X05000129

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131. doi: 10.1126/science.185.4157.1124

Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science* 211, 453–458. doi: 10.1126/science.7455683

Von Neumann, J., and Morgenstern, O. (2007). *Theory of games and economic behavior (60th anniversary ed)* Princeton University Press.

Warren, J. R. (2005). Helicobacter–the ease and difficulty of a new discovery. Nobel Lecture. Available at:https://www.nobelprize.org/uploads/2018/06/warren-lecture.pdf

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Q. J. Exp. Psychol.* 12, 129–140. doi: 10.1080/17470216008416717

Wason, P. C. (1968). Reasoning about a rule. *Q. J. Exp. Psychol.* 20, 273–281. doi: 10.1080/14640746808400161

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122. doi: 10.1177/1745691617693393

Yeshurun, Y., Nguyen, M., and Hasson, U. (2017a). Amplification of local changes along the timescale processing hierarchy. *Proc. Natl. Acad. Sci.* 114, 9475–9480. doi: 10.1073/pnas.1701652114

Yeshurun, Y., Nguyen, M., and Hasson, U. (2021). The default mode network: where the idiosyncratic self meets the shared social world. *Nat. Rev. Neurosci.* 22, 181–192. doi: 10.1038/s41583-020-00420-w

Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C. J., et al. (2017b). Same story, different story: the neural representation of interpretive frameworks. *Psychol. Sci.* 28, 307–319. doi: 10.1177/0956797616682029

Zadbood, A., Nastase, S. A., Chen, J., Norman, K. A., and Hasson, U. (2021). Here's the twist: how the brain updates the representations of naturalistic events as our understanding of the past changes. *Neuroscience*. doi: 10.1101/2021.09.28.462068

Zhou, Z.-X., and Zuo, X.-N. (2023). A Paradigm Shift in Neuroscience Driven by Big Data: State of art, Challenges, and Proof of Concept. doi: 10.48550/ARXIV.2212.04195

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 15 June 2023
DOI 10.3389/fpsyg.2023.1145969

Combatting negative bias: a mental contrasting and implementation intentions online intervention to increase help-seeking among individuals with elevated depressive symptomatology

Amanda R. Keeler[1, 2, 3]*, Liesl A. Nydegger[4, 5] and William D. Crano[6]

[1]Penn State Primary Care Research Laboratory, Department of Family and Community Medicine, Penn State College of Medicine, Hershey, PA, United States, [2]Depression and Persuasion Research Laboratory, School of Social Science, Policy and Evaluation, Claremont Graduate University, Claremont, CA, United States, [3]Mood Disorder Research Lab, Department of Psychiatry and Behavioral Health, Penn State College of Medicine, Hershey, PA, United States, [4]Department of Health, Behavior and Society, Johns Hopkins University, Baltimore, MD, United States, [5]Department of Kinesiology & Health Education, The University of Texas at Austin, Austin, TX, United States, [6]Institute of Health Psychology and Prevention Science, School of Social Science, Policy and Evaluation, Claremont Graduate University, Claremont, CA, United States

OPEN ACCESS

EDITED BY
Rashmi Gupta,
Indian Institute of Technology Bombay, India

REVIEWED BY
Frank Wieber,
Zurich University of Applied Sciences, Switzerland
Calandra Lindstadt,
Colorado State University, United States

*CORRESPONDENCE
Amanda R. Keeler
✉ mandakeeler@me.com

RECEIVED 16 January 2023
ACCEPTED 16 May 2023
PUBLISHED 15 June 2023

CITATION
Keeler AR, Nydegger LA and Crano WD (2023) Combatting negative bias: a mental contrasting and implementation intentions online intervention to increase help-seeking among individuals with elevated depressive symptomatology.
Front. Psychol. 14:1145969.
doi: 10.3389/fpsyg.2023.1145969

Background: There are many reasons why individuals with depression may not seek help. Among those with elevated depressive symptomatology, some previous interventions aimed at increasing help-seeking have unintentionally decreased help-seeking intentions. Beck's cognitive theory of depression posits that individuals with elevated depressive symptomatology process information differently from those without depression (i.e., increased cognitive errors, negative bias); potentially explaining the iatrogenic results of previous interventions. Mental contrasting and implementation intentions (MCII; a self-regulatory strategy) interventions have successfully influenced physical and mental health behaviors. However, MCII has not been used specifically for initiating help-seeking for depression. The goal of this research was to ascertain whether an online MCII intervention could increase *actual* help-seeking or the *intention* to seek help for depression.

Method: Two online randomized pre-post experiments were conducted to measure the primary outcome measures 2weeks post-intervention (Study 1 collected Summer 2019: information-only control ["C"], help-seeking MCII intervention ["HS"], and comparison MCII intervention ["E"]; Study 2 collected Winter 2020: "C" and "HS"). At Time 1, adults recruited from MTurk had a minimum Beck Depression Inventory (BDI-II) score of 14 (mild depressive symptoms) and were not seeking professional help.

Results: Study 1 (*N*=74) indicated that the intervention was feasible, provided preliminary support, and clarified intervention components for Study 2. Study 2 (*N*=224) indicated that the HS group reported greater *intentions* to seek help and *actual* help-seeking than the C group. Proportionally, *actual* help-seeking was more likely among individuals who received the HS intervention and either did not *perceive* themselves as depressed at Time 2 or had BDI-II scores indicating that their depressive symptomatology decreased from Time 1.

Limitations: Participation was limited to US residents who self-reported data.

Discussion: These studies indicate that a brief online MCII intervention to encourage help-seeking is feasible and preliminarily successful. Future studies should consider using ecological momentary assessment measurements to establish the temporal precedence of intervention effects and whether MCII is effective for encouraging help-seeking among individuals prone to experiencing cognitive errors who may not be experiencing negative bias (e.g., bipolar disorder or anxiety). Clinicians may find this method successful in encouraging ongoing treatment engagement.

KEYWORDS

depression, implementation intentions, mental contrasting, negative bias, help-seeking, online intervention, MCII

1. Introduction

Although depression is a serious condition that affects millions worldwide (James et al., 2018) and is a leading risk factor for suicide (World Health Organization, 2021); with help, depression can be effectively treated (Linde et al., 2015). However, many who experience symptoms of depression do not seek treatment (Mekonen et al., 2022): a recent meta-analysis indicated that spontaneous remission rarely occurs (Mekonen et al., 2022). The goal of the current studies was to investigate the utility of a brief, theory-based online intervention designed to increase help-seeking (interpersonal or professional) initiation for those with elevated depressive symptomatology.

There are many reasons why an individual with depression may not seek help, such as not knowing how to seek help (e.g., Henderson et al., 2013; Keeler and Siegel, 2016), general lack of knowledge about depression (e.g., Rüsch et al., 2011), fear of stigma (Corrigan, 2004; e.g., Henderson et al., 2013), or structural barriers (e.g., Carbonell et al., 2020). Interventions can address these issues successfully, inducing individuals to seek help, potentially even saving lives (e.g., Siegel et al., 2015; Parikh et al., 2018). However, some interventions aimed at increasing help-seeking for depression (or addressing common barriers to help-seeking) have indicated iatrogenic results (for examples see Christensen et al., 2006; Klimes-Dougan and Lee, 2010; Sin et al., 2011; Klimes-Dougan et al., 2013; Lienemann et al., 2013; Keeler and Siegel, 2016). This is particularly problematic when an intervention appeared to have been successful for individuals who were not depressed (e.g., Keeler and Siegel, 2016) and indicates the necessity to consider how individuals with depression may respond to interventions differently from general populations (Siegel et al., 2017). Beck's (Beck, 1964; Beck, 1967; Rush and Beck, 1978; Beck, 1987; Beck and Alford, 2009; Beck and Bredemeier, 2016) cognitive theory of depression (CTD) helps to explain why this phenomenon may occur.

Beck's CTD (see Clark and Beck, 2010) describes how elevated depressive symptomatology can alter how an individual may process information differently from an individual who is not depressed. Beck reasoned this is due to the depressogenic schema that involves faulty patterns in attitudes and cognitions (negative bias) leading to cognitive errors. The negative bias that leads to the negative triad of thinking negatively about themselves, the world, and the future, may amplify the perception of barriers to seeking care [e.g., not knowing how to seek help (Henderson et al., 2013; Keeler and Siegel, 2016)]. Beck's CTD illustrates the importance of choosing a sample comprised of individuals with elevated depressive symptomatology to test an intervention. Individuals without depression cannot be expected to think or experience the world in the same manner (Ingram et al., 2008). Further, Beck's CTD suggests that individuals with elevated depressive symptomatology make cognitive errors indicating an intervention requiring a decreased cognitive load may be optimal (Bowie et al., 2017). However, understanding CTD alone does not solve the problem of the iatrogenic results.

Although Gollwitzer's theory of implementation intentions (Gollwitzer, 1990; Gollwitzer, 1993; Gollwitzer and Bargh, 1996) and Oettingen and Gollwitzer's (2010) addition of mental contrasting to implementation intentions (MCII) is relatively new, the literature suggests MCII may provide an optimal theoretical basis to overcome these barriers for the proposed studies. MCII is a meta-cognitive, self-regulatory strategy that can be employed to initiate behavior change (Duckworth et al., 2013). On their own, implementation intentions are concise action plans using "if-then" statements set in advance between a given situation and the planned goal, bridging the gap between setting a goal (intentions) and outlining the exact mechanisms of how one plans to achieve the goal when a critical cue occurs (see Gollwitzer and Brandstätter, 1997; Gollwitzer et al., 2005; Gollwitzer and Sheeran, 2006). In theory, once the implementation intention is set, when the critical cues are experienced at a later time, individuals will complete the action plan quickly and without conscious effort (Bayer et al., 2009). Mental contrasting requires a modification in the *process* of how the implementation intention is *formed*. It includes using imagery to help elaborate on the positive future of goal achievement as well as on the negative reality of what is required to attain the positive future goal (Oettingen and Gollwitzer, 2010; Oettingen and Gollwitzer, 2018). Therefore, the goal of mental contrasting in implementation intentions is to motivate and prepare individuals cognitively to form and engage in implementation intentions realistically to achieve their personalized goal (Oettingen and Gollwitzer, 2010; Oettingen and Reininger, 2016). In practice, MCII has also recently been recently rereferred to as a WOOP

Abbreviations: BDI-II, Beck Depression Inventory-second edition; C, Information-Only Control; CTD, Cognitive theory of depression; E, Exercise; EMA, Ecological momentary assessment; HS, Help-seeking; ITT, Intention to Treat; MCII, Mental contrasting and implementation intentions; PP, Per Protocol; SIIS, Strength of Implementation Intentions Scale; T1, Time 1; T2, Time 2.

strategy (wish [i.e., goal], outcome [i.e., positive future of goal] obstacle [i.e., elaboration on negative reality or barrier], plan [i.e., implementation intention]) when used in interventions (e.g., Oettingen and Reininger, 2016; Gollwitzer et al., 2018; Mutter et al., 2020; Monin et al., 2021).

Although MCII has effectively encouraged a wide range of goal achievement including physical health behaviors with moderate success (see Wang G. et al., 2021), its use among specific populations with mental health concerns is comparatively sparse (Toli et al., 2016). Of the existing literature in the field of mental health, MCII has shown initial promise with increasing specific goal attainment in samples of individuals with ADHD (e.g., Gawrilow et al., 2012), depression (Fritzsche et al., 2016), and schizophrenia (e.g., Sailer et al., 2015). Additional implementation intentions studies focused on general mental health populations including reducing self-harm (Armitage et al., 2016). Using MCII techniques may compensate for any personality features (e.g., perfectionism) that can otherwise lessen the effectiveness of forming implementation intentions alone due to mental contrasting's ability to enhance the strength of commitment to forming and following through with implementation intentions (Oettingen and Gollwitzer, 2010). The focus on remaining realistic about the barriers to goal achievement and finding ways to overcome them may be especially useful for encouraging individuals with mental health concerns to achieve their personal goals (Fritzsche et al., 2016). This is particularly true for those experiencing the negative bias and cognitive errors associated with depression that can hinder the ability to process positive thoughts without acknowledging the current pervasive negative thoughts difficult (Beck, 1987; Clark and Beck, 2010). Despite calls for increased use of MCII interventions among individuals with mental health concerns (see Gollwitzer and Sheeran, 2006; Toli et al., 2016) and a theoretical model that appears well suited for use with depression, MCII remains underutilized in the current literature.

Another gap in the field is whether MCII can help encourage individuals experiencing troubling mental health symptoms to *initiate* help-seeking. The current mental health implementation intention literature is nearly devoid of help-seeking interventions, with only one study that focused on helping individuals *follow through* with a previously scheduled mental health appointment (i.e., Sheeran et al., 2007). Given that MCII inductions have been useful to help create a strong commitment to initiate other physical health behaviors (e.g., Christiansen et al., 2010), it seems plausible that using MCII to initiate help-seeking for mental health concerns may be a valid way to overcome the personality boundary conditions and ensure the individual forms sufficiently strong links between the critical cues and responses (Gollwitzer and Sheeran, 2006) through the more intense inductions used with MCII.

Based on the rationale that MCII shows promise to overcome the negative bias as proposed by CTD and that the personalized approach can be tailored to address the multitude of help-seeking barriers, this set of studies was designed to test a novel online intervention. The overarching goal was to increase mental health help-seeking initiation intentions and behaviors among individuals with elevated depressive symptomatology. A subset of analyses from a larger set of studies collected as part of a dissertation (Keeler, 2021), the focus of the current publication is narrowed in scope to focus exclusively on testing the efficacy of a newly designed online MCII intervention across 2 studies.

The primary goal of Study 1 was to test the feasibility of a novel online MCII help-seeking intervention to help pinpoint components that would be necessary to include in Study 2. The first study included both a comparison group that completed an MCII for an exercise goal and an information-only control group to ascertain if any group differences could result from either the length of the intervention or if completing any MCII intervention would be effective in increasing help-seeking for depression. The second study focused on testing the effects of the intervention with a larger sample once the online intervention was optimized based on Study 1 findings.

2. Study 1

There were several specific goals for the preliminary study including: (a) establishing the initial feasibility of implementing a pre-post online MCII intervention; (b) clarifying components to optimize the help-seeking MCII intervention for Study 2 including the need to use both a comparison and a control condition; and (c) testing the two hypotheses using three conditions (information-only control ["C"], help-seeking MCII intervention ["HS"], and comparison exercise MCII intervention ["E"]).

The first hypothesis focused on terminal help-seeking. The second examined if the intervention could influence help-seeking intentions, which could indicate a willingness to seek help in the future. H1: Participants in the HS group (i.e., those who received the HS MCII) intervention would be more likely to report initiation of help-seeking for depression during the intervention period than those who were in either the E MCII or the C conditions. H2: Participants in the HS group would be more likely to report greater *intentions to seek help* than those in the E MCII and the C conditions *regardless of whether they actually sought help* during the intervention.

2.1. Method

2.1.1. Participants

Participants were recruited from Amazon's Mechanical Turk platform (MTurk; Amazon Web Services, RRID: SCR_012854) with data collected from June – August 2019. All potential participants completed an IRB committee-approved online informed consent attesting to general inclusion criteria (18 or older, United States Resident, English fluent) before prescreening for full inclusion criteria in accordance with Siegel and Navarro's (2019) recommendations to prescreen MTurk populations. Individuals were prescreened for elevated depressive symptomatology (having a minimum score of 14 on the BDI-II) and current help-seeking practices for depression (not currently seeking professional help) before being immediately invited to participate in the intervention.

A G*Power (RRID: SCR_013726) analysis for a between-within repeated measures ANOVA with interactions using three groups and two time points (i.e., $f = 0.15$, $\alpha = 0.05$, $1\text{-}\beta = 0.95$, $r = 0.80$ between measurements) indicated a total sample size of 72 would be required for adequate power to detect a moderately small effect. However, to account for a 20% dropout (approximately 5 participants per condition), the initial aim was 30 participants for each condition (MCII HS, E, and C). The plan to recruit 90 individuals with elevated depressive symptomatology was comparable to previously successful

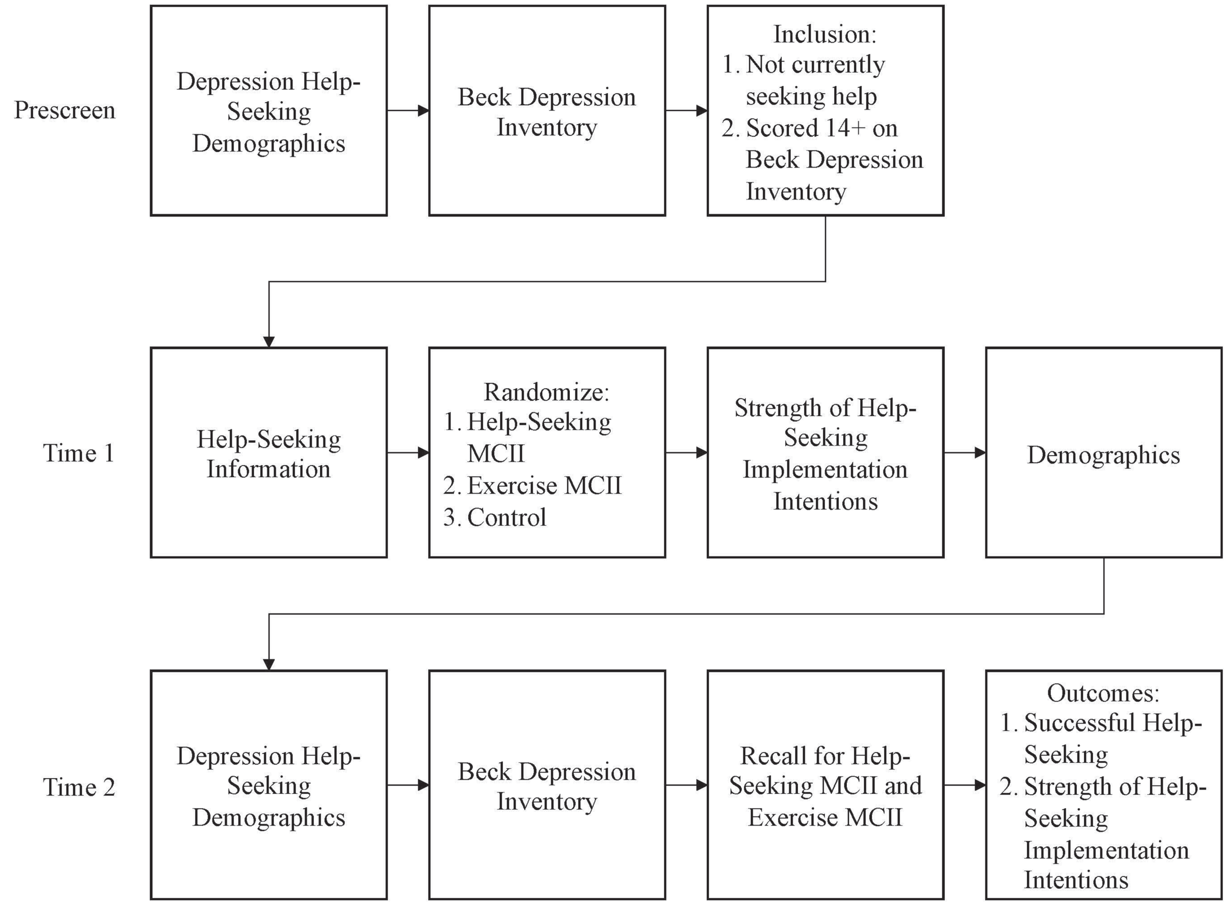

Prescreen
Depression Help-Seeking Demographics
Beck Depression Inventory
Inclusion:
1. Not currently seeking help
2. Scored 14+ on Beck Depression Inventory
Time 1
Help-Seeking Information
Randomize:
1. Help-Seeking MCII
2. Exercise MCII
3. Control
Strength of Help-Seeking Implementation Intentions
Demographics
Time 2
Depression Help-Seeking Demographics
Beck Depression Inventory
Recall for Help-Seeking MCII and Exercise MCII
Outcomes:
1. Successful Help-Seeking
2. Strength of Help-Seeking Implementation Intentions

FIGURE 1
Study 1 layout. Time 2 occurred 2 weeks post intervention.

MCII interventions' total sample sizes that included individuals with depression (e.g., $N = 47$, Fritzsche et al., 2016; $N = 36$, Sailer et al., 2015). The 2 week duration of the intervention was based on the lengths of previous MCII mental health interventions have ranged from two (Gawrilow et al., 2012) to 4 weeks (Sailer et al., 2015); the decision to err on a shorter follow-up period was due to the online nature of the study to minimize attrition.

2.1.2. Design

The study included three steps: prescreening, baseline (Time 1 [T1]), and termination (Time 2 [T2]) surveys (see Figure 1).

2.1.2.1. Prescreen

All potential participants were invited to join a new longitudinal study testing a new goal achievement strategy. Those who chose to click the link were directed to affirm their provision of consent. Consenting participants were then asked to complete the BDI-II and a battery of help-seeking questions to determine eligibility. Those with a BDI-II score of 14 or above (indicating mild depressive symptomatology) and indicated that they had not sought help for their current bout of depression from a professional were immediately invited to participate in the main longitudinal study. Those who did not qualify or chose not to continue were provided with help-seeking information and paid $0.15 for their time.

2.1.2.2. T1

Participants who agreed to continue the study read about the signs and symptoms of depression and completed several measures unrelated to the current analyses. After completing the scales, all participants were provided with information about the benefits of help-seeking for depression as well as a variety of low-cost and free help-seeking resources (see Supplementary material). All participants were asked to save this information as a memory aid in case the need to seek help arose. At this point, participants were randomly assigned to one of the three groups: (a) HS MCII intervention, (b) E comparison MCII, or (c) information-only (C). The HS and E groups each completed an MCII intervention ending with a personalized implementation intention. The HS group completed a personalized implementation intention to seek help for depression should the need arise. The E comparison group received similar information and resources regarding increasing exercise before completing a personalized implementation intention to increase their exercise by 20 min a week. The HS and E groups were directed to write down or print a copy of their implementation intention to keep as a reminder.

All groups were asked to complete two versions of the Strength of Implementation Intention Scale (SIIS) to assess strength of implementation intentions to increase exercise and one for help-seeking before proceeding to the demographic questionnaire. All participants were thanked, compensated $1.50 for their time, and notified that they would be contacted in 2 weeks to complete the follow-up.

2.1.2.3. T2

Two weeks later, all participants who consented to follow-up were notified via MTurk to complete the online battery of surveys in addition to a question to ask if they had sought help for depression during the previous 2 weeks. Both the HS and E groups were asked to reiterate the implementation intentions they established at T1 based on their reminder (HS group participants were prompted about help-seeking and the comparison E group was asked about exercise). All participants were directed to complete the help-seeking and exercise SIIS measures, asked if they had sought help for depression, and asked if they had increased their physical activity by 20 min during the previous 2 weeks. All participants were debriefed (including a tutorial on how to use the MCII technique) and paid $2.00.

2.1.3. Materials

2.1.3.1. Beck Depression Inventory-II

The BDI-II (Beck et al., 1996) is one of the most used depression assessments. The scale includes 21 groups of questions to determine levels of depressive symptomatology over the past 2 weeks. A composite score was calculated by summing the score of each question. Composite scores can range from 0 to 63 with higher scores indicating greater depressive symptomatology. Scores of 0–13 represent no to minimal, 14–19 mild, 20–28 moderate, and 29–63 severe depressive symptomatology (Beck et al., 1996).

2.1.3.2. Depression history and demographics

A series of items were used to assess participants' depression history in the screening survey. Items included: previous diagnosis of depression, current diagnosis of depression, current professional treatment for depression, and history of help-seeking from close others and mental health professionals. The T1 Survey ended with demographic items. The general items included characteristics such as gender, age, race, and insurance status that includes mental health coverage.

2.1.3.3. Mental contrasting and implementation intentions related measures

Both the HS (intervention) and the E (comparison) groups completed an MCII exercise. The HS group focused on help-seeking for depression should the need arise and the E group focused on increasing exercise by 20 min a day. Additionally, all participants completed quantitative measures of intentions to seek help for depression.

2.1.3.3.1. Induction

Two-thirds of the participants were randomly assigned to complete an MCII intervention (either help-seeking intervention [i.e., HS group] or physical activity comparison [i.e., E group]). For participants randomly assigned to complete either the HS or E MCII, this study used a multistep process implementation intention induction modeled after Sailer et al. (2015) aimed at increasing exercise for individuals with schizophrenia. All participants started by reading a short informational text that indicated behavior change was desirable, feasible, and how obstacles could be overcome. Afterward, participants were led through the following writing prompts:

- Participants identified a specific goal (the HS group were asked to pick a goal related to depression help-seeking; the E group were asked to choose a goal to increase exercise by 20 min per day).
- Asked to take a moment to imagine and write down the positive future of achieving their goal and to list four positive outcomes related to achieving goal (e.g., feeling happier or healthier).
- Mentally contrast the positive future with the current barrier or obstacle to achieving goal by listing four barriers (e.g., too tired, scared).
- Think about their biggest barrier or obstacle and write down ways to overcome it (e.g., enlist a friend's help).
- Formulate an implementation intention plan to overcome the barrier in the form of an "if-then" plan.

This plan was copied three times in accordance to Sailer et al.'s methodology (p. 5). Similar to Fritzsche et al. (2016), participants were asked to screenshot or write down their implementation intention to act as a reminder.

2.1.3.3.2. Follow-up questions

Inspired by Fleig et al. (2017), two questions assessed participants' perceptions of both the viability (i.e., does the participant have the resources to carry out their implementation intention plan) and instrumentality (i.e., belief their action plan can help them achieve the goal of seeking help if needed) of the implementation intention to seek help they created, which was proposed to influence the likelihood of implementation intentions leading to the enactment of goal-directed behavior. Both questions were asked at T1 and T2 and were rated on a 7-point Likert scale ranging from 1 ("Strongly Disagree") to 7 ("Strongly Agree").

2.1.3.3.3. Strength of implementation intentions scales

To quantitatively measure implementation intentions, we used a modified version of Nydegger's Strength of Implementation Intentions Scale (SIIS; Nydegger et al., 2013, 2017). The scale was developed to assess the perceived strength of the link between the critical cue (e.g., when, where, and specific emotional trigger) and the precise action the individual is planning in response in order to achieve their goal (Nydegger et al.,2017). The SIIS has demonstrated acceptable internal consistency reliability ($\alpha = 0.96$) with a focus on condom use and it was originally written so that it could be modified to fit different study and sample requirements by changing the target wording to be appropriate to various goals. The questions for these studies were modified with Dr. Nydegger's guidance for the goals of help-seeking (SIIS HS; 7 questions) and increasing exercise (SIIS E; 5 questions) with questions rated on a 6-point Likert scale ranging from 1 ("Strongly Disagree") to 6 ("Strongly Agree").

2.1.3.3.4. Success of MCII

At T2, all participants were asked to rate their perceived level of success related to increasing their exercise and help-seeking behaviors. Both questions were rated on a 6-point Likert scale ranging from 1 ("Strongly Disagree") to 6 ("Strongly Agree") with a "Not Applicable" option.

2.1.3.3.5. Attention checks

In addition to the prescreening and VPN/geolocation, these studies utilized both quantitative and qualitative attention checks to prevent noted issues with fraudulent MTurk data (Kennedy et al., 2020). Three scale-embedded, quantitative attention checks directed participants to select specific response options as a measure of attention. Additionally, individuals in the E and HS groups had their MCII answers examined for whether the topic of MCII's were on target, if directions were followed, and at T2, whether the participant remembered their MCII. All participants regardless of the data analysis plan (modified intention-to-treat or per-protocol) were required to pass the simple quantitative attention checks for quality control.

2.1.4. Data cleaning and analysis plan

The pre-established data analysis plan required that the hypotheses would be analyzed in two ways: modified intention-to-treat (ITT) and per-protocol (PP). The rationale for using both is that ITT provides a more conservative estimate of the effectiveness of the intervention and is preferred by the Federal Food and Drug Administration for randomized control trials (see Day, 2008; Gupta, 2011). ITT analyses include all participants who had been randomized regardless of whether they dropped out of the study. This study pre-established the modification that participants would need to pass all quantitative attention checks (see Day, 2008) for inclusion for quality control due to the online nature of the study (Kennedy et al., 2020).

For the PP analyses, participants were excluded if they dropped out of the intervention before completion or missed any of the attention checks. Additionally, for the PP analyses, the content of the MCII was examined for those in the HS and E groups to explore whether the participants appeared to take the exercise seriously (e.g., were they on topic?) and at follow-up, did the participants indicate they remembered the general theme of their MCII? When the results of the ITT and PP analyses were the same, only the ITT results are reported. Multivariate outliers were removed based on Mahalanobis distance and univariate outliers based on Cook's distance. The intervention and information-only control groups' data were compared using non-parametric measures to ensure equivalence at T1.

2.2. Study 1 results

2.2.1. Data cleaning

Data were analyzed with SPSS 27 software (IBMCorp, Released 2020; RRID:SCR_019096) in two ways: ITT and PP according to the established data plan.

2.2.1.1. ITT

Of the total individuals ($N = 981$) who responded to the MTurk post, $n = 117$ did not consent, $n = 461$ had BDI-II scores lower than 14, $n = 85$ of individuals with BDI-II scores 14 and above but sought help for depression, and $n = 179$ participants did not pass one or both attention checks. The total number of participants per ITT at T1 was $N = 139$ with $n = 38$ HS, $n = 39$ E, and $n = 62$ C agreeing to be contacted for follow-up. Due to attrition of 55 between T1 and T2, the total possible number of participants for whom follow-up was possible was 83 (BDI-II T1 = 25.92 ± 9.59 (min 14, max 59), BDI-II T2 = 23.80 ± 10.59 (min 6, max 53), 36 ± 11.7 years, 57% Female, 77% white) with $n = 19$ HS, $n = 23$ E, and $n = 41$ C.

2.2.1.2. PP

To establish the PP sample, the final T1 ITT sample was used as the initial starting point ($N = 139$). When examining the data to establish the PP sample, 55 participants were lost to attrition, eight participants were excluded based on blatantly not taking the MCII exercise seriously (e.g., "If I win the lottery, then I guess I'll exercise more") or showing no indication that they remembered forming an MCII (e.g., "I do not recall" or "I do not have a copy"). Two univariate outliers were identified using Cook's Distance; no multivariate outliers were found via Mahalanobis distance. The final PP total was $N = 74$ [HS = 17, $E = 17$, $C = 40$, BDI-II T1 = 24.74 ± 9.32 (min 14, max 59), BDI-II T2 = 23.54 ± 10.26 (min 6, max 53), 35.4 ± 11.8 years, 57% Female, 76% white].

The HS and C groups' data were compared to ensure equivalence at T1. Kruskal-Wallis for independent samples tests were used to assess group independence to determine any significant differences between groups for age, and chi-square tests were used to test for group differences in gender; no significant differences were observed (PP or ITT). See Table 1 for the full demographics and Table 2 depression help-seeking demographics for both the PP and ITT samples. Table 3 reports the descriptive information, and reliability information for the measures obtained in Study 1.

2.2.2. H1

The first hypothesis proposed that the HS group would be more likely to report that they *initiated* help-seeking for depression during the intervention period than those in either the comparison (E) or the control conditions (C). Individuals who responded with "Not Applicable" were treated as missing (ITT $n = 8$; PP $n = 2$) resulting in there being no differences between the ITT and PP participant samples and therefore, no differences in the analysis outcome. The one-way ANOVA analyses did not support this hypothesis $F(2,74) = 0.046$, $p = 0.995$; see Figure 2A for ITT and Figure 2B for PP. Individuals in the HS group were no more likely to report *actually* seeking help at T2 than those in the E or C groups.

2.2.3. H2

Two-way mixed ANOVAs were used to test if there were group differences in *intentions to seek help* for depression from T1 to T2 *regardless* of whether participants *actually* sought help. H2 predicted that the HS group would report greater intentions to seek help from T1 to T2 for depression as measured by the SIIS HS than the E and C groups. GLM repeated measures function was used to conduct the 2 (SIIS HS: T1 and T2) X 3 (group: HS, E, C) ANOVA. Sphericity can be assumed since there were only two levels of the repeated measure. Bonferroni corrections were used for *post hoc* contrasts to account for multiple testing (Field, 2018). The results of this analysis varied depending on the sampling method.

TABLE 1 Study 1 sample demographics.

Age *M* (*SD*)	PP *n* (%)	ITT Time 1 *n* (%)	ITT Time 2 *n* (%)
	35.4 (11.8)	35.7 (11.7)	36 (11.7)
Group assignment			
Help-seeking	17 (23.0)	38 (27.3)	19 (22.9)
Exercise	17 (23.0)	39 (28.1)	23 (27.7)
Control	40 (54.0)	62 (44.6)	41 (49.4)
Gender			
Male	31 (41.9)	56 (40.3)	34 (41.0)
Female	42 (56.8)	81 (58.3)	47 (56.6)
Prefer not to say	1 (1.4)	2 (1.4)	2 (2.4)
Ethnicity/Race			
African American/Black	5 (6.8)	15 (10.8)	5 (6.0)
Asian	7 (9.5)	13 (9.4)	7 (8.4)
Hispanic/Latinx	3 (4.1)	5 (3.6)	3 (3.6)
White	56 (75.7)	100 (71.9)	64 (77.1)
Other	0 (0)	2 (1.4)	1 (1.2)
Highest level of education			
Some high school	0 (0)	1 (0.7)	0 (0)
Graduated high school	11 (14.9)	17 (12.2)	13 (15.7)
Some college	18 (24.3)	40 (28.8)	20 (24.1)
Associate degree	5 (6.8)	11 (7.9)	5 (6.0)
Bachelor's degree	34 (45.9)	62 (44.6)	38 (45.8)
Master's degree or higher	6 (8.1)	8 (5.8)	7 (8.4)
Marital status			
Single	36 (48.6)	63 (45.3)	38 (45.8)
Married/committed relationship	30 (40.5)	55 (39.6)	34 (41)
Divorced	6 (8.1)	17 (12.2)	8 (9.6)
Other	0 (0)	1 (0.7)	0 (0)
Prefer not to say	2 (2.7)	3 (2.2)	3 (3.6)
Insurance includes ANY mental health			
No	20 (27.0)	43 (30.9)	22 (26.5)
Yes	42 (55.4)	74 (53.2)	46 (55.4)
I do not know	13 (17.6)	22 (15.8)	15 (18.1)

ITT, Modified intention-to-treat; PP, Per Protocol. PP $n = 74$, ITT T1 $n = 139$, ITT T2 $n = 83$.

When examining the ITT sample, the analyses indicated that there were no significant main effects for scores on the SIIS HS over time ($F(1,80) = 0.035$, $p = 0.853$, partial $\eta^2 < 0.001$) or group ($F(2,80) = 2.361$, $p = 0.101$, partial $\eta^2 = 0.06$). Additionally, there was no significant interaction between group and scores on the SIIS HS over time ($F(2,80) = 2.035$, $p = 0.137$, partial $\eta^2 = 0.05$). The results indicate that the hypothesis was not supported using the ITT sample; completing the HS MCII had no effect on SIIS scores over time. See Figure 3A for means and standard error of scores.

For the PP sample, the results indicate that completing the HS MCII influenced SIIS HS scores over time. Analyses revealed no main effect for the SIIS HS, $F(1,73) = 1.067$, $p = 0.305$, partial $\eta^2 = 0.01$, indicating that individual's scores on the SIIS HS measure did not vary significantly from T1 to T2. However, there was a main effect for group differences, $F(2,73) = 3.364$, $p = 0.04$, $\eta^2 = 0.08$), and a significant interaction between group and scores on the SIIS HS over time ($F(2,73) = 3.713$, $p = 0.029$, partial $\eta^2 = 0.09$). Examining the pairwise comparisons, there was a significant difference between the HS and C groups (*M*diff = 5.090, $p = 0.035$, 95% CI:(0.278; 9.090). There were no significant differences between the HS and the E groups, nor between the E and C groups. See Figure 3B for means and standard error of scores.

2.3. Study 1 discussion

Study 1 was designed as a preliminary study to accomplish three primary goals: (a) establish the feasibility of conducting an entirely

TABLE 2 Study 1 depression help-seeking demographics.

	PP *n* (%)	ITT T1 *n* (%)	ITT T2 *n* (%)
Have you ever believed you were depressed but did not seek help?			
No	15 (20.3)	30 (21.6)	16 (19.3)
Yes	59 (79.7)	109 (78.4)	67 (80.7)
Have you ever sought help for depression from a loved one?			
No	42 (56.8)	75 (54)	48.8 (57.8)
Yes	32 (43.2)	64 (46)	35 (42.2)
Have you ever sought help for depression from a professional?			
No	42 (56.8)	76 (54.7)	45 (54.2)
Yes	32 (43.2)	63 (45.3)	38 (45.8)
Do you believe you currently have depression?			
No	26 (35.1)	49 (35.3)	30 (36.1)
Yes	48 (64.9)	90 (64.7)	53 (63.9)
Are you currently diagnosed with depression?			
No	60 (81.1)	111 (79.9)	66 (79.5)
Yes	14 (18.9)	28 (20.1)	17 (20.5)
Have you ever been diagnosed with depression?			
No	46 (62.2)	85 (61.2)	49 (59.0)
Yes	28 (37.8)	54 (38.8)	34 (41.0)
Are you currently seeking professional help for depression?			
No	74 (100)	139 (100)	83 (100)
Yes	0 (0)	0 (0)	0 (0)

ITT, Modified intention-to-treat; PP, Per Protocol. T1, Time 1; T2, Time 2. PP *n* = 74, ITT T1 *n* = 139, ITT T2 *n* = 83.

online MCII intervention, (b) clarify areas of improvement for Study 2, and (c) test the primary intervention hypotheses. Despite high attrition, it was possible to translate an MCII intervention to work online for individuals with depression. During the intervention, there were clearly features of the study that required improvement. The following sections of the discussion briefly examine the results of the hypothesis tests and their implications for Study 2.

2.3.1. Hypotheses

It was expected that individuals in the HS group would be more likely to seek help for their depressive symptoms (H1) and would be more likely to report a greater intention to seek help over time (H2) when compared to the E and C groups. Like all the group-based analyses, H1 and H2 may have suffered by the unequal and smaller proportion of participants in the HS (ITT *n* = 19, PP *n* = 17) and E (ITT *n* = 23, PP *n* = 17) groups compared to the C group (ITT *n* = 41, PP *n* = 40).[1] Although there were no significant group differences in reported success in actual help-seeking for depression (H1), the hypothesized significant group differences and interaction was found for the PP sample in the Strength of Implementation Intentions Scale for Help-Seeking (SIIS HS). Since there were no significant differences between the E and C groups or the HS and E groups for H2, it seems reasonable to limit Study 2 to control and experimental groups. By limiting the number of groups, it will be possible to achieve greater statistical power in the forthcoming analyses.

One alternative explanation for the lack of intervention effects is whether participants in the HS group failed to achieve their goals due to a lack of perceived utility or a perceived lack of resources. Despite being conducted, an exploratory mixed ANOVA using the two questions inspired by Fleig et al. (2017), assessing instrumentality and resources was not included in the results section due to the highly speculative nature of the underpowered results (eight participants affirmed they achieved success at their goal and eight participants did not). Each of the questions were scored on a 1 ("Strongly Disagree") to 7 ("Strongly Agree") measure with average item scores of the non-successful individuals ranging from $M = 4.38$ ($SD = 2.39$) to $M = 5.75$ ($SD = 1.28$) and successful individuals were similar for both questions ranging from $M = 5.63$ ($SD = 0.916$) to $M = 6.13$ ($SD = 0.84$) indicating that individuals generally believed they saw utility in goal setting and had the resources to carry them out at both T1 and T2. However, despite the small number of participants, the analysis seemed to suggest that utility and resources did not make a significant difference in level of success, nor did it appear that the scores changed significantly over time. Although this analysis examining an alternative explanation was underpowered, and thus, the likelihood of a Type II error is high (Crano et al., 2014), the means of perceived utility and perceived resources all being rated positively indicate it was likely that ceiling effects occurred. Depending on the results of Study

[1] Although the power analysis indicated that a sample size of 72 was required and the total sample sized met this goal, the small and unequal group sizes may have resulted in inadequate power.

TABLE 3 Study 1 measure means, standard deviations, and internal consistency (if applicable).

Measure	N items	PP T1		PP T2		ITT T1		ITT T2	
		M (*SD*)	*α*	*M* (*SD*)	*α*	*M* (*SD*)	*α*	*M* (*SD*)	*α*
BDI	21	24.74 (9.32)	0.87	23.54 (10.26)	0.90	25.92 (9.59)	0.88	23.80 (10.59)	0.90
SIIS SH	–	26.64 (7.89)	–	26.47 (7.36)	–	26.62 (8.35)	–	25.98 (7.96)	–
HS	7	30.71 (5.26)	0.88	29.76 (5.89)	0.85	30.26 (6.62)	0.91	29.21 (8.22)	0.93
EX	7	27.47 (8.00)	0.95	24.94 (7.44)	0.93	27.21 (9.05)	0.96	23.48 (7.80)	0.94
C	7	24.58 (8.20)	0.93	25.73 (7.60)	0.94	24.02 (8.09)	0.93	25.88 (7.57)	0.94
SIIS EX	–	19.84 (5.24)	–	19.92 (5.77)	–	19.80 (6.32)	–	19.34 (6.34)	–
HS	5	20.47 (5.57)	0.92	20.71 (6.81)	0.96	19.50 (6.95)	0.96	19.05 (8.11)	0.98
EX	5	18.68 (5.36)	0.77	21.06 (5.30)	0.90	22.00 (5.99)	0.90	19.70 (6.34)	0.94
C	5	18.68 (5.36)	0.91	19.10 (5.51)	0.90	18.60 (5.84)	0.92	19.27 (5.55)	0.90
MCII SH success	1	–	–	3.58 (1.73)	–	–	–	3.45 (1.71)	–
MCII EX success	1	–	–	3.53 (1.66)	–	–	–	3.59 (1.71)	–

Means, standard deviations, observed alpha (if applicable). T1, Time 1; T2, Time 2; BDI, Beck Depression II inventory scores; C, information only control; SIIS, Strength of Implementation Intentions Scale; EX, Exercise; SH, Seek Help; ITT, Intend to Treat; PP, Per Protocol; MCII, Mental Contrasting and Implementation Intentions Intervention.

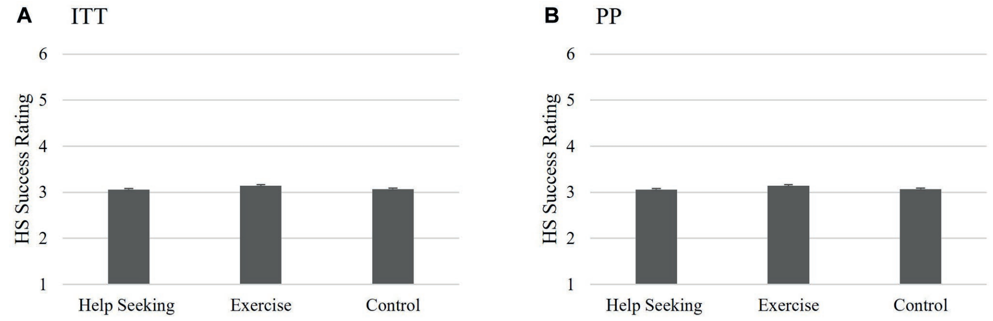


FIGURE 2
Study 1 Hypothesis 1: Success of help-seeking intervention. Non-significant interaction between groups on help-seeking success indicating that completing the HS MCII intervention had no significant effect on perceived help-seeking success 2weeks post-intervention for either the ITT **(A)** or PP **(B)** analyses. Scores reported are the means for perceived success of help-seeking scores and the bars are the standard error.

2 these should be examined again with a larger sample to rule out the alternative explanation that perceived utility or resources may influence help-seeking outcomes.

Despite the ~40% attrition between T1 and T2 for the modified ITT MTurk sample, it was possible to obtain an adequate sample of individuals with elevated depressive symptomatology based on the power calculations. Other studies have also described issues with attrition in MTurk samples (Zhou and Fishbach, 2016; Hauser et al., 2018) and stressed the importance of tempering conclusions drawn with elevated attrition levels. Several modifications to the study design were implemented to decrease attrition in Study 2.

2.3.2. Modifications for Study 2

Though the results of the hypotheses from Study 1 were weak, several key lessons were used to design Study 2. For example, the results of this preliminary study suggested answers to questions such as: (a) what level of attrition should be expected? (~40% without modifications), (b) is there evidence that new HS SIIS scale is reliable? (yes), and (c) did the data indicate the necessity for both a control and comparison group? (no). Although there was limited success with changing intentions to seek help in 2 weeks using the more liberal estimate of SIIS HS in the PP analysis, Zhou and Fishbach (2016) noted the importance of tempering the excitement due to the large percentage of attrition–especially considering the small sample size. Due to the high attrition noted in this study, the authors decided that a longer duration between Times 1 and 2 or adding a booster session between time points would only exacerbate attrition.

Planned changes to avoid low power in the second study included conducting new power analyses based on observed measure correlations from Study 1 rather than using estimates and oversampling with the expectation of attrition. Further, Study 2 could conserve power by focusing only on the two groups that illustrated significant differences: C and HS. To decrease the burden on participants, the questionnaires were shortened to only necessary measures to test for the intervention effects. After analyzing Study 1 data, it seemed plausible that an additional alternative explanation for intervention effects to seek help could be related to participants' perceptions of their depression; prompting the question to be asked at both time points in Study 2.

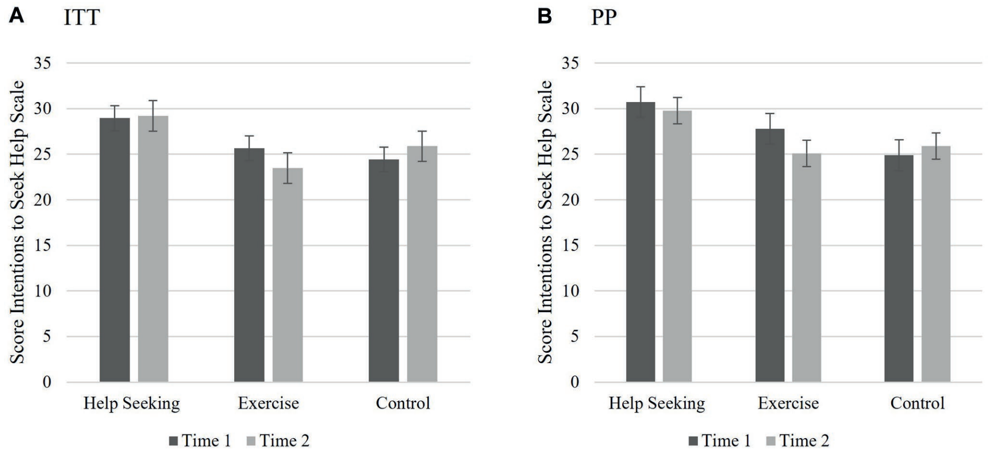


FIGURE 3
Study 1 Hypothesis 2: Changes to intentions to seek help. **(A)** ITT: Illustrating the non-significant interaction between groups and SIIS HS over time; completing the HS MCII intervention had no significant effect on intentions to seek help over time for the ITT analyses; However, **(B)** PP: indicated significant main effect for group and interaction between group and SIIS HS over time. Those who received HS intervention had greater intentions seek help across both time points compared to the other conditions. Scores reported are the means for the SIIS HS scores T1 and T2 and the bars are the standard error.

3. Study 2

With the study modifications in place, Study 2 aimed to explore the two intervention-based hypotheses explored in Study 1: H1: Participants in the HS group (i.e., those who receive the HS MCII intervention) would be more likely to report that they initiated help-seeking for depression during the intervention period than those who were in C. H2: Participants in the HS group would be more likely to report greater *intentions to seek help* (as measured by the SIIS HS) than those in C *regardless of whether they actually sought help* during the intervention.

3.1. Method and materials

3.1.1. Participants

Participants were again recruited from Amazon's MTurk (RRID:SCR_012854) with data collected between February – early March 2020. To participate in the prescreening, participants were notified that they must be an English-speaking US resident and at least 18 years old. Again, to be consistent with Siegel and Navarro's (2019) recommendation of prescreening MTurk populations rather than explicitly listing inclusion criteria for online surveys, the sample was prescreened for depressive symptomatology (having a minimum score of 14 on the BDI-II) and current help-seeking practices for depression (not currently seeking professional help) prior to being immediately invited to participate in the first part of the intervention. Individuals who participated in Study 1 were not eligible to participate. A G*power analysis for a between-within repeated measures ANOVA with interactions using two groups (HS and C), two measurements (*r* based on Study 1's correlation between T1 and T2 SIIS HS; $f = 0.15$, $\alpha = 0.05$, $1\text{-}\beta = 0.95$, $r = 0.65$), indicated a total sample size of 104 (Faul et al., 2009). Due to the large proportion of participants in the first study who did not pass the attention checks (Study 1 = 56%) and the significant attrition from T1 to T2 (40%), screening continued until 345 participants passed the initial screening.

3.1.2. Design

The design of Study 2 was a slightly modified version of Study 1. Study 2 used only the experimental (HS) and information-only control (C) groups and focused on the measures necessary for testing the intervention effects (i.e., removal of unrelated scales, additional depression demographics at T2). Otherwise, the design mirrored Study 1 (see Figure 4).

3.1.3. Measures

To reduce time and cognitive fatigue, the measures were limited to what was needed to study the intervention effects. The measures included depressive symptomatology (i.e., BDI-II, depression demographics) and MCII measurements (i.e., induction, intervention success, MCII instrumentality, MCII viability, and SIIS HS). Full descriptions of the measures are in Study 1.

3.1.4. Data cleaning and analysis plan

The data analysis plan mirrored Study 1 and included both the modified ITT and PP analyses. When both ITT and PP analyses are both significant, only the more conservative ITT results are provided in text.

3.2. Study 2 results

Of the 2,134 respondents to the MTurk post, 382 did not consent, and 1,005 had BDI-II scores lower than 14. Of the 712 who had BDI-II scores of 14 or above, 137 were excluded for

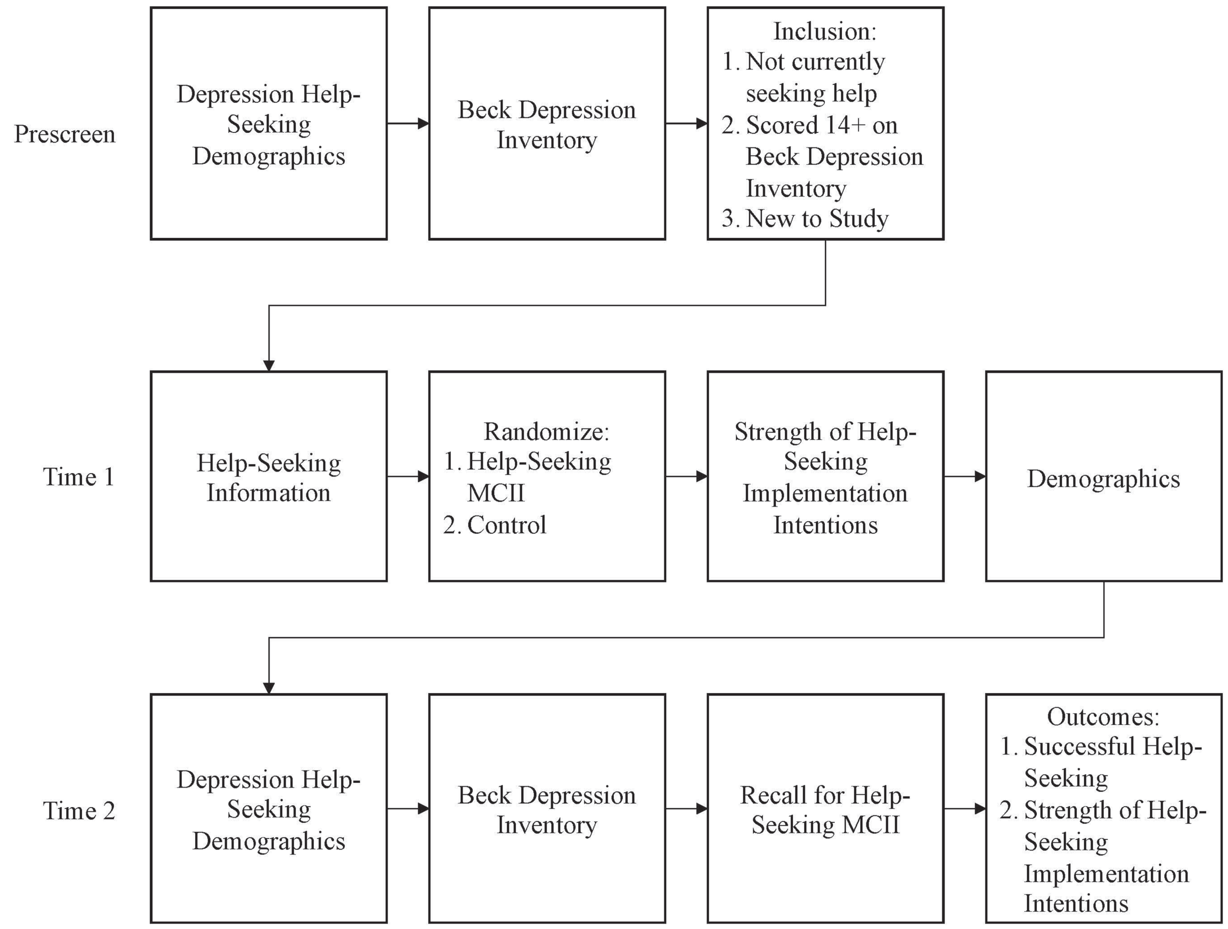

Prescreen
Depression Help-Seeking Demographics
Beck Depression Inventory
Inclusion:
1. Not currently seeking help
2. Scored 14+ on Beck Depression Inventory
3. New to Study
Time 1
Help-Seeking Information
Randomize:
1. Help-Seeking MCII
2. Control
Strength of Help-Seeking Implementation Intentions
Demographics
Time 2
Depression Help-Seeking Demographics
Beck Depression Inventory
Recall for Help-Seeking MCII
Outcomes:
1. Successful Help-Seeking
2. Strength of Help-Seeking Implementation Intentions

FIGURE 4
Study 2 layout. Time 2 occurred 2 weeks post intervention.

responding that they previously took a similar survey and 206 were disqualified for currently seeking professional help. Of the 345 individuals who consented to continue with the study, an additional 24 were removed for missing the attention check, resulting in T1 ITT total *N* = 321 (HS = 149, C = 172). For the T2 ITT analyses, 44 were lost in the HS group (32 attrition, 12 failed attention check; *n* = 105) and 44 participants in the C condition were lost to attrition (4 additional for failed attention check; *n* = 124). Therefore, the total ITT sample analyses were conducted with 228 [HS = 105, C = 123, BDI-II T1 = 24.32 ± 9.07 (min 14, max 54), BDI-II T2 = 24.01 ± 10.62 (min 2, max 54), 37.2 ± 12.1 years, 66% Female, 67% white] participants with the total attrition from T1 to T2 reduced, compared to Study 1, to 24% and quality control removals were reduced to 5%.[2]

For the PP analyses, an additional 19 participants in the HS condition were excluded after reading through the MCII for either blatantly not taking the exercise seriously (e.g., copy and pasting the question in the answer box), or indicating no recollection of completing the exercise. One univariate outlier and four multivariate outliers were removed bringing the total for the PP sample to 205 [HS = 86, C = 119, BDI-II T1 = 24.44 ± 9.41 (min 14, max 54), BDI-II T2 = 23.93 ± 10.93 (min 2, max 54), 37.2 ± 12.4 years, 68% Female, 69% white]. See Table 4 for the full demographics and Table 5 depression help-seeking demographics.

Kruskal-Wallis for independent samples was used to test for group independence to determine any significant differences between groups for age and Chi square tests were used to test for group differences in gender. No significant differences were observed (PP or ITT). Please see Tables 4, 5 for the demographics and reported depression help-seeking demographics for both the PP and ITT samples that were used for analyses. Scale analyses were completed for the study for both analysis methods (see Table 6).

[2] After an additional reminder to complete the final survey, the T2 surveys were deactivated on 3/18/20 due to the growing COVID-19 pandemic since the effects of the worsening global situation might be impossible to disentangle from the intervention effect.

TABLE 4 Study 2 sample demographics.

	PP *n* (%)	ITT Time 1 *n* (%)	ITT Time 2 *n* (%)
Age *M* (*SD*)	37.4 (12.4)	36.0 (11.8)	37.2 (12.1)
Group assignment			
Help-seeking	86 (42.0)	149 (46.4)	105 (45.9)
Control	119 (58.0)	172 (53.6)	123 (54.1)
Gender			
Male	66 (32.2)	122 (38.0)	76 (33.3)
Female	139 (67.8)	198 (61.7)	151 (66.3)
Prefer not to say	0 (0)	1 (0.3)	1 (0.4)
Ethnicity/Race			
African American/Black	17 (8.3)	31 (9.7)	19 (8.3)
Asian	21 (10.2)	33 (10.3)	26 (11.4)
Caucasian/White	142 (69.3)	213 (66.4)	153 (66.8)
Hispanic/Latinx	18 (8.8)	33 (10.3)	20 (8.7)
Multiethnic	5 (2.5)	8 (2.5)	7 (3.1)
Other	2 (1.0)	3 (0.9)	3 (1.2)
Highest level of education			
Some high school	0 (0)	2 (0.6)	0 (0)
Graduated high school	25 (12.2)	30 (9.3)	26 (11.4)
Some college	53 (25.9)	89 (27.7)	62 (27.2)
Associate degree	19 (9.3)	30 (9.3)	22 (9.6)
Bachelor's degree	77 (37.6)	131 (40.8)	86 (37.7)
Master's degree or higher	31 (15.1)	39 (12.1)	32 (14.0)
Marital status			
Single	72 (35.1)	119 (37.1)	79 (34.6)
Married/committed relationship	114 (55.6)	168 (52.3)	128 (56.1)
Divorced	15 (7.3)	27 (8.4)	16 (7.0)
Other	4 (2.0)	5 (1.6)	4 (1.8)
Prefer not to say	0 (0)	2 (0.6)	1 (0.4)
Insurance includes ANY mental health			
No	65 (31.7)	113 (35.2)	77 (33.8)
Yes	98 (47.8)	149 (46.4)	106 (46.5)
I do not know	42 (20.5)	59 (18.4)	45 (19.7)

PP $n = 205$, ITT Time 1 $n = 321$, ITT Time 2 $n = 228$.

3.2.1. H1

H1 predicted that participants in the HS group would be more likely to report greater success in initiating help-seeking for depression during the intervention period than those in the C group. Individuals who answered this item as "Not Applicable" were excluded from this analysis (ITT $n = 24$, PP $n = 23$). The independent samples t-test analyses supported this hypothesis regardless of the sampling measure ($t[202] = 2.509$, $p = 0.013$, CI: 0.103, 0.860): Participants in the HS group were more likely to report greater mean help-seeking success at T2 than those in the C group. See Figure 5A (ITT) and Figure 5B (PP) for means and standard error of scores.

3.2.2. H2

H2 predicted that participants in the HS group (i.e., those who completed the HS MCII) would be more likely to report greater *intentions to seek help* (as measured by the SIIS HS) than those in C *regardless of whether they actually sought help* during the intervention. Repeated measure mixed ANOVAs were used to test this hypothesis (SIIS HS scores T1, T2 x group). The results of this analysis supported this hypothesis regardless of testing the ITT or PP sample. The results found that completing the HS MCII positively influenced SIIS HS scores over time. Analyses revealed a statistically significant main effect for the SIIS HS, $F(1,226) = 7.979$, $p = 0.005$, partial $\eta^2 = 0.034$, indicating that individual's scores on the SIIS HS measure significantly

TABLE 5 Study 2 depression help-seeking demographics.

	PP *n* (%)	ITT Time 1 *n* (%)	ITT Time 2 *n* (%)
Have you ever believed you were depressed but did not seek help?[a]			
No	53 (25.9)	90 (28.0)	61 (26.6)
Yes	152 (74.1)	231 (72.0)	167 (72.9)
Have you ever sought help for depression from a loved one?[a]			
No	94 (45.9)	161 (50.2)	112 (48.9)
Yes	111 (54.1)	160 (49.8)	116 (50.7)
Have you ever sought help for depression from a professional?[a]			
No	110 (53.7)	179 (55.8)	125 (54.6)
Yes	95 (46.3)	142 (44.2)	103 (45.0)
Do you believe you currently have depression?[a]			
No	85 (41.5)	138 (43.0)	96 (41.9)
Yes	120 (58.5)	183 (57.0)	132 (57.6)
Are you currently diagnosed with depression?[a]			
No	170 (82.9)	265 (82.6)	192 (83.8)
Yes	14 (17.1)	56 (17.4)	36 (15.7)
Have you ever been diagnosed with depression?[a]			
No	128 (62.4)	210 (65.4)	146 (63.8)
Yes	77 (37.6)	111 (35.6)	82 (35.8)
Are you currently seeking professional help for depression?[a]			
No	205 (100)	321 (100)	228 (100)
Yes	0 (0)	0 (0)	0 (0)
Have you sought help for depression from a loved one in the past 2 weeks?[b]			
No	135 (65.9)	–	155 (68.1)
Yes	70 (34.1)	–	73 (31.9)
Have you sought help for depression from a professional in the past 2 weeks?[b]			
No	183 (89.3)	–	204 (89.5)
Yes	22 (10.7)		24 (10.5)
Do you believe you currently have depression?[b]			
	82 (40)	–	92 (40.6)
	123 (0)	–	136 (59.4)

[a]Time 1 question; [b]Time 2 depression question. ITT, Modified intention-to-treat; PP, Per Protocol. PP $n = 74$, ITT T1 $n = 139$, ITT T2 $n = 83$.

TABLE 6 Study 2 measure means, standard deviations, and internal consistencies (if applicable).

Measure	N items	PP T1		PP T2		ITT T1		ITT T2	
		M (*SD*)	*α*	*M* (*SD*)	*α*	*M* (*SD*)	*α*	*M* (*SD*)	*α*
BDI	21	24.44 (9.41)	0.89	23.93 (10.48)	0.92	24.32 (9.07)	0.88	24.01 (10.62)	0.92
SIIS HS	–	26.64 (7.89)	–	26.47 (7.36)	–	26.83 (8.39)	–	25.90 (8.04)	–
HS	7	28.80 (7.16)	0.92	28.80 (7.16)	0.94	27.84 (7.37)	0.92	27.84 (7.37)	0.92
C	7	24.22 (8.36)	0.93	24.22 (8.36)	0.93	24.26 (8.25)	0.93	24.26 (8.25)	0.93
MCII HS success	1	–	–	3.19 (1.39)	–	–	–	3.21 (1.38)	–

Measures Means, standard deviations, observed alpha (if applicable). T1, Time 1; T2, Time 2; BDI, Beck Depression II inventory scores; SIIS, Strength of Implementation Intentions Scale; SH, Seek Help; ITT, Intend to Treat; PP, Per Protocol; MCII, Mental Contrasting and Implementation Intentions Intervention.

varied from T1 to T2 for all participants. There was a main effect for group differences, (F(1,226) = 24.084, $p < 0.001$, $\eta^2 = 0.096$). Examining the pairwise comparisons, there was a significant difference between the HS and C groups indicating that the HS groups scored significantly higher (*M*diff = 4.803, $p < 0.001$, 95% CI: 2.875; 6.732). There was a significant group x SIIS HS interaction, F(1,226) = 11.468, $p = 0.001$,

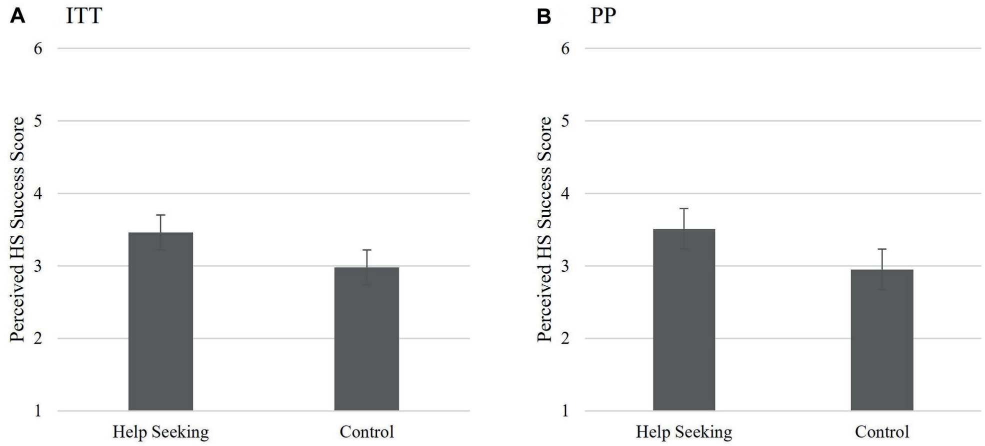


FIGURE 5
Study 2 Hypothesis 1: Success of help-seeking intervention. Those who received the intervention were more likely to report seeking help in the past 2 weeks than those who received information alone. This was supported for both the IIT **(A)** and **(B)** PP analyses Scores reported are the means for perceived success of help-seeking scores and the bars are the standard error.

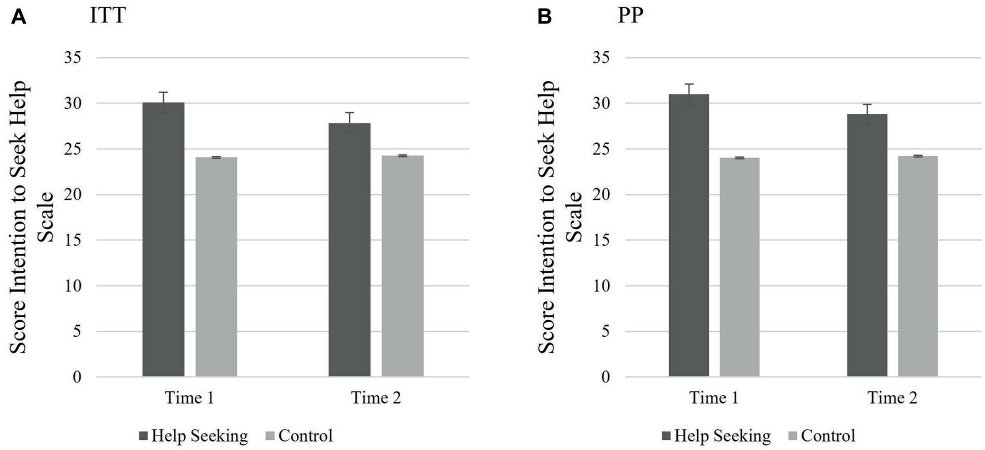


FIGURE 6
Study 2 Hypothesis 2: Changes to intentions to seek help. Illustrating the significant main effect for group and interaction between group and SIIS HS over time. Those who received HS intervention had greater intentions seek help across both time points compared to those who received information alone. This was supported for both the IIT **(A)** and PP **(B)** analyses. Scores reported are the means for the SIIS HS T1 and T2 and the bars are the standard error.

partial $\eta^2 = 0.048$. Although the HS group scored higher on the SIIS HS measure than the C group at both T1 and T2, the SIIS HS scores of the HS group decreased at T2. This decrease was not observed in the C group. See Figure 6A (ITT) and Figure 6B (PP) for means and standard error of scores.

3.2.3. Exploratory Analysis 1: effect of perceived utility and perceived resources on successful help-seeking

Fleig et al. (2017) proposed that the success of MCII could be dependent on perceived utility and perceived resources. A mixed ANOVA tested a 2 (utility T1, utility T2) × 2 (resources T1, resources T2) x 2 (success/nonsuccess of intervention) for participants in the HS group. Individuals who rated their MCII success as 1–3 were considered unsuccessful (ITT $n = 44$, PP $n = 35$) and those who rated their MCII success as 4–6, (ITT $n = 51$, PP $n = 42$) were considered successful. Those who indicated that the question was not applicable (ITT $n = 10$, PP $n = 9$) were treated as missing. The goal was to explore the relationship among perceived success of their HS MCII, utility of the interventions, and the perceived availability of the resources to carry out the interventions.

The results of the analyses did not vary significantly based on analysis methodology, therefore the results reported in text are the from the ITT sample. There were no significant differences between scores on perceived utility, ($F(1,93) = 3.054$, $p = 0.084$, partial $\eta^2 = 0.032$) or scores on perceived resources ($F(1,93) = 1.066$, $p = 0.305$, partial $\eta^2 = 0.011$) between T1 and T2. Additionally, there was no significant differences between those who rated themselves as successful and those who did not, $F(1,93) = 1.445$, $p = 0.232$, partial $\eta^2 = 0.015$. There were no interactions between perceived success and either perceived utility, $F(1,93) = 1.622$, $p = 0.206$, partial $\eta^2 = 0.017$, nor

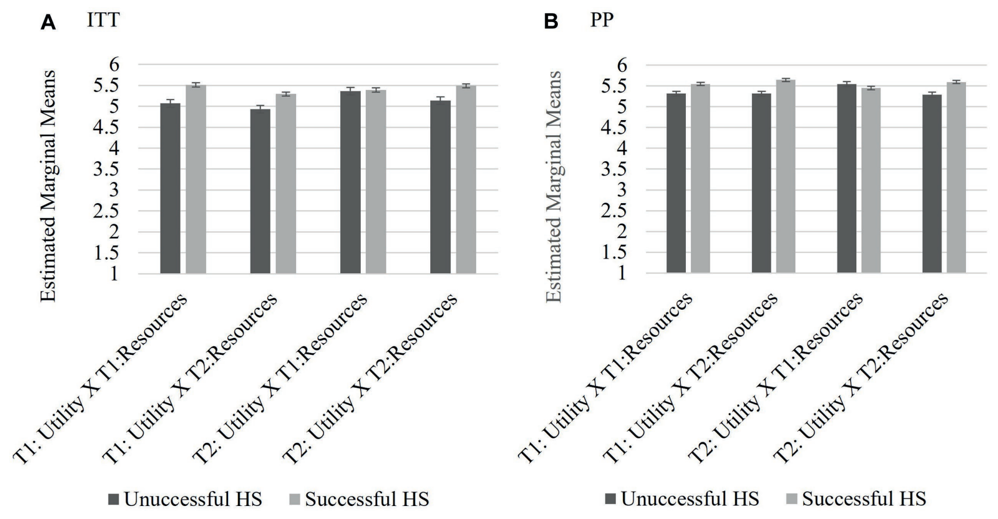


FIGURE 7
Study 2 Exploratory Analysis 1 3-way interaction: Utility x Resource x Success Ratings. For those who completed the HS MCII, there were no significant relationships between perceived utility, resources, nor interaction between utility and resources on perceived help-seeking success. Scores reported are the estimated marginal means for the interaction terms and the bars are the standard error for both the IIT **(A)** and PP **(B)** analyses.

perceived resources, $F(1,93) = 0.278$, $p = 0.599$, partial $\eta^2 = 0.003$, over time. There were no significant interactions between perceived utility and perceived resources ($F(1,93) = 0.560$, $p = 0.456$, partial $\eta^2 = 0.006$). Further, there was not a three-way interaction between perceived utility over time, perceived resources over time, and perceived success, $F(1,93) = 1.848$, $p = 0.177$, partial $\eta^2 = 0.019$. Please see Figure 7A (ITT) and Figure 7B (PP) for estimated marginal means for the three-way interaction term and the standard error bars.

3.2.4. Exploratory Analysis 2: effect of actual and perceived depression on successful help-seeking

Neither the perceived utility of goal setting nor having resources to carry out the MCII had a significant effect on perceived help-seeking success; thus, it seemed plausible that success was based on whether individuals believed help-seeking was needed. This alternative explanation was tested using both conditions by examining the potential role of perceived and actual depressive symptomatology in outcome ratings of perceived success for help-seeking.

The following analyses were conducted to determine if there were differences in frequencies in help-seeking success (dichotomous: 1 = success defined as reporting a score of 4 “slightly agree” to 6 “strongly agree” on the item relating to success of help-seeking; below 4 = 0) as a result of perceiving one was experiencing depression *regardless* of *actual depression* score (T2 “Are you currently depressed”; 1 = no, 0 = yes), actual depression score (BDI-II T2 dichotomous; 13 and below = 1 no to minimal; 14 and above = 0 at least mild symptomatology), and group (C; HS). The decision to dichotomize depression at a score indicating at least a mild level of depressive symptomatology (Hautzinger et al., 2009) was based on using the same cutoff in the screening measure. The rationale for dichotomizing and scoring all positive results (i.e., success, lower perceived depression, no-to-minimal BDI) in the same direction provided the clearest basis for understanding the participants’ perception of their success in achieving their goal of help-seeking and lowering depression. Individuals who indicated that the goal of increasing help-seeking for elevated depressive symptomatology was “not applicable” were excluded from these analyses ($n = 24$). For both sets of analyses, crosstab analyses were conducted ITT and PP as well as comparing the total sample to each of the conditions individually. If the results for the HS and C conditions indicated the same pattern, the total sample was utilized. MedCalc (RRID:SCR_015044) was used to calculate the odds ratios and confidence intervals. When both sample statistics were significant, only the ITT sample is listed in the text.

3.2.4.1. Exploratory Analysis 2a: the effect of actual depression at T2 on successful help-seeking

The first chi-square analysis explored whether the categorical level of depressive symptomatology as measured by the BDI-II at T2 (no or minimal vs. mild or greater), influenced help-seeking success (successful vs. non-successful). The ITT and PP results followed the same pattern of results and indicated significant differences based on condition assigned at randomization. At least one cell had an expected count below 5 for the C group.

For the HS group, the pattern of results varied significantly based on whether the participant had a score of at least a mild level of depressive symptomatology at T2 on the BDI-II and perceived help-seeking success, ITT $X^2(1) = 7.969$, $p = 0.005$ OR 11.825, CI: 1.460 to 95.790. Regardless of the sampling method, the odds ratio indicated the likelihood of reporting successful help-seeking was proportionally over 10 times greater for individuals whose depression scores fell to 13 or below on the BDI-II (no to minimal depression) than those continuing to score 14 and above indicating at least mild depression at T2.

For the C group, a Bayes factor correction was used as a correction for the low expected cell count. For those in the control condition, the categorical level of depressive symptomatology did not affect whether a participant reported seeking help, ITT $X^2(1) = 0.001$, $p = 0.990$, $BF = 2.775$, OR = 0.981, CI: 0.270 to 3.594. The Bayes factor indicated the probability of the data was only 0.36 times greater given the alternative hypothesis that BDI-II classification influences success

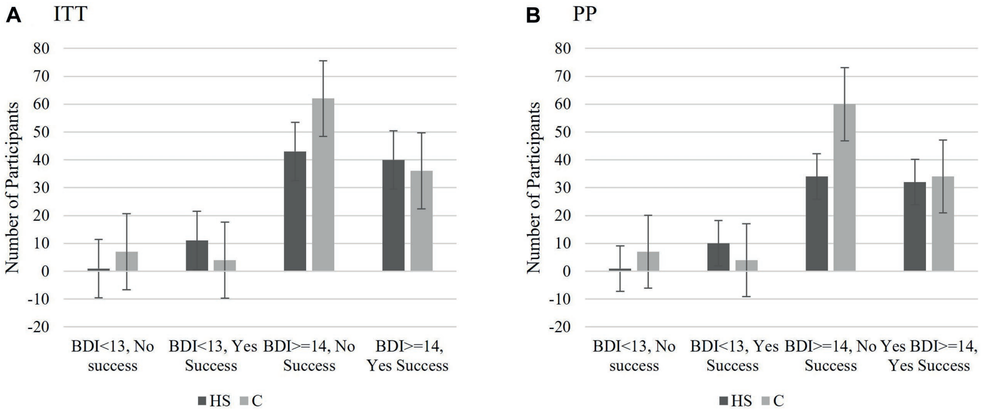


FIGURE 8
Study 2 Exploratory Analysis 2a: The effect of actual depression at T2 on successful help-seeking. Those who received the HS MCII and whose actual depression scores fell below the threshold for mild symptomatology at T2 (BDI-II=0–13), were proportionally the most likely to report seeking help. Depression scores did not significantly influence the success in the control group. This was true for both the IIT **(A)** and PP **(B)** analyses. BDI scores 0–13 indicates no to minimal symptomatology; BDI scores 14 or higher indicate mild or greater symptomatology. The scores represent number of participants who fell into each categorical classification and the bars are the standard errors.

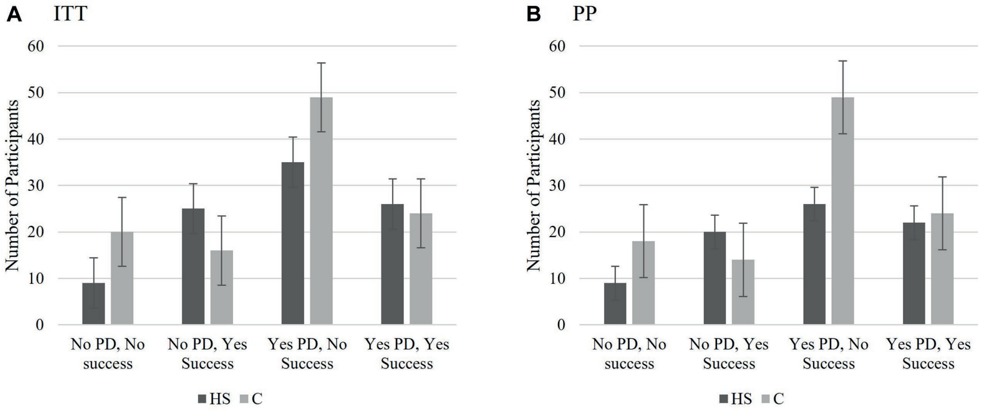


FIGURE 9
Study 2 Exploratory Analysis 2b: The effect of perceived depression at T2 on successful help-seeking at T2. **(A)** The significant odds ratio for the ITT sample indicated that the probability of reporting success at help-seeking was 3.741 times higher for individual in the HS group who did not perceive themselves as having elevated depressive symptomatology (PD) at T2 than those who did. **(B)** Despite a significant chi square, the confidence interval for the odds ratio was not significant for the PP sample for the HS group. There were no significant relationships between the PD at T2 and perceived success for the C group (ITT or PP). The scores represent number of participants who fell into each categorical classification and the bars are the standard errors.

rather than the null hypothesis for individuals in the control condition (ITT or PP). Please see Figure 8A (ITT) and Figure 8B (PP) for the number of participants in each classification and the standard error bars.

3.2.4.2. Exploratory Analysis 2b: the effect of perceived depression at T2 on successful help-seeking at T2

The ITT and PP samples followed the same pattern of results when examining the influence of perceived depression (regardless of actual depressive symptomatology scores), and help-seeking success as EA2a. There was a significant relationship between perceived depression status and help-seeking success for individuals in the HS group, *ITT* $X^2(1) = 8.387$, $p < 0.001$ OR 3.739, CI: 1.497 to 9.340; PP $X^2(1) = 3.902$, $p < 0.05$, OR = 2.626, CI 0.995 to 6.929. The significant odds ratio for the ITT sample indicated that the probability of reporting success at help-seeking was 3.741 times higher for individuals in the HS group who did not perceive themselves as depressed at T2 than those who did. Despite a significant chi-square, the odds ratio for the PP fell short of reaching significance.

However, for the control group, there was no significant difference between perceived depression status and help-seeking success regardless of sample, *ITT* $X^2(1) = 1.389$, $p = 0.292$, OR = 1.633, CI: 0.720 to 3.705. Please see Figure 9A (ITT) and Figure 9B (PP) for the number of participants in each classification and the standard error bars.

3.3. Study 2 discussion

The goal of Study 2 was to test the MCII intervention using a streamlined set of procedures given the results of Study 1. Specifically, the hypotheses were designed to test whether an online MCII intervention could increase *actual* help-seeking or the *intentions* to seek help for depression. Despite discontinuing survey collection early due to the COVID-19 pandemic, attrition in the ITT sample was reduced to 24% with the modifications made from Study 1 (e.g., limiting the scales to the ones necessary for testing the intervention, updating the power analysis calculation, oversampling).

3.3.1. Planned analyses

H1 indicated that the HS group was more likely to report seeking help at T2 than those in the control group. In addition, there was a significant interaction such that individuals in the HS group were more likely to report greater intentions to seek help for depression (as measured by SIIS HS) than those in the control condition over both time points; the individuals in the HS group had a dip in SIIS HS scores over time, which was not observed in the C group (H2). The literature is mixed regarding how long implementation intentions will last without a booster; Martin et al. (2011) found that a simple implementation intention intervention that involved similar implementation intention formation focusing on contraception reduced unplanned pregnancy and emergency contraception consultations among young women over the course of 2 years for the intervention group. Martin et al.'s results could be an anomaly; illustrative of differences in group demographics (Martin et al.'s participants were not depressed), or indicate that the young women in the original study were already highly motivated to begin contraception use.

3.3.2. Unplanned analyses

To examine whether participants in the HS group's perceived utility or perceived resources influenced success of the MCII intervention, a mixed ANOVA utilized the questions inspired by Fleig et al. (2017). Despite Study 2 boasting a larger sample size of individuals in the HS group ($n = 95$ ITT and $n = 75$ PP) the perceived utility of setting a goal and perceived resources to achieve the goal did not appear to significantly influence the success of the intervention over time. Considering the means for the perceived resources and perceived utilities at both T1 and T2 were all greater than five ("Agree") out of a possible score of seven for each item for both the PP and ITT analyses, it is arguable that a ceiling effect occurred. Although not mentioned in text, this was also examined using PROCCESS mediation analyses to examine whether utility and resources mediated the relationship between SIIS HS and HS MCII success as a continuous variable. Regardless of running with T1 or T2 SIIS (ITT and PP), none of the indirect effects in the analyses were significant. Further, when focusing only the regression analyses for utility, resources, and interactions of utility and resources on HS success, all analyses were insignificant.

The results of Exploratory Analyses 2a and 2b that individuals in the HS group who did not perceive themselves as depressed at T2 or whose BDI-II scores fell below the mild threshold for depression were proportionally *more likely* to report success at accomplishing their help-seeking goal were interesting findings. Conversely, perceived and actual depression scores were *not predictive of reported success* for individuals randomly assigned to the C group.

As a reminder, all participants entered the study with at least a score of 14 on the BDI-II indicating a mild level of depressive symptomatology (Hautzinger et al., 2009) at T1, 2 weeks prior to the success being measured at T2. For the individuals whose BDI-II scores at T2 indicated they no longer met a minimal threshold of depression (scores less than 14) and that they were successful at seeking help, it is impossible to establish order effects. It is just as possible that their episode of depression decreased to the point that seeking help was perceived as manageable *or* that because of seeking help, their scores on the BDI-II decreased. Using an ecological momentary assessment (EMA: Shiffman et al., 2008; Wenze and Miller, 2010) as a follow-up that can consistently measure depressive symptomatology and help-seeking actions similar to Kenny et al. (2016) could provide a clearer picture of causal relationships.

As far as perceived depression–one's *felt* depression *regardless of their objective level of depression*– the observed results are more curious. Of the 25 individuals in the HS group who perceived themselves as non-depressed and successful at seeking help at T2, 16 had scores on the BDI-II at T2 indicating at least a mild level of depressive symptomatology. Of those 16 individuals, half did not perceive themselves as depressed at T1. A larger sample should be used to further elucidate the relationships between actual and perceived depression and any potential influence it may have on help-seeking outcomes.

It may be that an individual's perception of "being depressed" could influence the effects of cognitive bias. Perhaps the individuals who did not perceive themselves as depressed despite meeting the mild criteria were comparing their current symptoms to others they have known who have experienced depression (e.g., social comparison theory of Festinger, 1954). An alternative is since the BDI-II inquires about depression symptoms during the previous 2 weeks (Hautzinger et al., 2009) and success did not inquire about *when* participants sought help during the intervention period, their symptoms of depression could have been subsiding for some time. Additionally, it is possible that the sample included individuals who have been diagnosed with other mental health disorders that include depressive symptoms (e.g., bipolar), potentially limiting their self-perception of "having depression" and providing a potential area for future studies.

3.3.3. Summary

With the modifications from Study 1, a larger sample was obtained and retained despite an early termination due to concerns about how the emerging COVID-19 pandemic could affect the results in March 2020. Over the past few years, the pandemic and the effects of quarantining have placed mental health awareness in the spotlight due to the observed rising depression rates (see Ettman et al., 2020; Hyland et al., 2020). The inability to access care face-to-face increased the necessity to reach out to others in different ways due to social distancing. Many individuals face new life stressors such as illness or death in their families or loss of income (Nelson et al., 2020). It is difficult to predict how the intervention results of this study may have varied if the data were collected 6 or 9 months later, but it is highly likely depression ratings would be greater (see Ettman et al., 2020; Hyland et al., 2020). The general discussion includes contributions to literature, limitations, and suggestions for future studies.

4. General discussion

Few would deny that depression is a serious condition, and emerging evidence indicates that the rates of depression (along with other mental health issues) has risen exponentially during the COVID-19 pandemic (Cao et al., 2020; Ettman et al., 2020; Vindegaard and Benros, 2020; Wang M. et al., 2021). Although Ettman et al. (2020) noted that increasing rates of mental health concerns is common during times of uncertainty and disruption such as after the attacks on the World Trade Towers and stock market crashes, the novel coronavirus may provide unique and lasting challenges for a wider segment of the population. Dubey et al.'s (2020) review of the literature found that factors such as forced quarantines, job insecurities, pervasive negative feelings (including guilt, inadequacy, or fear), scarcity of basic resources, on top of fear of the illness contributed to reduced feelings of wellbeing among individuals with little regard for age, gender, or occupation. In addition to confirming the role of the psychosocial factors listed above leading to increased depression, Wang M. et al. (2021) noted that for individuals with preexisting conditions, there is increased stress due to the inability to schedule or attend appointments. For these reasons, a simple online intervention that can encourage help-seeking for depression from a multitude of sources seems particularly timely.

Despite random assignment to conditions and technically reaching the target number of participants required by the power analysis after attrition and data cleaning in Study 1 (g*power estimate $n=72$, achieved PP $n=74$), the control group (C $n=40$) was still twice as large as the experimental (HS $n=17$) and comparison groups (E $n=17$). It was surprising that even with the small number of participants that completed the HS MCII, there was a statistically significant difference such that those in the HS group were more likely to report an increase in their implementation intentions to seek-help (as measured by the SIIS HS) than those in the E or C groups regardless of their actual help-seeking behaviors for those in the PP sample. Study 1 established that although modifications were necessary, the online HS MCII intervention for use with MTurk populations was feasible.

With modifications to the length of survey and over-sampling, attrition was reduced from ~40% in Study 1 to ~25% in Study 2. The significant intervention results in H1 and H2a appeared more promising in Study 2. By adding the depression demographic measures to T2, it was possible to perform the supplementary analyses. However, interpreting the results of those analyses requires caution since the individuals who were proportionally the most likely to be successful were in the HS group and did not *perceive* themselves as having elevated depressive symptomatology or their depressive symptomatology decreased from T1 to T2. An alternative hypothesis proposed by Nisbett and Wilson (1977) is that individuals may not have cognitive access or awareness of their depression status and therefore, any reports of their depression status may be suspect (see also Johansson et al., 2005; Petitmengin et al., 2013). Without further exploration with a larger sample and using a technique such as EMA (Shiffman et al., 2008; Wenze and Miller, 2010), it would be impossible to establish a timeline of changes in depression scores (or changes in negative bias) and individual help-seeking actions (see Kenny et al., 2016).

4.1. Limitations and strengths

Although samples collected via MTurk's TurkPrime, which is based on self-selection, offers more diversity than an average college class, as noted by Casler et al. (2013), samples consistently more likely to be white (see also McCredie and Morey, 2019) and suffer from high attrition rates (Zhou and Fishbach, 2016). However, it is also possible that the demographics were homogeneous because of limiting the sample to US residents who were proficient in English. McCredie and Morey (2019) found that MTurk participants may vary from community samples by being more socially isolated, having a more limited social support network, and having higher depression scores (p. 764). Although their findings suggest limitations for some topics, these characteristics suggest that MTurk samples may provide a fertile ground to test the present intervention. Although the initial MTurk HIIT did not advertise that this study was specifically related to help-seeking for depression, no deception was used in the consent form indicating that there was self-selection into this study since all individuals could opt out at any point. This was also reflected by rates of self-reporting at least mild levels of depressive symptomatology recruited for both Studies (Study 1: 32.4%; Study 2: 33.8%) that are a little higher than recently reported statistics on pre-pandemic depressive symptomatology (pre-pandedemic: 24.7%, March 2020–June 2021: 36%; See Ettman et al., 2020, 2023). It should also be noted that the BDI-II is a depression screening measure, but it does not take the place of a clinical diagnosis of depression. Relevance of the intervention to the MTurk population could also help explain why the current studies collected slightly larger sample sizes than previous MCII mental health studies (e.g., $N=47$, Fritzsche et al., 2016; $N=36$, Sailer et al., 2015).

The rationale for using ITT analyses is to reduce bias when conducting randomized control trials. As was noted, rather than using a traditional ITT method where all participants who were randomized are analyzed (see Day, 2008; Gupta, 2011), this study preestablished the necessity of passing the quantitative attention checks. This modification seemed pertinent to add as a layer of protection against fraudulent data often used with MTurk samples (e.g., Kennedy et al., 2020). A key challenge for the current studies is that with only two time points and significant attrition rates, making plausible assumptions to impute the missing data (e.g., using the carry forward method, multiple imputation) would also create undo bias. White et al. (2011) offers several suggestions to combat these issues including vigilant follow ups and reducing attrition using study design; both methods these studies employed. To assess potential bias in attrition rates in the current studies, the modified ITT samples for each study was compared by condition (i.e., Study 1: SH, E, C; Study 2: SH, C) and time point (i.e., T1, T2) and no statistically significant differences were found. To be clear, there are many ways to approach ITT analyses, each with their own caveats; it is possible using another method might have returned different results.

There are many methods that can be used to form and reinforce implementation intentions, ranging from simply reading a preformed implementation intention (e.g., Gawrilow and Gollwitzer, 2007; Parks-Stamm et al., 2010) to the more in-depth MCII interventions (e.g., Sailer et al., 2015; Fritzsche et al., 2016) including the specific WOOP strategy (e.g., Mutter et al., 2020; Monin et al., 2021). Other studies have successfully used tools such as volitional help sheets whereby the participants were aided in forming implementation intentions for multiple situations related to the topic (e.g., Armitage, 2015; Armitage et al., 2016). Given the many ways implementation intentions can be formed and reinforced, one limitation of this set of studies is the sole utilization of the online MCII. The decision to do so was based on the success of MCII with individuals with elevated depressive symptomatology (e.g., Sailer et al., 2015; Fritzsche et al., 2016) but future studies may consider exploring if this approach is best by comparing it with other techniques for establishing implementation intentions. For example, a volitional help sheet that outlines multiple help-seeking options may be just as–or more–useful for this population.

All measures used in this study and the measure of success was dependent on self-report; a possible limitation. This procedure is not unusual in the field of MCII (Webb and Sheeran, 2006) or depression research (Keeler and Siegel, 2016), but should be noted. Although it is certainly more resource efficient (i.e., time, money), it is possible that individuals will not be honest or will forget about their help-seeking practices. However, while self-report is less than perfect, the added opportunity to reiterate the MCII in the form of a quantitative scale, directions to keep a copy of the help-seeking information, and the personalized implementation on top of the relatively short time period may reduce memory errors across groups. In general, few individuals were removed due to completely forgetting their implementation intention. It should also be noted that apart from asking two questions in Study 2 regarding whether participants sought help from interpersonal or professional sources, we did not require participants to qualify what help-seeking behaviors they initiated in the past 2 weeks; only if they were successful at initiating their personalized help-seeking plan.

Studying the HS MCII intervention effects add significantly to the literature in several ways. As mentioned, although the literature using implementation intentions and mental health has grown substantially since Gollwitzer and Sheeran's (2006) meta-analysis, interventions directed specifically at individuals with depression are quite rare (e.g., Fritzsche et al., 2016). Quite often, individuals with depression may be included in studies (e.g., Sheeran et al., 2007; Pomp et al., 2013; Armitage et al., 2016) or depression may be considered as a variable (e.g., Sailer et al., 2015; Mutter et al., 2020) but it is not the exclusive focus. Given the conflicting results of the previous studies that measured depression explicitly (i.e., successful: Fritzsche et al., 2016; unsuccessful: Pomp et al., 2013), this study provided much needed support for the viability of using implementation intentions among individuals with elevated depressive symptomatology. This study may serve as a step in supporting the premise of how the process of completing an MCII can limit the effects of negative bias to encourage achieving a goal beyond receiving information alone. This appeared to happen with those in the HS MCII condition.

Although the results should be replicated to fully consider the differences in reporting timeframes of the BDI-II, perceived depression, and goal achievement, the most significant addition to the literature is the potential of adding MCII as another tool to initiate help-seeking for depression. As mentioned, the current mental health implementation intention literature is nearly devoid of help-seeking interventions. The only implementation intention study located focused specifically on helping individuals *follow through* with a previously scheduled appointment with a mental health professional (i.e., Sheeran et al., 2007). A recent literature search in mid 2022 failed to find any studies that expressly examined MCII for depression using MTurk. Though an MCII intervention was successful at initiating help-seeking for individuals to establish behaviors to help overcome chronic back pain (Christiansen et al., 2010), the use for the initiation of mental health support has not been previously explored. Given the number of interventions designed to encourage help-seeking that have had iatrogenic effects (e.g., Lienemann et al., 2013; Keeler and Siegel, 2016), it is important to find innovative and reliable ways to encourage formal and informal help-seeking in relatively simple to disseminate that is easy to tailor to individuals. Such research could make a practical and important impact considering escalating depression rates due to the COVID-19 pandemic (Cao et al., 2020; Ettman et al., 2020; Vindegaard and Benros, 2020; Wang M. et al., 2021).

4.2. Future directions

Above all, we suggest replication and expansion studies that address the limitations noted with larger, more diverse samples and via different recruitment modalities (i.e., cloud research, community, health clinic). With further research, it will be possible to delineate exactly how this method can be applied in the future to examine possible dissemination to a wider audience to encourage various help-seeking initiation behaviors like other mental health interventions including through mass media (e.g., Ort and Fahr, 2022), clinics (e.g., Monin et al., 2021), or workplace initiatives (e.g., Gollwitzer et al., 2018). Thus, more research is needed to replicate findings and expand reach.

As a first step, future studies should seek to both generalize to symptomatology beyond depression as well as seek to explore combinations of symptom severity. It is currently unknown if a mental health help-seeking MCII intervention would be effective for individuals who may not be actively experiencing negative bias (e.g., anxiety, mania).

Considering the results regarding the differences in BDI-II scores in the exploratory analyses, future studies may want to explore the nuance between the severity of disorders and the type of implementation intention interventions that work best at each level for encouraging help-seeking. Currently, assessing the level of severity of mental health disorders has been largely unexplored with the exception of Parks-Stamm et al. (2010), who examined the effectiveness of the type of implementation intention intervention most useful for shielding individuals with varying levels of test anxiety from unwanted distraction. Their results indicated it is vital to explore severity of conditions because what is useful at a mild level of impairment may not be effective at a more severe level. In the case of depression, a simpler implementation intention ("If I am feeling depressed, I will call my loved one") may be sufficient to increase help-seeking for depression for someone with mild symptoms, whereas the more complex MCII interventions (i.e., elaborating on the positive and negative aspects of calling a loved one for help) with multiple follow ups or in-person training may be needed for moderate to severe levels of depression. This knowledge would be useful in that resources (e.g., time, money for training) can be more efficiently directed to the populations that need it the most.

Galea et al.'s (2020) recent call for innovative ways to increase training for non-traditional mental health first aid suggests an alternative population interest for future MCII aimed to care for individuals with elevated depressive symptomatology. HS MCII interventions could potentially be utilized for the family member to recognize when they should intervene with their loved one with depression and have pre-established responses to best offer support to their loved ones (e.g., "If my loved one talks about suicide, then I will call the crisis intervention hotline saved on my phone"). Due to the high recurrence rate of depression (Judd et al., 2000), it may be interesting and useful to design an MCII intervention for the loved ones of individuals with mental health symptoms as part of an aftercare plan. For example, the loved ones could form specific implementation intentions for what, when, and how they will provide their loved one with help if they see the symptoms of depression reoccur ("If I notice my loved one stops going to their aftercare treatment, then I will offer to drive them to their next appointment"). Although there is limited evidence in this area, a recent study found that focusing on caregivers can benefit both the caregiver and their loved one's mental health (Monin et al., 2021).

Although the current intervention seeks to explore if MCII techniques can help initiate help-seeking for depression over a 2 week time period, it would be interesting to test if combining MCII techniques with aspects of a complementary model such as Siegel et al. (2010) IIFF model could bolster intervention effectiveness. The IIFF Model, designed to increase organ donation registration, advocates for increasing information, favorable activation of the desired behavior, focused engagement, and an immediate and complete opportunity to engage in the behavior. A joint MCII and IIFF help-seeking for depression intervention would include providing information that is specifically favorable to the multiple avenues of help-seeking, which is already standard in many help-seeking campaigns. Due to mental contrasting's proposed tempering of negative bias due to the more balanced approach to deciding to seek help and acknowledging the negative reality, a favorable view of help-seeking may not have the same boomerang effect previous studies have found (e.g., Keeler and Siegel, 2016). Mental contrasting and implementation intentions are designed to provide focused engagement with an issue by having the individual actively contemplate not only their positive and negative realities of initiating help-seeking but also setting a plan.

Where IIFF could add to MCII is by answering the question of whether providing an *immediate* opportunity for help-seeking would be efficacious. This could be accomplished by using an online study and providing participants with depression the opportunity to go directly to a link for a national crisis center information page, directly link to the suicide lifeline chat service, connect to a professional, or to open a window to send an email to a loved one for help.

4.3. Conclusion

At a time when depression rates are increasing because of the lingering COVID-19 pandemic (Cao et al., 2020; Ettman et al., 2020; Vindegaard and Benros, 2020; Wang M. et al., 2021), it is vital to develop remote, affordable, scalable, and effective interventions to encourage help-seeking. Together, this set of studies offers support that a brief online MCII intervention to increase help-seeking initiation and intentions to seek help is feasible and offers preliminary evidence of success. However, whether actual help-seeking initiation success is based *solely* on the intervention requires further investigation. Future studies should address the limitations of this study and consider using EMA measurements to establish temporal precedence regarding the finding that those who were more likely to report successful help-seeking reported decreased BDI-II scores at T2. Examining larger samples of participants with varying severities and differing mental health symptoms could also provide insight into the effectiveness of MCII for encouraging mental health help-seeking among individuals prone to experiencing cognitive errors who may not be experiencing negative bias (e.g., bipolar disorder or anxiety). Clinicians may find this method successful for encouraging continued attendance to treatment sessions as well as targeting loved ones to recognize warning signs and plan a strategy for intervening with a loved one with depression.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Claremont Graduate University. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

AK, LN, and WC contributed to the conception and design of this manuscript. LN and AK modified the key measurement. AK created the database, performed the statistical analyses, and wrote the first draft of the manuscript. WC provided the statistical support. LN completed tables and figures. All authors contributed to the manuscript revision, read, and approved the submitted version.

Funding

This publication was supported by the Health Resources and Services Administration (HRSA) of the U.S. Department of Health and Human Services (HHS) as part of an award totaling $590,270.00 with 14% financed with non-governmental sources. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by HRSA, HHS, or the U.S. Government. For more information, please visit HRSA.gov.

Acknowledgments

These hypotheses were initially included in the AK's dissertation (Keeler, 2021) and is uploaded in the Proquest database as cited in the references. The intervention effects from Study 2 were presented in poster format at both the Sept 2022 National Network of Depression Center's Annual Conference in Ann Arbor Michigan as well as at the November 2022 North American Primary Care Research annual conference, Phoenix, AZ (Keeler and Nydegger, 2023).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1145969/full#supplementary-material

References

Armitage, C. J. (2015). Randomized test of a brief psychological intervention to reduce and prevent emotional eating in a community sample. *J. Public Health* 37, 438–444. doi: 10.1093/pubmed/fdv054

Armitage, C. J., Rahim, W. A., Rowe, R., and O'Connor, R. C. (2016). An exploratory randomised trial of a simple, brief psychological intervention to reduce subsequent suicidal ideation and behaviour in patients admitted to hospital for self-harm. *Br. J. Psychiatry* 208, 470–476. doi: 10.1192/bjp.bp.114.162495

Bayer, U. C., Achtziger, A., Gollwitzer, P. M., and Moskowitz, G. B. (2009). Responding to subliminal cues: do if-then plans facilitate action preparation and initiation without conscious intent? *Soc. Cogn.* 27, 183–201. doi: 10.1521/soco.2009.27.2.183

Beck, A. T. (1964). Thinking and depression. Ii. Theory and therapy. *Arch. Gen. Psychiatry* 10, 561–571. doi: 10.1001/archpsyc.1964.01720240015003

Beck, A. T. (1967) *Depression: clinical, experimental, and theoretical aspects*. New York: Hoeber Medical Division, Harper & Row.

Beck, A. T. (1987). Cognitive models of depression. *J. Cogn. Psychother.* 1, 5–37.

Beck, Aaron T. (2009). *Depression: causes and treatment, 2nd*. Philadelphia, PA: University of Pennsylvania Press.

Beck, A. T., and Alford, B. A. (2009). *Depression: causes and treatment*. Philadelphia, PA: University of Pennsylvania Press.

Beck, A. T., and Bredemeier, K. (2016). A unified model of depression. *Clin. Psychol. Sci.* 4, 596–619. doi: 10.1177/2167702616628523

Beck, A. T., Steer, R. A., and Brown, G. K. (1996). *BDI-II, Beck depression inventory: manual. 2nd*. San Antonio, TX: Psychological Corporation.

Bowie, C. R., Milanovic, M., Tran, T., and Cassidy, S. (2017). Disengagement from tasks as a function of cognitive load and depressive symptom severity. *Cogn. Neuropsychiatry* 22, 83–94. doi: 10.1080/13546805.2016.1267617

Cao, W., Fang, Z., Hou, G., Han, M., Xu, X., Dong, J., et al. (2020). The psychological impact of the COVID-19 epidemic on college students in China. *Psychiatry Res.* 287:112934. doi: 10.1016/j.psychres.2020.112934

Carbonell, A., Navarro-Pérez, J. J., and Mestre, M. V. (2020). Challenges and barriers in mental healthcare systems and their impact on the family: a systematic integrative review. *Health Soc. Care Community* 28, 1366–1379. doi: 10.1111/hsc.12968

Casler, K., Bickel, L., and Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Comput. Hum. Behav.* 29, 2156–2160. doi: 10.1016/j.chb.2013.05.009

Christensen, H., Leach, L. S., Barney, L., Mackinnon, A. J., and Griffiths, K. M. (2006). The effect of web based depression interventions on self reported help seeking: randomised controlled trial [ISRCTN77824516]. *BMC Psychiatry* 6, 1–11. doi: 10.1186/1471-244X-6-13

Christiansen, S., Oettingen, G., Dahme, B., and Klinger, R. (2010). A short goal-pursuit intervention to improve physical capacity: a randomized clinical trial in chronic back pain patients. *Pain* 149, 444–452. doi: 10.1016/j.pain.2009.12.015

Clark, D. A., and Beck, A. T. (2010). Cognitive theory and therapy of anxiety and depression: convergence with neurobiological findings. *Trends Cogn. Sci.* 14, 418–424. doi: 10.1016/j.tics.2010.06.007

Corrigan, P. (2004). How stigma interferes with mental health care. *Am. Psychol.* 59, 614–625. doi: 10.1037/0003-066X.59.7.614

Crano, W. D., Brewer, M. B., and Lac, A. (2014) *Principles and methods of social research*. New York: Routledge.

Day, S. (2008). *Analysis issues, ITT, post-hoc, and subgroups*. Baltimore: Johns Hopkins University Bloomberg School of Public Health. Available at: http://ocw.jhsph.edu/courses/BiostatMedicalProductRegulation/biomed_lec7_day.pdf (Accessed February 17, 2019).

Dubey, S., Biswas, P., Ghosh, R., Chatterjee, S., Dubey, M. J., Chatterjee, S., et al. (2020). Psychosocial impact of COVID-19. *Diabetes Metab. Syndr. Clin. Res. Rev.* 14, 779–788. doi: 10.1016/j.dsx.2020.05.035

Duckworth, A. L., Kirby, T., Gollwitzer, A., and Oettingen, G. (2013). From fantasy to action: mental contrasting with implementation intentions (MCII) improves academic performance in children. *Soc. Psychol. Personal. Sci.* 4, 745–753. doi: 10.1177/1948550613476307

Ettman, C. K., Abdalla, S. M., Cohen, G. H., Sampson, L., Vivier, P. M., and Galea, S. (2020). Prevalence of depression symptoms in US adults before and during the COVID-19 pandemic. *JAMA Netw. Open* 3:e2019686. doi: 10.1001/jamanetworkopen.2020.19686

Ettman, C. K., Fan, A. Y., Subramanian, M., Adam, G. P., Badillo Goicoechea, E., Abdalla, S. M., et al. (2023). 'Prevalence of depressive symptoms in U.S. adults during the COVID-19 pandemic: a systematic review', SSM - Population. *Health* 21:101348. doi: 10.1016/j.ssmph.2023.101348

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using G* power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/BRM.41.4.1149

Festinger, L. (1954). A theory of social comparison processes. *Hum. Relat.* 7, 117–140. doi: 10.1177/001872675400700202

Field, A. P. (2018). *Discovering statistics using IBM SPSS statistics. 5th*. London: SAGE Publications.

Fleig, L., Gardner, B., Keller, J., Lippke, S., Pomp, S., and Wiedemann, A. U. (2017). What contributes to action plan enactment? Examining characteristics of physical activity plans. *Br. J. Health Psychol.* 22, 940–957. doi: 10.1111/bjhp.12263

Fritzsche, A., Schlier, B., Oettingen, G., and Lincoln, T. M. (2016). Mental contrasting with implementation intentions increases goal-attainment in individuals with mild to moderate depression. *Cogn. Ther. Res.* 40, 557–564. doi: 10.1007/s10608-015-9749-6

Galea, S., Merchant, R. M., and Lurie, N. (2020). The mental health consequences of COVID-19 and physical distancing: the need for prevention and early intervention. *JAMA Intern. Med.* 180, 817–818. doi: 10.1001/jamainternmed.2020.1562

Gawrilow, C., and Gollwitzer, P. M. (2007). Implementation intentions facilitate response inhibition in children with ADHD. *Cogn. Ther. Res.* 32, 261–280. doi: 10.1007/s10608-007-9150-1

Gawrilow, C., Morgenroth, K., Schultz, R., Oettingen, G., and Gollwitzer, P. M. (2012). Mental contrasting with implementation intentions enhances self-regulation of goal pursuit in schoolchildren at risk for ADHD. *Motiv. Emot.* 37, 134–145. doi: 10.1007/s11031-012-9288-3

Gollwitzer, P. M. (1990). *'Action phases and mind-sets', handbook of motivation and cognition: Foundations of social behavior*, Vol. *2*. New York, NY: The Guilford Press, pp. 53–92.

Gollwitzer, P. M. (1993). Goal achievement: the role of intentions. *Eur. Rev. Soc. Psychol.* 4, 141–185. doi: 10.1080/14792779343000059

Gollwitzer, P. M., and Bargh, J. A. (1996) *The psychology of action: linking cognition and motivation to behavior*. New York: Guilford Press.

Gollwitzer, P. M., Bayer, U. C., and McCulloch, K. C. (2005) *'The control of the unwanted', the new unconscious. Oxford series in social cognition and social neuroscience*. New York, NY: Oxford University Press, pp. 485–515.

Gollwitzer, P. M., and Brandstätter, V. (1997). Implementation intentions and effective goal pursuit. *J. Pers. Soc. Psychol.* 73, 186–199. doi: 10.1037/0022-3514.73.1.186

Gollwitzer, P. M., Mayer, D., Frick, C., and Oettingen, G. (2018). Promoting the self-regulation of stress in health care providers: an internet-based intervention. *Front. Psychol.* 9:838. doi: 10.3389/fpsyg.2018.00838

Gollwitzer, P. M., and Sheeran, P. (2006). Implementation intentions and goal achievement: a meta-analysis of effects and processes. *Adv. Exp. Soc. Psychol.* 38, 69–119. doi: 10.1016/S0065-2601(06)38002-1

Gupta, S. K. (2011). Intention-to-treat concept: a review. *Perspect. Clin. Res.* 2, 109–112. doi: 10.4103/2229-3485.83221

Hauser, D., Paolacci, G., and Chandler, J. J. (2018). “Common concerns with MTurk as a participant pool: evidence and solutions” in *Handbook of research methods in consumer psychology*. eds. F. R. Kardes, P. M. Herr and N. Schwarz (New York: Routledge)

Hautzinger, M., Beck, A. T., Keller, F., and Kühner, C. (2009). *Beck Depressions-Inventar: BDI II; Manual*. Revision, 2. Aufl. Edn. Frankfurt am Main: Pearson Assessment.

Henderson, C., Evans-Lacko, S., and Thornicroft, G. (2013). Mental illness stigma, help seeking, and public health programs. *Am. J. Public Health* 103, 777–780. doi: 10.2105/AJPH.2012.301056

Hyland, P., Shevlin, M., McBride, O., Murphy, J., Karatzias, T., Bentall, R. P., et al. (2020). Anxiety and depression in the Republic of Ireland during the COVID-19 pandemic. *Acta Psychiatr. Scand.* 142, 249–256. doi: 10.1111/acps.13219

Ingram, R. E., Steidtmann, D. K., and Bistricky, S. L. (2008). “Chapter 7 - information processing: attention and memory” in *Risk factors in depression*. eds. K. S. Dobson and D. J. A. Dozois (San Diego: Elsevier), 145–169.

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., et al. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *Lancet* 392, 1789–1858. doi: 10.1016/S0140-6736(18)32279-7

Johansson, P., Hall, L., Sikström, S., and Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science* 310, 116–119. doi: 10.1126/science.1111709

Judd, L. L., Akiskal, H. S., Zeller, P. J., Paulus, M., Leon, A. C., Maser, J. D., et al. (2000). Psychosocial disability during the long-term course of unipolar major depressive disorder. *Arch. Gen. Psychiatry* 57, 375–380. doi: 10.1001/archpsyc.57.4.375

Keeler, A. R. (2021). Help-seeking for depression: can a mental-contrasting and implementation-intentions intervention overcome the curse of the boomerang? Ann Arbor: Claremont graduate university. Available at: https://ccl.on.worldcat.org/atoztitles/link?sid=ProQ:&issn=&volume=&issue=&title=Help-Seeking+for+Depression%3A+Can+a+Mental-Contrasting+and+Implementation-Intentions+Intervention+Overcome+the+Curse+of+the+Boomerang%3F&spage=&date=2021-01-01&atitle=Help-Seeking+for+Depression%3A+Can+a+Mental-Contrasting+and+Implementation-Intentions+Intervention+Overcome+the+Curse+of+the+Boomerang%3F&au=Keeler%2C+Amanda&id=doi.

Keeler, A. R., and Nydegger, L. (2023). “An RCT to Assess Whether an Online Behavioral Intervention Increases Help-Seeking Behavior Initiation for Depression” in *The Annals of Family Medicine*. 21:4344. Available at: https://go.gale.com/ps/i.do?id=GALE%7CA737170839&sid=sitemap&v=2.1&it=r&p=EAIM&sw=w&userGroupName=anon%7E2559ab5&aty=open+web+entry

Keeler, A. R., and Siegel, J. T. (2016). Depression, help-seeking perceptions, and perceived family functioning among Spanish-dominant Hispanics and non-Hispanic whites. *J. Affect. Disord.* 202, 236–246. doi: 10.1016/j.jad.2016.05.017

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., and Winter, N. J. G. (2020). The shape of and solutions to the MTurk quality crisis. *Polit. Sci. Res. Methods* 8, 614–629. doi: 10.1017/psrm.2020.6

Kenny, R., Dooley, B., and Fitzgerald, A. (2016). Ecological momentary assessment of adolescent problems, coping efficacy, and mood states using a mobile phone app: an exploratory study. *JMIR Mental Health* 3:e6361. doi: 10.2196/mental.6361

Klimes-Dougan, B., Klingbeil, D. A., and Meller, S. J. (2013). The impact of universal suicide-prevention programs on the help-seeking attitudes and behaviors of youths. *Crisis* 34, 82–97. doi: 10.1027/0227-5910/a000178

Klimes-Dougan, B., and Lee, C.-Y. S. (2010). Suicide prevention public service announcements. *Crisis* 31, 247–254. doi: 10.1027/0227-5910/a000032

Lienemann, B. A., Siegel, J. T., and Crano, W. D. (2013). Persuading people with depression to seek help: respect the boomerang. *Health Commun.* 28, 718–728. doi: 10.1080/10410236.2012.712091

Linde, K., Sigterman, K., Kriston, L., Rücker, G., Jamil, S., Meissner, K., et al. (2015). Effectiveness of psychological treatments for depressive disorders in primary care: systematic review and meta-analysis. *Ann. Fam. Med.* 13, 56–68. doi: 10.1370/afm.1719

Martin, J., Sheeran, P., Slade, P., Wright, A., and Dibble, T. (2011). Durable effects of implementation intentions: reduced rates of confirmed pregnancy at 2 years. *Health Psychol.* 30, 368–373. doi: 10.1037/a0022739

McCredie, M. N., and Morey, L. C. (2019). Who are the Turkers? A characterization of MTurk workers using the personality assessment inventory. *Assessment* 26, 759–766. doi: 10.1177/1073191118760709

Mekonen, T., Ford, S., Chan, G. C. K., Hides, L., Connor, J. P., and Leung, J. (2022). What is the short-term remission rate for people with untreated depression? A systematic review and meta-analysis. *J. Affect. Disord.* 296, 17–25. doi: 10.1016/j.jad.2021.09.046

Monin, J. K., Oettingen, G., Laws, H., David, D., DeMatteo, L., and Marottoli, R. (2021). A controlled pilot study of the wish outcome obstacle plan strategy for spouses of persons with early-stage dementia. *J. Gerontol. Ser. B* 77, 513–524. doi: 10.1093/geronb/gbab115

Mutter, E. R., Oettingen, G., and Gollwitzer, P. M. (2020). An online randomised controlled trial of mental contrasting with implementation intentions as a smoking behaviour change intervention. *Psychol. Health* 35, 318–345. doi: 10.1080/08870446.2019.1634200

Nelson, L. M., Simard, J. F., Oluyomi, A., Nava, V., Rosas, L. G., Bondy, M., et al. (2020). US public concerns about the COVID-19 pandemic from results of a survey given via social media. *JAMA Intern. Med.* 180, 1020–1022. doi: 10.1001/jamainternmed.2020.1369

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychol. Rev.* 84, 231–259. doi: 10.1037/0033-295X.84.3.231

Nydegger, L. A., Ames, S. L., and Stacy, A. W. (2017). Predictive utility and measurement properties of the strength of implementation intentions scale (SIIS) for condom use. *Soc. Sci. Med.* 185, 102–109. doi: 10.1016/j.socscimed.2017.05.035

Nydegger, L. A., Keeler, A. R., Hood, C., Siegel, J. T., and Stacy, A. W. (2013). Effects of a one-hour intervention on condom implementation intentions among drug users in Southern California. *AIDS Care* 25, 1586–1591. doi: 10.1080/09540121.2013.793271

Oettingen, G., and Gollwitzer, P. M. (2010). “Strategies of setting and implementing goals: mental contrasting and implementation intentions” in *Social psychological foundations of clinical psychology*. eds. J. E. Maddux and J. P. Tangney (New York: Guilford Press), 114–135.

Oettingen, G., and Gollwitzer, P. M. (2018) 'Health behavior change by self-regulation of goal pursuit: mental contrasting with implementation intentions'. In RidderD. de, M. Adriaanse and K. Fujita (Eds.), *The Routledge international handbook of self-control in health and well-being* (pp. 418–430). New York, NY: Routledge.

Oettingen, G., and Reininger, K. M. (2016). The power of prospection: mental contrasting and behavior change. *Soc. Personal. Psychol. Compass* 10, 591–604. doi: 10.1111/spc3.12271

Ort, A., and Fahr, A. (2022). Mental contrasting with implementation intentions as a technique for media-mediated persuasive health communication. *Health Psychol. Rev.* 16, 602–621. doi: 10.1080/17437199.2021.1988866

Parikh, S. V., Taubman, D. S., Antoun, C., Cranford, J., Foster, C. E., Grambeau, M., et al. (2018). The Michigan peer-to-peer depression awareness program: school-based prevention to address depression among teens. *Psychiatr. Serv.* 69, 487–491. doi: 10.1176/appi.ps.201700101

Parks-Stamm, E. J., Gollwitzer, P. M., and Oettingen, G. (2010). Implementation intentions and test anxiety: shielding academic performance from distraction. *Learn. Individ. Differ.* 20, 30–33. doi: 10.1016/j.lindif.2009.09.001

Petitmengin, C., Remillieux, A., Cahour, B., and Carter-Thomas, S. (2013). A gap in Nisbett and Wilson's findings? A first-person access to our cognitive processes. *Conscious. Cogn.* 22, 654–669. doi: 10.1016/j.concog.2013.02.004

Pomp, S., Fleig, L., Schwarzer, R., and Lippke, S. (2013). Effects of a self-regulation intervention on exercise are moderated by depressive symptoms: a quasi-experimental study. *Int. J. Clin. Health Psychol.* 13, 1–8. doi: 10.1016/S1697-2600(13)70001-2

Rüsch, N., Evans-Lacko, S. E., Henderson, C., Flach, C., and Thornicroft, G. (2011). Knowledge and attitudes as predictors of intentions to seek help for and disclose a mental illness. *Psychiatr. Serv.* 62, 675–678. doi: 10.1176/ps.62.6.pss6206_0675

Rush, A. J., and Beck, A. T. (1978). Cognitive therapy of depression and suicide. *Am. J. Psychother.* 32, 201–219. doi: 10.1176/appi.psychotherapy.1978.32.2.201

Sailer, P., Wieber, F., Propster, K., Stoewer, S., Nischk, D., Volk, F., et al. (2015). A brief intervention to improve exercising in patients with schizophrenia: a controlled pilot study with mental contrasting and implementation intentions (MCII). *BMC Psychiatry* 15:211. doi: 10.1186/s12888-015-0513-y

Sheeran, P., Aubrey, R., and Kellett, S. (2007). Increasing attendance for psychotherapy: implementation intentions and the self-regulation of attendance-related negative affect. *J. Consult. Clin. Psychol.* 75, 853–863. doi: 10.1037/0022-006X.75.6.853

Shiffman, S., Stone, A. A., and Hufford, M. R. (2008). Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* 4, 1–32. doi: 10.1146/annurev.clinpsy.3.022806.091415

Siegel, J. T., Alvaro, E. M., Crano, W. D., Gonzalez, A. V., Tang, J. C., and Jones, S. P. (2010). Passive-positive organ donor registration behavior: a mixed method assessment of the IIFF model. *Psychol. Health Med.* 15, 198–209. doi: 10.1080/13548501003623922

Siegel, J. T., Lienemann, B. A., and Rosenberg, B. D. (2017). Resistance, reactance, and misinterpretation: highlighting the challenge of persuading people with depression to seek help. *Soc. Personal. Psychol. Compass* 11:e12322. doi: 10.1111/spc3.12322

Siegel, J. T., Lienemann, B. A., and Tan, C. N. (2015). Influencing help-seeking among people with elevated depressive symptomatology: mistargeting as a persuasive technique. *Clin. Psychol. Sci.* 3, 242–255. doi: 10.1177/2167702614542846

Siegel, J. T., and Navarro, M. (2019). A conceptual replication examining the risk of overtly listing eligibility criteria on Amazon’s mechanical Turk. *J. Appl. Soc. Psychol.* 49, 239–248. doi: 10.1111/jasp.12580

Sin, N. L., Della Porta, M. D., and Lyubomirsky, S. (2011). “Tailoring positive psychology interventions to treat depressed individuals” in *Applied positive psychology: improving everyday life, health, schools, work, and society*. eds. S. I. Donaldson, M. Csikszentmihalyi and J. Nakamura (New York, NY: Routledge), 79–96.

Toli, A., Webb, T. L., and Hardy, G. E. (2016). Does forming implementation intentions help people with mental health problems to achieve goals? A meta-analysis of experimental studies with clinical and analogue samples. *Br. J. Clin. Psychol.* 55, 69–90. doi: 10.1111/bjc.12086

Vindegaard, N., and Benros, M. E. (2020). COVID-19 pandemic and mental health consequences: systematic review of the current evidence. *Brain Behav. Immun.* 89, 531–542. doi: 10.1016/j.bbi.2020.05.048

Wang, G., Wang, Y., and Gai, X. (2021). A meta-analysis of the effects of mental contrasting with implementation intentions on goal attainment. *Front. Psychol.* 12:565202. doi: 10.3389/fpsyg.2021.565202

Wang, M., Zhao, Q., Hu, C., Wang, Y., Cao, J., Huang, S., et al. (2021). Prevalence of psychological disorders in the COVID-19 epidemic in China: a real world cross-sectional study. *J. Affect. Disord.* 281, 312–320. doi: 10.1016/j.jad.2020.11.118

Webb, T. L., and Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychol. Bull.* 132, 249–268. doi: 10.1037/0033-2909.132.2.249

Wenze, S. J., and Miller, I. W. (2010). Use of ecological momentary assessment in mood disorders research. *Clin. Psychol. Rev.* 30, 794–804. doi: 10.1016/j.cpr.2010.06.007

White, I. R., Horton, N. J., Carpenter, J., and Pocock, S. J. (2011). Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ* 342:d40. doi: 10.1136/bmj.d40

World Health Organization (2021). Depression: WHO. Available at: https://www.who.int/news-room/fact-sheets/detail/depression (Accessed: November 29 2022).

Zhou, H., and Fishbach, A. (2016). The pitfall of experimenting on the web: how unattended selective attrition leads to surprising (yet false) research conclusions. *J. Pers. Soc. Psychol.* 111, 493–504. doi: 10.1037/pspa0000056

frontiers | Frontiers in Psychology

TYPE Brief Research Report
PUBLISHED 05 July 2023
DOI 10.3389/fpsyg.2023.1127535

OPEN ACCESS

EDITED BY
Rashmi Gupta,
Indian Institute of Technology Bombay, India

REVIEWED BY
Giulia Maria Giordano,
University of Campania Luigi Vanvitelli, Italy
Francisco Reyes-Madrigal,
Manuel Velasco Suárez National Institute of Neurology and Neurosurgery, Mexico

*CORRESPONDENCE
Vanessa Sanchez-Gistau
✉ sanchezv@peremata.com

RECEIVED 19 December 2022
ACCEPTED 26 May 2023
PUBLISHED 05 July 2023

CITATION
Sanchez-Gistau V, Cabezas A, Manzanares N, Sole M, Corral L, Vilella E and Gutierrez-Zotes A (2023) Cognitive biases in first-episode psychosis with and without attention-deficit/hyperactivity disorder. *Front. Psychol.* 14:1127535.
doi: 10.3389/fpsyg.2023.1127535

Cognitive biases in first-episode psychosis with and without attention-deficit/hyperactivity disorder

Vanessa Sanchez-Gistau[1,2,3,4]*, Angel Cabezas[1,2,3,4], Nuria Manzanares[1,4], Montse Sole[1,2,3,4], Lia Corral[1,2,3,4], Elisabet Vilella[1,2,3,4] and Alfonso Gutierrez-Zotes[1,2,3,4]

[1]Hospital Universitari Institut Pere Mata of Reus, Reus, Spain, [2]Institut d'Investigació Sanitària Pere Virgili (IISPV- CERCA), Tarragona, Spain, [3]Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), ISCIII, Madrid, Spain, [4]Universitat Rovira i Virgili (URV), Reus, Spain

Introduction: Psychotic disorders such schizophrenia and attention-deficit/hyperactivity disorder (ADHD) are neurodevelopmental disorders with social cognitive deficits. Specifically, biased interpretation of social information can result in interpersonal difficulties. Cognitive biases are prevalent in psychosis, but no previous study has investigated whether the type and severity of cognitive biases differ between subjects experiencing first-episode psychosis (FEP) with (FEP-ADHD$^+$) and without ADHD (FEP-ADHD$^-$).

Methods: A total of 121 FEP outpatients at the Early Intervention Service of Reus were screened for childhood ADHD through the Diagnostic Interview for ADHD (DIVA). Cognitive biases were assessed by the Cognitive Biases Questionnaire for Psychosis (CBQp). CBQp scores of FEPs groups were compared with those of healthy controls (HCs) with an analysis of covariance. Spearman correlation analysis explored associations between CBQp scores and psychopathology.

Results: Thirty-one FEPs met the criteria for childhood ADHD and reported significantly more cognitive bias [median (interquartile range): 47 (38–56)] than FEP-ADHD$^-$ [42 (37–48)] and HCs [38 (35.5–43)]. CBQp scores did not differ between FEP-ADHD-and HCs when adjusted for age and sex. After controlling for clinical differences, Intentionalising (F =20.97; p <0.001) and Emotional Reasoning biases (F =4.17; p =0.04) were more strongly associated with FEP-ADHD$^+$ than FEP-ADHD$^-$. Cognitive biases were significantly correlated with positive psychotic symptoms in both groups but only with depressive symptoms in FEP-ADHD$^-$ (r =0.258; p =0.03) and with poor functioning in FEP-ADHD$^+$ (r =−0.504; p =0.003).

Conclusion: Cognitive bias severity increased from HCs to FEP-ADHD-patients to FEP-ADHD$^+$ patients. FEP-ADHD$^+$ patients may be a particularly vulnerable group in which metacognitive targeted interventions are needed.

KEYWORDS

cognitive bias, CBQp, ADHD, first episode psychosis (FEP), schizophrenia, social cognition

Introduction

The cognitive model of psychosis suggests that psychotic symptoms may arise because of biased information processing (Garety et al., 2007). In accordance with this model, people with sub threshold (Livet et al., 2020) and full psychotic symptoms (Freeman et al., 2001) are more prone to cognitive biases. Cognitive biases refer to automatic errors in both cognitive processing and content across specific situations (Beck, 1963). A substantial body of research has demonstrated that cognitive biases contribute to the processes of reasoning and metacognition (Garety et al., 2001; Freeman, 2007; Morrison et al., 2007; Bob et al., 2016). From this perspective, the development and maintenance of delusions may be due to the presence of dysfunctional patterns of thought that leads to incorrect judgments and abnormal interpretations or perceptions. The cognitive biases in psychosis that have been most extensively studied are jumping to conclusions (JTC) (Ross et al., 2015; Dudley et al., 2016; McLean et al., 2017) attributional biases (Langdon et al., 2010; Sanford and Woodward, 2017), and belief inflexibility (Moritz and Woodward, 2006). However, it is evident that subjects with psychotic disorder present varied cognitive biases (De Rossi and Georgiades, 2022) including Beck's emotional biases (Beck, 1963). Thus, the emotional biases of catastrophising (C) and dichotomous thinking (DT) have also been involved in psychoses (Gawęda and Prochwicz, 2015). Peters E and colleagues developed the Cognitive Biases Questionnaire for Psychosis (CBQp) (Peters et al., 2014) to easily and comfortably assess cognitive biases in psychosis. The CBQp was based on the Blackburn Cognitive Styles Test (Blackburn et al., 1986) which was designed to assess frequent cognitive biases in depression. The cognitive biases included in the CBQp are jumping to conclusions (JTC), dichotomous thinking (DT), intentionalising (Int), emotional reasoning (ER), and catastrophizing (C).

Attention-deficit/hyperactivity disorder (ADHD) is a neurodevelopmental disorder affecting 3–7% of school-age children (Polanczyk et al., 2015). It is characterized by motor hyperactivity and impulsiveness and inattention or distractibility that produces functioning problems in the family and school environments' and in the relationship with peers; frequently, these difficulties persist in adulthood (Barnett, 2016). In addition to impairments in cognitive function, deficits in social cognition and interpersonal difficulties are also important features of ADHD. Within social cognition, deficits in theory of mind and emotion recognition and processing are the domains that have been most investigated; however, findings have been ambiguous (Morellini et al., 2022). Therefore, research on cognitive distortions in ADHD remains scarce, with some studies focusing on attentional and attribution bias (Hartmann et al., 2020; Jenness et al., 2021).

In addition to the genetic overlap between some risk alleles (Hamshere et al., 2013), psychotic disorders and ADHD share some clinical manifestations. Males are overrepresented, both have a high comorbidity with substance abuse, and both manifest difficulties in emotional regulation and peer relationships. Deficits in cognition are central symptoms of neurodevelopmental disorders and have been associated with poor functional outcomes and poor response to treatment. A previous report by our group (Sanchez-Gistau et al., 2020) compared cognitive performance between patients in their first episode of psychosis (FEP) with and without childhood ADHD (c-ADHD) and healthy controls (HCs); we found a gradient in the severity of cognitive impairment, with FEP patients with ADHD (FEP-ADHD$^+$ patients) being the most impaired. Compared to FEP-ADHD$^-$, FEP-ADHD$^+$ were more frequently men, showed a worse antipsychotic response and had a higher risk of drug consumption. The present study builds on this previous study by aiming to determine whether the type and severity of different cognitive biases (measured with the CBQp) differ between FEP patients with and without c-ADHD relative to a control group. In addition, we aimed to investigate the relationship between cognitive bias and psychopathological symptoms in both FEP-ADHD$^+$ and FEP-ADHD-patients.

Methods

Participants

We invited all consecutive outpatients referred to the Early Intervention Programme (EIP) at the University Hospital Institut Pere Mata of Reus, Spain, from January 2015 to July 2019 fulfilling the following inclusion criteria: age between 14 and 35 years; FEP, defined as the *"onset of full psychotic symptoms within the last 12 months"*; and less than 6 months of antipsychotic treatment. The exclusion criteria were as follows: psychosis induced by substances or other medical conditions, intellectual disability, severe head injury or a lack of fluency in Spanish. During the target period, 152 FEP subjects were referred to the EIP. Six subjects refused to participate and 15 did not fulfil the inclusion criteria. Of the 133 FEP subjects included in our previous study, 11 did not complete the CBQp. Therefore, the final sample consisted of 122 FEP subjects, 31 FEP-ADHD$^+$ subjects and 91 FEP-ADHD$^-$subjects. The sample of HCs ($N=26$) was drawn from our previous validation study of the CBQp in the Spanish language (Corral et al., 2020).

Ethical approval was obtained by the Committee for Ethical Clinical and Pharmacological Investigation of the Pere Virgili Research Institute (IISPV). After a complete description of the study was given to the subjects, written informed consent was obtained.

Assessments

Clinical assessments

Clinical assessments were administered by two experienced psychiatrists of the team. Clinical variables related to psychosis, such as the duration of untreated psychosis, current pharmacological treatment, and frequency of drug use in the past 6 months, were assessed through a direct interview. The dose of each antipsychotic was converted to chlorpromazine (CPMZ) equivalents in mg/day (Gardner et al., 2010). We defined as drug users those individuals who used a specific drug *"at least several times a week."* The severity of psychotic symptoms was assessed using the Positive and Negative Syndrome Scale (PANSS) (Kay et al., 1990) and the severity of affective symptoms by the Calgary Depression Scale for Schizophrenia (CDSS) (Addington et al., 1993) and the Young Mania Rating scale (YMRS) (Young et al., 1978). Finally the level of functioning was assessed by the Global Assessment of Functioning (GAF) (American Psychiatric Association, 1994).

The Spanish version of the Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders, 4th edition

(DSM-IV) for Axis I disorders (SCID-I) (First et al., 1997) confirmed the diagnosis of psychosis following DSM criteria. For descriptive purposes we grouped the diagnoses of schizophrenia, schizophreniform and schizoaffective disorders as "schizophrenia spectrum disorders"; manic and depressive episodes with psychotic symptoms as "affective psychoses" and brief psychotic disorders and psychosis not otherwise specified were categorized as "other psychosis."

Assessment of ADHD

A child and adolescent psychiatrist blind to clinical assessments administered the Spanish version of the Diagnostic Interview for ADHD in Adults (DIVA) (Ramos-Quiroga et al., 2019). Additional information was provided by a parent or close relative. The DIVA is the gold-standard assessment for ADHD in adults assessing the severity of each of the 18 symptoms required to meet the DSM-IV diagnostic criteria for ADHD in both childhood and adulthood. Those symptoms must cause impairment in at least two settings and must not be better explained by another psychiatric disorder. The diagnosis is considered definite when six or more criteria are met for each of the symptom domains of hyperactivity-impulsivity and/or attention deficit. We specifically asked about childhood-onset ADHD (c-ADHD) in order to avoid confusion with recent prodromal or current full psychotic symptoms. Adult-onset symptoms were therefore not considered for ADHD diagnoses. FEP subjects fulfilling the criteria for a definite diagnosis of c-ADHD were categorized as FEP-ADHD$^+$ otherwise; they were categorized as FEP-ADHD$^-$.

Assessment of cognitive biases

The CBQp (Peters et al., 2014) has been recently translated and validated for the Spanish population by our group (Corral et al., 2020). It is a self-report questionnaire containing 30 scenarios, 15 involving the theme of Anomalous Perception (AP) and 15 involving the theme of Threatening Events (TE); for each vignette, the subject is asked to choose one of three options that best describe that situation. Six vignettes are included for each of the five cognitive biases: Intentionalising (Int), Catastrophising (C), Dichotomous thinking (DT), Jumping to conclusions (JTC) and Emotional reasoning (ER). The score of each cognitive bias subscale ranges from 5 to 18, and the CBQp total score ranges from 30 to 90 points.

Statistical analyses

Demographic data were compared among the FEP-ADHD$^+$, FEP-ADHD-and HC groups using the chi-squared test with Yates' correction or Fisher's exact test for discrete variables and one-way analysis of variance (ANOVA) followed by *post hoc* Tukey pairwise comparisons. Differences in clinical variables between the two FEP groups were explored by the chi-squared test with Yates' correction or Fisher's exact test for discrete variables and Student's t test or the Mann–Whitney U test for continuous variables.

First, differences in cognitive bias (total scores, theme scores and bias scores) between the three groups were analysed by between-subject univariate analysis of variance (ANOVA) followed by *post hoc* Tukey pairwise comparisons. Significant group differences (at $p < 0.05$) were controlled by age and sex by subsequent univariate analysis of covariance (ANCOVA).

Second, differences in cognitive bias scores between clinical groups were analysed while controlling for the effects of clinical variables on which FEP-ADHD$^+$ and FEP-ADHD-groups differed at $p < 0.10$. Finally, Spearman correlation analyses were used to separately explore associations between the CBQp total score, TE and SA theme scores, cognitive bias scores and psychopathological symptoms in each clinical group.

All analyses were conducted using IBM SPSS for Windows, version 20.0 (IBM Corp., Armonk, NY).

Results

Sociodemographic and clinical characteristics

Thirty-one FEP subjects (25.4%) fulfilled the criteria for c-ADHD: 35.4% as the inattentive subtype, 25.8% as the hyperactive–impulsive subtype, and 38.8% as the combined subtype. Ten out of 31 ADHD subjects (32.25%) had been previously diagnosed with c-ADHD in a child and adolescent mental health unit, but only three were taking treatment for ADHD: one was on methylphenidate, one on guanfacine and one on atomoxetine.

As it can be seen in Table 1, the three groups significantly differed in age and sex. Both clinical groups were significantly younger than the HC group, and males were overrepresented in the FEP-ADHD$^+$ group (90.3%). The severity of clinical variables did not differ between the FEP groups, but FEP-ADHD$^+$ subjects used tobacco and cannabis more frequently and were treated with a higher dose and a greater number of antipsychotics than FEP-ADHD$^-$subjects.

Differences in cognitive biases

As shown in Table 2, the ANOVA revealed group differences in all cognitive biases except for JTC. *Post hoc* pairwise comparisons showed that FEP-ADHD$^+$ patients scored significantly higher than HCs on all scores. After adjusting for age and sex, these differences remained significant except for the C bias. FEP-ADHD-patients exhibited significantly higher scores than HCs on the CBQp total score, the TE theme and the ER bias; however, after adjusting for age and sex, these differences were no longer significant.

The two clinical groups were further directly compared after adjusting for sociodemographic and clinical differences at a threshold of $p < 0.10$, that is, after adjusting for sex, years of education, tobacco and cannabis use and antipsychotic dose (See Table 3). Compared to FEP-ADHD-patients, FEP-ADHD$^+$ patients scored significantly higher on the CBQp total score ($F = 7.11$; $p = 0.009$), TE ($F = 4.10$; $p = 0.04$) and AP ($F = 8.94$; $p = 0.003$) themes and Int ($F = 20.97$; $p < 0.001$) and RE ($F = 4.17$; $p = 0.04$) biases.

Correlation between psychopathological symptoms and cognitive biases

In the correlation analyses (Table 4), the CBQp total score was significantly correlated with the PANSS positive symptoms subscale score (PANSS-P) in both clinical groups and was correlated with the

TABLE 1 Socio-demographic and clinical characteristics of the groups.

	FEP Sample (*N* =122)	FEP-ADHD^{+} (*N* =31) (A)	FEP-ADHD$^{−}$ (*N* =91) (B)	Controls *N* =26 (C)	Statistic (χ^2, *t*)	*p*	*Post hoc*
Socio demographic variables							
Age, years (Mean, SD)	22.2 (5.4)	22.1 (4.8)	22.2 (5.61)	33.4 (6.5)	42.14	<0.001	A = B A < C***B < C***
Sex (*N*, % of male)	80 (65.6)	28 (90.3)	52 (57.1)	13 (50)	12.9	0.002	A > B**A < C**B = C
Education years (Median, IQR)	10 (8.5–13)	9 (8–10)	10 (9–13)		-1.82[a]	0.06	
Premorbid IQ	99.4 (15.6)	95.2 (14.9)	100.2 (16.1)		−1.39	0.16	
Clinical variables							
DUP (Median, IQR)	25 (10–60)	30 (15–87)	21 (10–60)		0.67[a]	0.47	
Diagnoses (*n*, %)					2.744	0.254	
Schizophrenia spectrum disorders	64 (52.2)	20 (64.5)	44 (48.4)				
Affective psychoses	26 (21.3)	4 (12.9)	22 (24.2)				
Other psychoses	32 (26.2)	7 (22.6)	25 (27.5)				
PANSS (Median, IQR)							
Positive	12 (9–18)	13 (8.5–18)	12 (9–18)		0.27[a]	0.78	
Negative	16.5 (11–26)	17 (11.5–26.5)	16 (10.5–26)		0.15[a]	0.87	
General	32.5 (25–44.5)	30 (23.5–44)	33 (26–45.5)		−0.29[a]	0.77	
total	63 (48–84.5)	55 (45.5–92.0)	64 (48.5–81.5)		0.19[a]	0.84	
GAF (Mean, SD)	57.60 (10.7)	59.7 (11.7)	56.9 (10.3)		1.23	0.21	
CDSS (Median, IQR)	1 (0–6)	1 (0–6.5)	1 (0–6)		0.83[a]	0.40	
YMRS (Median, IQR)	1 (0–9)	2 (0–13)	0 (0–9)		0.41 [a]	0.68	
Drug abuse (*n*, %)							
tobacco	67 (54.9)	25 (80.6)	42 (46.2)		9.76	0.002	
cannabis	23 (18.9)	11 (35.5)	12 (13.2)		6.12	0.010	
alcohol	23 (18.9)	9 (29.0)	14 (15.4)		1.99	0.16	
Treatment (*n*, %)							
Antipsychotics	115 (94.3)	30 (96.8)	85 (93.4)		0.06	0.80	
Number of APs (Median, IQR)	1 (1–1)	1 (1–2)	1 (1–1)		3.05[a]	0.002	
CPZE (mg/day) (Median, IQR)	300 (200–442.4)	399 (200–600)	300 (200–400)		1.71[a]	0.08	
Antidepressants	28 (23.0)	5 (16.1)	23 (25.3)		0.63	0.42	
Mood stabilizers	21 (17.2)	4 (12.9)	17 (18.7)		0.21	0.64	
Benzodiazepines	44 (36.1)	9 (29.0)	35 (38.5)		0.53	0.46	

SD, standard deviation; IQR, interquartile range; FEP, first episode of psychosis; ADHD, attention-deficit/hyperactivity disorder; DUP, duration of untreated psychoses; IQ, intellectual coefficient; PANSS, positive and negative syndrome scale; CDSS, calgary depression scale for schizophrenia; YMRS, young mania rating scale; GAF, global assessment of functioning; AP, antipsychotic treatment; CPZE, estimated equivalent amount of chlorpromazine.

[a] Mann-Whitney *U* test.

**p* < 0.05; ***p* < 0.01; ****p* < 0.001.

PANSS general symptoms subscale score (PANSS-G) and CDSS score in only the FEP-ADHD$^{−}$ group. In the FEP-ADHD^{+} group, the CBQp total score was also correlated with lower GAF scores. Regarding the TE and AP themes, TE was correlated with the PANSS-P scores in both clinical groups and with the PANSS-G and CDSS scores in the FEP-ADHD$^{−}$ group. The AP theme score was associated with PANSS-P and PANSS-G scores in only the FEP-ADHD$^{−}$ group.

Regarding cognitive biases, Int was positively correlated with the PANSS-P score in FEP-ADHD-patients and with worse GAF scores in FEP-ADHD^{+} patients. Positive correlations were found between the C and DT bias scores and the PANSS-P score in both clinical groups and between the CDSS score in the FEP-ADHD$^{−}$ group. The DT bias score also correlated with the PANSS-G score in the FEP-ADHD$^{−}$ group. JTC correlated with PANSS-P score in FEP-ADHD^{+} patients and with general PANSS-G scores, and CDSS score in FEP-ADHD-patients. Finally, the ER bias score was associated with positive PANSS-P and PANSS-G scores in only in the FEP-ADHD$^{−}$ group.

Discussion

As far as we know, this is the first study to assess and compare the severity of cognitive bias (measured by -the CBQp) between FEP-ADHD^{+} and FEP-ADHD$^{−}$ subjects relative to HCs. Our results therefore must be considered preliminary and caution is required when interpreting the subsequent findings.

TABLE 2 Comparison of cognitive biases between the groups.

				ANOVA results				ANCOVA ^{a}results			
CBQp (median/ IQR)	FEP-ADHD+ *N*=31 (A)	FEP-ADHD- *N*=91 (B)	Controls *N*=26 (HC)	F (p) *Post-hoc* Between subjects				F (p) *Post-hoc* Between subjects			
CBQ total	47 (38–56)	42 (37–48)	38 (35.7–43)	9.11***	A > B*	A > C***	B > C*	5.88*	A > B**	A > C**	B = C
Themes											
TE (Threatening events)	24 (21–32)	22 (19–25)	20 (18–22)	8.23***	A > B*	A > C***	B > C*	4.31**	A > B*	A > C**	B = C
AP (Anomalous perception)	23 (18–26)	20 (18–23)	18 (17–20)	7.44**	A > B*	A > C**	B = C	5.85***	A > B**	A > C**	
Biases											
Int (Intentionalising)	9 (7–12)	8 (6–8)	7 (6–8)	17.33***	A > B***	**A > C*****	B = C	12.61***	A > B***	**A > C****	
C (Catastrophism)	9 (8–11)	8 (7–10)	8 (7–9)	3.38*	A = B	**A > C***	B = C	2.76			
DT (Dichotomous thinking)	10 (8–12)	8 (7–10)	7 (7–8)	8.86***	A > B*	**A > C***	**B = C**	4.40**	A > B*	**A > C****	
JTC (Jumping to conclusions)	9 (9–11)	9 (8–11)	9 (9–10)	0.83				0.91			
ER (Emotional reasoning)	9 (7–11)	8 (7–10)	7 (6–8)	7.72**	A > B*	**A > C*****	**B > C***	2.80*	A > B*	**A > C****	B = C

IQR, interquartile range; FEP, first episode of psychosis; ADHD, attention-deficit/hyperactivity disorder; HC, healthy control; A, PEP-ADHD^{+}; B, PEP-ADHD^{-}; C, healthy controls.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.
[a]Differences between 3 groups adjusting for sex and age.

We found a gradient of cognitive bias severity from the FEP-ADHD^{+} group: [median (interquartile range) 47 (38–56)], to the FEP-ADHD^{-} group [42 (37–48)] to the HC group [38 (35.5–43)]. The FEP-ADHD^{+} group presented significantly higher scores not only on the total score but also in the CBQp themes of TE and AP than both the FEP-ADHD-and HC groups. Regarding specific cognitive biases, FEP-ADHD^{+} patients exhibited greater Int, DT, C and ER bias scores than HCs, while FEP-ADHD-patients presented only greater ER bias scores than HCs. The FEP-ADHD^{+} group therefore showed the most marked differences from HCs. However, the HC group used as a control group was not specifically recruited for the present study but was used in our previous study to validate the questionnaire in the Spanish population (Corral et al., 2020). Consequently, groups differed in terms of age and sex ratios. When the results were further adjusted for these differences, differences between FEP-ADHD-patients and HCs were no longer significant. In contrast, FEP-ADHD^{+} patients still significantly scored higher than HCs, except for the C bias score. Unexpectedly, the three groups did not significantly differ in the JTC bias score. JTC is the most investigated cognitive bias in psychosis, and when assessed by the probabilistic behavior task (the beads task) (Garety et al., 2005), JTC bias has been frequently observed in patients with established schizophrenia (Ross et al., 2015; Dudley et al., 2016; McLean et al., 2017) and FEP (Falcone et al., 2015) as well as in those at clinical risk of psychosis (Livet et al., 2020). However, when assessed by self-report questionnaires, mixed results have been reported (Bastiaens et al., 2013; Ahuir et al., 2021; Pena-Garijo et al., 2022; Pugliese et al., 2022). One possible explanation is that patients with psychosis often have a little awareness of their cognitive deficits and biases (Moritz et al., 2004). Therefore, a dissociation between the objective (assessed by task performance) and subjective measures (assessed by self-report) cannot be ruled out. In addition, Beck's cognitive biases have an emotional component rather than a psychotic cognitive–perceptual component, which may explain why they are also present in the healthy population (Bastiaens et al., 2018). Unfortunately, depressive and anxiety symptoms were not evaluated in the HC group. Thus, it would have been useful in order to identify a potential relationship between the emotional component and the cognitive distortions in that group.

A novel finding of the present study is that FEP-ADHD^{+} patients showed more severe cognitive biases than FEP-ADHD-patients, even after controlling for clinical and sociodemographic differences. Apart from scoring higher on the two themes (AP and TE), it is particularly interesting that the FEP-ADHD^{+} group exhibited more DT and significantly more Int and ER than the FEP-ADHD^{-} group. Consistent with these findings, in their original validation report comparing subjects with psychosis with HCs and depressed subjects, Peters and colleagues (Peters et al., 2014) reported that the psychosis group scored higher than the depressed group on the Int and ER biases. Moreover, Int was the only bias where the depressed group and HCs did not score differently. The authors suggested that these two specific biases may represent a particular "paranoid thinking style" that distinguishes individuals with psychosis from other clinical populations. In accordance, we found that FEP-ADHD^{+} patients were more likely to exhibit the Int bias ($F = 20.97$; $p < 0.001$). The Int bias refers to the implicit and automatic inclination to interpret human actions as intentional and to think that negative actions toward oneself were committed on purpose (i.e., intentionally). In the ADHD literature, research on interpretation bias is very scarce. There is

TABLE 3 Comparison of CBQ between FEP ADHD+ and FEP-ADHD-.

Cognitive Biases (median, IQR)	FEP sample (*N*=122)	FEP-ADHD$^+$ (*N*=31)	FEP-ADHD$^-$ (*N*=91)	ANCOVA^a Statistic	*p*
CBQp Total	43 (37.5–50.5)	47 (38–56)	42 (37–48)	7.11	0.009
Themes					
TE (Threatening Events)	22 (20–26)	24 (21–32)	22 (19–25)	4.10	0.04
AP (Anomalous Perception)	20 (18–24)	23 (18–26)	20 (18–23)	8.94	0.003
Biases					
Int (Intentionalising)	8 (7–9)	9 (7–12)	8 (6–8)	20.97	<0.001
C (Catastrophism)	8 (7–10)	9 (8–11)	8 (7–10)	2.95	0.10
DT (Dichotomous thinking)	9 (7–10)	10 (8–12)	8 (7–10)	2.53	0.11
JTC (Jumping to conclusions)	9 (8–11)	9 (9–11)	9 (8–11)	0.34	0.56
ER (Emotional reasoning)	8 (7–10)	9 (7–11)	8 (7–10)	4.17	0.04

IQR, interquartile range; CBQp, cognitive biases questionnaire for psychosis; IQR, interquartile range.
FEP, first episode of psychosis; ADHD, attention-deficit/hyperactivity disorder.
[a]adjusted by: sex, years of education, cannabis and tobacco use and CPZE estimated equivalent amount of chlorpromazine dose.

inconsistent evidence that hostile attribution bias (HAB) with ambiguous situations and ambiguous faces occurs more frequently in children and adults with ADHD than in HCs (King et al., 2009; Sibley et al., 2010; Schneidt et al., 2019). No previous study to date has addressed this issue in FEP patients with and without ADHD; thus, further investigation is needed to replicate or refute our results and disentangle whether FEP patients with c-ADHD exhibit greater cognitive bias than FEP patients without c-ADHD.

Notably, when exploring the relationship between cognitive biases and psychopathological symptoms, only weak correlations were found. Regarding specific biases, the severity of positive symptoms was associated with the C and DT biases in both clinical groups; however, the severity of positive symptoms was associated with the JTC bias in the FEP-ADHD$^+$ group and with the Int and ER biases in the FEP-ADHD$^-$ group. Nevertheless, positive correlations between cognitive biases and depressive symptoms and general symptoms were found only in FEP-ADHD-patients. Although neither clinical group differed in psychopathological symptoms or functioning, our study suggests a different pattern of biases related to positive, general and depressive symptoms in FEP-ADHD$^+$ patients compared to FEP-ADHD-patients. Moreover, poor functioning was associated with the CBQp total score and the Int bias in only the FEP-ADHD$^+$ group. Given the lack of previous studies, we can only speculate that the dominant cognitive biases in FEP-ADHD-patients may be related to depressive and anxious symptoms. On the other hand, the severity, type of cognitive biases and the lack of relationship with depressive and general symptoms in the FEP-ADHD$^+$ group may reflect traditional psychotic thinking, which in turn might be associated with worse functioning. Moreover, children with ADHD present deficits in recognizing facial emotions and others' emotional states in addition to deficits in emotional processing (Pishyareh et al., 2015). Thus, we speculate that greater cognitive biases in the FEP-ADHD$^+$ group may interact with the social processing difficulties already present in ADHD to pose a higher risk of impaired functioning.

Our results extend previous findings suggesting that young adults with c-ADHD and FEP suffer additional impairments (Peralta et al., 2011; Rho et al., 2015; Sanchez-Gistau et al., 2020). Specifically, we report for the first time greater cognitive biases (in general) and more severe Int and ER biases (in particular) in FEP-ADHD$^+$ patients. Together, these findings indicate that FEP-ADHD$^+$ subjects may be a particularly vulnerable group and a high-priority target for interventions addressing both cognitive biases. Metacognitive training therapy (MCT) was developed by Moritz S and colleagues two decades ago to address problems related to cognitive biases and social cognition in psychosis (Moritz and Woodward, 2007). Previous research has indicated that MCT is an effective psychological intervention for people with schizophrenia (Moritz and Lysaker, 2018; Moritz et al., 2022). Specifically, in patients with recent-onset psychosis, MCT has demonstrated to be effective for improving psychotic symptoms, cognitive insight, and attributional style, as well as for reducing cognitive distortions (Ochoa et al., 2017; Ahuir et al., 2018). Our findings therefore may have clinical relevance for treatment recommendations. As reported, ADHD is a prevalent condition in FEP accompanied by prominent cognitive bias; thus, an adapted intervention for this subgroup aiming to reduce the most prevalent bias can be recommended. Our findings indicate the necessity of conducting metacognitive intervention studies specifically designed to assess the effectiveness of these particular interventions in this particular subgroup of patients.

Limitations and strengths

Some limitations must be taken into account when interpreting our finding. With regards the ADHD diagnoses, despite we used an structured interview for assessing ADHD symptoms recall bias regarding childhood onset symptoms cannot be entirely ruled out. We tried to avoid the possibility of overlapping ADHD symptoms with psychotic symptoms by restricting the diagnoses of ADHD to childhood-onset, that is, onset of symptoms before the age of 7 years according to the DSM-IV criteria. Moreover, the healthy control group HCs who participated in the previous validation study was not specifically assessed for ADHD and differed in terms of age and sex distribution. Second, despite controlling for clinical and

TABLE 4 Association of cognitive biases with psychopathological symptoms and daily functioning in the clinical groups.

	FEP-ADHD⁺ (N=31)					FEP-ADHD⁻ (N=91)				
	PANSS-P	PANSS-N	PANSS-G	CDSS	YMRS	GAF	PANSS-P	PANSS-N	PANSS-G	CDSS	YMRS	GAF
CBQt	0.410*	0.202	0.253	0.225	0.167	−0.504**	0.379**	0.092	0.260**	0.258*	0.207	−0.188
CBQ-TE	0.468*	0.232	0.357	0.248	0.171	−0.214	0.273*	0.023	0.186	0.257*	0.070	- 0.157
CBQ-AP	0.259	0.181	0.117	0.135	0.140	−0.390*	0.433**	0.184	0.293**	0.159	0.207	−0.183
CBQ-Int	0.208	0.157	0.072	0.158	0.003	−0.338*	0.236*	0.030	0.049	0.057	0.149	0.037
CBQ-C	0.415*	0.276	0.291	0.162	0.284	−0.306	0.211*	0.062	0.163	0.216*	0.133	−0.086
CBQ-DT	0.294*	0.004	0.251	0.226	−0.079	−0.243	0.377**	0.045	0.284**	0.364**	0.109	−0.265*
CBQ-JTC	0.444*	0.190	0.214	0.152	0.136	−0.458	0.196	0.089	0.226*	0.292**	0.052	−0.230*
CBQ-ER	0.334	0.286	0.260	0.098	0.328	−0.238	0.357**	0.158	0.272*	0.135	0.144	−0.129

Spearman correlation coefficients (r); *p value < 0.05; **p value < 0.01.

FEP, first episode of psychosis; ADHD, attention-deficit/hyperactivity disorder; HC, healthy control; PANSS, positive and negative syndrome scale; PANSS-P, PANSS positive symptom subscale; PANSS-N, PANSS negative symptom subscale; PANSS-G, PANSS general symptom subscale; CDSS, calgary depression scale for schizophrenia; YMRS, young mania rating scale; GAF, global assessment of functioning; CBQp, cognitive biases questionnaire for psychosis; CBQt, CBQp total score; CBQ-TE, CBQp threatening events (TE) theme; CBQ-AP, CBQp anomalous perceptions (AP) theme; CBQ-I, CBQp intentionalising (I) subscore; CBQ-C, CBQp catastrophising (C) subscore; CBQ-DT, CBQp dichotomous thinking (DT) subscore; CBQ-JTC, CBQp jumping to conclusions (JTC) subscore; CBQ-ER, CBQp emotional reasoning (ER) subscore.

socio-demographic differences, the scores of CBQp might have been influenced by other variables that were not adjusted for, such as variables related to stress and childhood trauma. It has also to be acknowledged that sample size limited our ability to conduct secondary analyses stratified by ADHD subtype or by psychotic diagnoses. Moreover, the low percentage of females prevented us to investigate sex differences in the studied variables and ADHD.

However, despite these limitations, we have included a real-world clinical practice sample in their early stages of the illness coming from a particular geographical area. Our relatively homogeneous sample, allows us to minimize the impact of the burden of a chronic disease and long-term antipsychotic treatment. Finally cross-sectional assessment did not allow us to infer a causal relationship between cognitive bias and ADHD in FEP patients.

In summary, we report a gradient of severity in CBQp scores among the three groups, with the FEP-ADHD⁺ group differing the most markedly from the FEP-ADHD-and HC groups. The severity of cognitive biases, however, did not differ between the FEP-ADHD-and HC groups after adjusting for age and sex. Importantly, the Int and ER biases were the most strongly associated with the FEP-ADHD⁺ group, but no bias was associated with the FEP-ADHD⁻ group.

Conclusion

Our present findings together with previous findings indicate that FEP-ADHD⁺ subjects represent a clinical subgroup with a worse potential prognosis than FEP-ADHD⁻ subjects. Further research on the relationships among cognitive biases, cognitive performance and environmental factors are needed to develop individualized pharmacological and psychological interventions, such as MCT, for FEP subpopulations. The relationship between ADHD and psychosis is still an important knowledge gap that requires further investigation.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon a reasonable request.

Ethics statement

The studies involving human participants were reviewed and approved by the Committee for Ethical Clinical and Pharmacological Investigation of the Pere Virgili Research Institute (CEIM of IISPV). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

VS-G designed the current study. AC, MS, VS-G, and LC contributed to the acquisition of clinical data. NM was blinded to the results of the clinical assessments and performed the ADHD interviews. VS-G, AG-Z, and EV undertook the statistical analysis and all authors took part in the interpretation of the results. VS-G wrote the first draft of the article. All authors critically revised the first draft and provided their contributions and have approved the final manuscript.

Funding

This work was supported by the AGAUR Agència de Gestió d'Ajuts Universitaris i de Recerca (SGR2017_444).

Acknowledgments

The authors would like to acknowledge the participants and families who gave freely of their time to take part in the research. We would also like to thank Encarna Peñas for its assistance in participant recruitment.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Addington, D., Addington, J., and Maticka-Tyndale, E. (1993). Assessing depression in schizophrenia: the calgary depression scale. *Br. J. Psychiatry Suppl.* 22, 39–44.

Ahuir, M., Cabezas, Á., Miñano, M. J., Algora, M. J., Estrada, F., Solé, M., et al. (2018). Improvement in cognitive biases after group psychoeducation and metacognitive training in recent-onset psychosis: a randomized crossover clinical trial. *Psychiatry Res.* 270, 720–723. doi: 10.1016/j.psychres.2018.10.066

Ahuir, M., Crosas, J. M., Estrada, F., Zabala, W., Pérez-Muñoz, S., González-Fernández, A., et al. (2021). Cognitive biases are associated with clinical and functional variables in psychosis: a comparison across schizophrenia, early psychosis and healthy individuals. *Revista de Psiquiatria y Salud Mental* 14, 4–15. doi: 10.1016/j.rpsm.2020.07.005

American Psychiatric Association. *DSM-IV. Diagnostic and statistical manual of mental disorders*. Washington: APA, (1994).

Barnett, R. (2016). Attention deficit hyperactivity disorder. *Lancet* 387:737. doi: 10.1016/s0140-6736(16)00332-9

Bastiaens, T., Claes, L., Smits, D., De Wachter, D., van der Gaag, M., and De Hert, M. (2013). The cognitive biases questionnaire for psychosis (CBQ-P) and the Davos assessment of cognitive biases (DACOBS): validation in a Flemish sample of psychotic patients and healthy controls. *Schizophr. Res.* 147, 310–314. doi: 10.1016/j.schres.2013.04.037

Bastiaens, T., Claes, L., Smits, D., Vanwalleghem, D., and De Hert, M. (2018). Self-reported cognitive biases are equally present in patients diagnosed with psychotic versus nonpsychotic disorders. *J. Nerv. Ment. Dis.* 206, 122–129. doi: 10.1097/NMD.0000000000000763

Beck, A. T. (1963). Thinking and depression I- idiosyncratic content and cognitive distortions. *Arch. Gen. Psychiatry* 9, 324–333. doi: 10.1001/archpsyc.1963.01720160014002

Blackburn, I. M., Jones, S., and Lewin, R. J. (1986). Cognitive style in depression. *Br. J. Clin. Psychol.* 25, 241–251. doi: 10.1111/j.2044-8260.1986.tb00704.x

Bob, P., Pec, O., Mishara, A. L., Touskova, T., and Lysaker, P. H. (2016). Conscious brain, metacognition and schizophrenia. *Int. J. Psychophysiol.* 105, 1–8. doi: 10.1016/j.ijpsycho.2016.05.003

Corral, L., Labad, J., Ochoa, S., Cabezas, A., Muntané, G., Valero, J., et al. (2020). Cognitive biases questionnaire for psychosis (CBQp): Spanish validation and relationship with cognitive insight in psychotic patients. *Front. Psych.* 11:596625. doi: 10.3389/fpsyt.2020.596625Ç

De Rossi, G., and Georgiades, A. (2022). Thinking biases and their role in persecutory delusions: a systematic review. *Early Interv. Psychiatry* 16, 1278–1296. doi: 10.1111/eip.13292

Dudley, R., Taylor, P., Wickham, S., and Hutton, P. (2016). Psychosis, delusions and the "jumping to conclusions" reasoning Bias: a systematic review and Meta-analysis. *Schizophr. Bull.* 42, 652–665. doi: 10.1093/schbul/sbv150

Falcone, M. A., Murray, R. M., Wiffen, B. D. R., O'Connor, J. A., Russo, M., Kolliakou, A., et al. (2015). Jumping to conclusions, neuropsychological functioning, and delusional beliefs in first episode psychosis. *Schizophr. Bull.* 41, 411–418. doi: 10.1093/schbul/sbu104

First, M.B., Spitzer, R., and Gibbon, M. *Structured clinical interview for DSM-IV Axis I disorders*. Washington, DC: American Psychiatric Press Inc;(1997). SCID-I. Entrevista clínica estructurada para los trastornos del eje I. del DSM-IV. Masson, Barcelona 1999

Freeman, D. (2007). Suspicious minds: the psychology of persecutory delusions. *Clin. Psychol. Rev.* 27, 425–457. doi: 10.1016/j.cpr.2006.10.004

Freeman, D., Garety, P. A., and Kuipers, E. (2001). Persecutory delusions: developing the understanding of belief maintenance and emotional distress. *Psychol. Med.* 31, 1293–1306. doi: 10.1017/s003329170100455x

Gardner, D. M., Murphy, A. L., O'Donnell, H., Centorrino, F., and Baldessarini, R. J. (2010). International consensus study of antipsychotic dosing. *Am. J. Psychiatry.* 167, 686–693. doi: 10.1176/appi.ajp.2009.09060802

Garety, P. A., Bebbington, P., Fowler, D., Freeman, D., and Kuipers, E. (2007). Implications for neurobiological research of cognitive models of psychosis: a theoretical paper. *Psychol. Med.* 37, 1377–1391. doi: 10.1017/S003329170700013X

Garety, P. A., Freeman, D., Jolley, S., Dunn, G., Bebbington, P. E., Fowler, D. G., et al. (2005). Reasoning, emotions, and delusional conviction in psychosis. *J. Abnorm. Psychol.* 114, 373–384. doi: 10.1037/0021-843X.114.3.373

Garety, P. A., Kuipers, E., Fowler, D., Freeman, D., and Bebbington, P. E. (2001). A cognitive model of the positive symptoms of psychosis. *Psychol. Med.* 31, 189–195. doi: 10.1017/s0033291701003312

Gawęda, Ł., and Prochwicz, K. (2015). A comparison of cognitive biases between schizophrenia patients with delusions and healthy individuals with delusion-like experiences. *Eur. Psychiatry* 30, 943–949. doi: 10.1016/j.eurpsy.2015.08.003

Hamshere, M. L., Stergiakouli, E., Langley, K., Martin, J., Holmans, P., Kent, L., et al. (2013). Shared polygenic contribution between childhood attention-deficit hyperactivity disorder and adult schizophrenia. *Br. J. Psychiatry J. Ment. Sci.* 203, 107–111. doi: 10.1192/bjp.bp.112.117432

Hartmann, D., Ueno, K., and Schwenck, C. (2020). Attributional and attentional bias in children with conduct problems and callous-unemotional traits: a case-control study. *Child Adolesc. Psychiatry Ment. Health* 14, 1–11. doi: 10.1186/s13034-020-00315-9

Jenness, J. L., Lambert, H. K., Bitrán, D., Blossom, J. B., Nook, E. C., Sasse, S. F., et al. (2021). Developmental variation in the associations of attention Bias to emotion with internalizing and externalizing psychopathology. *Res. Child Adolesc. Psychopathol.* 49, 711–726. doi: 10.1007/s10802-020-00751-3

Kay, S. R., Fiszbein, A., Vital-Herne, M., and Fuentes, L. S. (1990). The positive and negative syndrome scale--Spanish adaptation. *J. Nerv. Ment. Dis.* 178, 510–517. doi: 10.1097/00005053-199008000-00007

King, S., Waschbusch, D. A., Pelham, W. E. J., Frankland, B. W., Andrade, B. F., Jacques, S., et al. (2009). Social information processing in elementary-school aged children with ADHD: medication effects and comparisons with typical children. *J. Abnorm. Child Psychol.* 37, 579–589. doi: 10.1007/s10802-008-9294-9

Langdon, R., Ward, P. B., and Coltheart, M. (2010). Reasoning anomalies associated with delusions in schizophrenia. *Schizophr. Bull.* 36, 321–330. doi: 10.1093/schbul/sbn069

Livet, A., Navarri, X., Potvin, S., and Conrod, P. (2020). Cognitive biases in individuals with psychotic-like experiences: a systematic review and a meta-analysis. *Schizophr. Res.* 222, 10–22. doi: 10.1016/j.schres.2020.06.016

McLean, B. F., Mattiske, J. K., and Balzan, R. P. (2017). Association of the jumping to conclusions and evidence integration biases with delusions in psychosis: a detailed meta-analysis. *Schizophr. Bull.* 43, sbw056–sbw354. doi: 10.1093/schbul/sbw056

Morellini, L., Ceroni, M., Rossi, S., Zerboni, G., Rege-Colet, L., Biglia, E., et al. (2022). Social cognition in adult ADHD: a systematic review. *Front. Psychol.* 13, 1–10. doi: 10.3389/fpsyg.2022.940445

Moritz, S., Ferahli, S., and Naber, D. (2004). Memory and attention performance in psychiatric patients: lack of correspondence between clinician-rated and patient-rated functioning with neuropsychological test results. *J. Int. Neuropsychol. Soc.* 10, 623–633. doi: 10.1017/S1355617704104153

Moritz, S., and Lysaker, P. H. (2018). Metacognition research in psychosis: uncovering and adjusting the prisms that distort subjective reality. *Schizophr. Bull.* 45, 17–18. doi: 10.1093/schbul/sby151

Moritz, S., Menon, M., Balzan, R., and Woodward, T. S. (2022). Metacognitive training for psychosis (MCT): past, present, and future. *Eur. Arch. Psychiatry Clin. Neurosci.* 1–7, 1–7. doi: 10.1007/s00406-022-01394-9

Moritz, S., and Woodward, T. S. (2006). A generalized bias against disconfirmatory evidence in schizophrenia. *Psychiatry Res.* 142, 157–165. doi: 10.1016/j.psychres.2005.08.016

Moritz, S., and Woodward, T. S. (2007). Metacognitive training in schizophrenia: from basic research to knowledge translation and intervention. *Curr. Opin. Psychiatry* 20, 619–625. doi: 10.1097/YCO.0b013e3282f0b8ed

Morrison, A. P., French, P., and Wells, A. (2007). Metacognitive beliefs across the continuum of psychosis: comparisons between patients with psychotic disorders, patients at ultra-high risk and non-patients. *Behav. Res. Ther.* 45, 2241–2246. doi: 10.1016/j.brat.2007.01.002

Ochoa, S., López-Carrilero, R., Barrigón, M. L., Pousa, E., Barajas, A., Lorente-Rovira, E., et al. (2017). Randomized control trial to assess the efficacy of metacognitive training compared with a psycho-educational group in people with a recent-onset psychosis. *Psychol. Med.* 47, 1573–1584. doi: 10.1017/S0033291716003421

Pena-Garijo, J., Palop-Grau, A., Masanet, M. J., Lacruz, M., Plaza, R., Hernández-Merino, A., et al. (2022). Self-reported cognitive biases in psychosis: validation of the Davos assessment of cognitive biases scale (DACOBS) in a Spanish sample of psychotic patients and healthy controls. *J. Psychiatr. Res.* 155, 526–533. doi: 10.1016/j.jpsychires.2022.09.041

Peralta, V., de Jalón, E. G., Campos, M. S., Zandio, M., Sanchez-Torres, A., and Cuesta, M. J. (2011). The meaning of childhood attention-deficit hyperactivity symptoms in patients with a first episode of schizophrenia-spectrum psychosis. *Schizophr. Res.* 126, 28–35. doi: 10.1016/j.schres.2010.09.010

Peters, E. R., Moritz, S., Schwannauer, M., Wiseman, Z., Greenwood, K. E., Scott, J., et al. (2014). Cognitive biases questionnaire for psychosis. *Schizophr. Bull.* 40, 300–313. doi: 10.1093/schbul/sbs199

Pishyareh, E., Tehrani-Doost, M., Mahmoodi-Gharaie, J., Khorrami, A., and Rahmdar, S. R. (2015). A comparative study of sustained attentional bias on emotional processing in ADHD children to pictures with eye-tracking. *Iran. J. Child Neurol.* 9, 64–70.

Polanczyk, G. V., Salum, G. A., Sugaya, L. S., Caye, A., and Rohde, L. A. (2015). Annual research review: a meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. *J. Child Psychol. Psychiatry* 56, 345–365. doi: 10.1111/jcpp.12381

Pugliese, V., Aloi, M., Maestri, D., de Filippis, R., Gaetano, R., Pelizza, L., et al. (2022). Validation of the Italian version of the Davos assessment of cognitive biases scale (DACOBS) in a sample of schizophrenia spectrum disorder patients and healthy controls. *Riv. Psichiatr.* 57, 127–133. doi: 10.1708/3814.37991

Ramos-Quiroga, J. A., Nasillo, V., Richarte, V., Corrales, M., Palma, F., Ibanez, P., et al. (2019). Criteria and concurrent validity of DIVA 2.0: a semi-structured diagnostic interview for adult ADHD. *J. Atten. Disord.* 23, 1126–1135. doi: 10.1177/1087054716646451

Rho, A., Traicu, A., Lepage, M., Iyer, S. N., Malla, A., and Joober, R. (2015). Clinical and functional implications of a history of childhood ADHD in first-episode psychosis. *Schizophr. Res.* 165, 128–133. doi: 10.1016/j.schres.2015.03.031

Ross, R. M., McKay, R., Coltheart, M., and Langdon, R. (2015). Jumping to conclusions about the beads task? A meta-analysis of delusional ideation and data-gathering. *Schizophr. Bull.* 41, 1183–1191. doi: 10.1093/schbul/sbu187

Sanchez-Gistau, V., Manzanares, N., Cabezas, A., Sole, M., Algora, M., and Vilella, E. (2020). Clinical and cognitive correlates of childhood attention-deficit/hyperactivity disorder in first-episode psychosis: a controlled study. *Eur. Neuropsychopharmacol.* 36, 90–99. doi: 10.1016/j.euroneuro.2020.05.010

Sanford, N., and Woodward, T. S. (2017). Symptom-related attributional biases in schizophrenia and bipolar disorder. *Cogn. Neuropsychiatry* 22, 263–279. doi: 10.1080/13546805.2017.1314957

Schneidt, A., Jusyte, A., and Schönenberg, M. (2019). Interpretation of ambiguous facial affect in adults with attention-deficit/hyperactivity disorder. *Eur. Arch. Psychiatry Clin. Neurosci.* 269, 657–666. doi: 10.1007/s00406-018-0879-1

Sibley, M. H., Evans, S. W., and Serpell, Z. N. (2010). Social cognition and interpersonal impairment in Young adolescents with ADHD. *J. Psychopathol. Behav. Assess.* 32, 193–202. doi: 10.1007/s10862-009-9152-2

Young, R. C., Biggs, J. T., Ziegler, V. E., and Meyer, D. A. (1978). A rating scale for mania: reliability, validity and sensitivity. *Br. J. Psychiatry* 133, 429–435. doi: 10.1192/bjp.133.5.429

TYPE Original Research
PUBLISHED 12 July 2023
DOI 10.3389/fpsyg.2023.1200627

OPEN ACCESS

EDITED BY
Dario Monzani,
University of Palermo, Italy

REVIEWED BY
Henry Markovits,
Université du Québec à Montréal, Canada
Patrice Rusconi,
University of Messina, Italy

*CORRESPONDENCE
Marguerite Beattie
✉ marguerite.beattie@helsinki.fi

RECEIVED 05 April 2023
ACCEPTED 19 June 2023
PUBLISHED 12 July 2023

CITATION
Beattie P and Beattie M (2023) Political polarization: a curse of knowledge? *Front. Psychol.* 14:1200627.
doi: 10.3389/fpsyg.2023.1200627

Political polarization: a curse of knowledge?

Peter Beattie[1] and Marguerite Beattie[2]*

[1]MGPE Programme, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China, [2]Faculty of Social Sciences, Faculty of Educational Sciences, University of Helsinki, Helsinki, Finland

Purpose: Could the curse of knowledge influence how antagonized we are towards political outgroups? Do we assume others know what we know but still disagree with us? This research investigates how the curse of knowledge may affect us politically, i.e., be a cause of political polarization.

Background: Research on the curse of knowledge has shown that even when people are incentivized to act as if others do not know what they know, they are still influenced by the knowledge they have.

Methods: This study consists of five studies consisting of both experimental and non-experimental and within- and between-subjects survey designs. Each study collected samples of 152–1,048.

Results: Partisans on both sides overestimate the extent to which stories from their news sources were familiar to contrapartisans. Introducing novel, unknown facts to support their political opinion made participants rate political outgroup members more negatively. In an experimental design, there was no difference in judging an opponent who did not know the same issue-relevant facts and someone who did know the same facts. However, when asked to compare those who know to those who do not, participants judged those who do not know more favorably, and their ratings of all issue-opponents were closer to those issue-opponents who shared the same knowledge. In a debiasing experiment, those who received an epistemological treatment judged someone who disagreed more favorably.

Conclusion: This research provides evidence that the curse of knowledge may be a contributing cause of affective political polarization.

KEYWORDS

curse of knowledge, cognitive bias, debiasing, political polarization, political epistemology

1. Introduction

> "*[T]he opponent presents himself as the man who says, evil be thou my good. … [H]e who denies either my moral judgments or my version of the facts, is to me perverse, alien, dangerous. How shall I account for him? The opponent has always to be explained, and the last explanation that we ever look for is that he sees a different set of facts.*"
>
> Walter Lippman, *Public Opinion.*

The "curse of knowledge" fundamentally consists of an impaired ability to imagine the reasoning of others who do not share one's knowledge. This is caused by an implicit presumption that one's own knowledge is shared by others, a presumption that is largely impervious to evidence that others do *not* share this knowledge (Birch and Bloom, 2007; Dębska and Komorowska, 2013; *cf.* Ryskin and Brown-Schmidt, 2014). This bias is likely produced by a combination of fluency misattribution (mistaking the fluency or ease with which information comes to mind, with how widely shared that information is) and a failure of inhibitory control (inhibiting one's own knowledge while estimating what others know; Birch et al., 2017). Drayton and Santos (2018) found evidence that non-human primates do not exhibit curse-of-knowledge effects, suggesting it is unique to the theory of mind humans have evolved. As such, it may have been evolutionarily adaptive for its efficiency; as hyper-cooperative or eusocial animals living in small groups for most of our history (Wilson, 2012), we inhabit "knowledge communities" wherein we take the knowledge held by members of our own community, or even the internet, to be the same as our own knowledge (Fisher et al., 2015; Sloman and Rabb, 2016; Rabb et al., 2019). The curse of knowledge is essentially the inverse: assuming our knowledge is the same as that held by members of our community.

The term "curse of knowledge" was first coined in an economic experiment (Camerer et al., 1989). The study tested the uncontroversial assumption that in economic situations featuring asymmetric information, marketplace participants with more information than others would be able to accurately predict the judgments of participants without this information — and profit from the information asymmetry. What the experimenters found, however, is that even with real money to be gained, their subjects had difficulty predicting the judgments of people *lacking* the information they had. Although they had been instructed that other marketplace participants lacked certain key information they had been given, subjects made investment decisions influenced by an apparently unconscious assumption that all other marketplace participants also knew what they knew — and lost money as a result. Rather than serving as an advantage, the knowledge unique only to the subjects operated as a curse.

Paradoxically, the authors noted that in economic settings, this curse of knowledge may actually *increase* social welfare (Camerer et al., 1989, p. 1245). Information asymmetries are conventionally thought to produce economic inefficiencies, as when a used car dealer overcharges for a "lemon," because only the seller knows hidden defects of which the buyer is ignorant. In such economic settings, the individual curse of knowledge might be a blessing for society, by making information asymmetries invisible to, and less likely to be exploited by, the party with more information. Hence in these economic settings, the curse of knowledge may be an example of psychological vice producing public virtue.

Yet in political settings, there is reason to expect the curse of knowledge to *reduce* social welfare. The curse of knowledge may describe a psychological "default" setting, or an innate theory of mind we use to understand others, in which everything we know is considered common knowledge, shared by all of our interlocutors (Nickerson, 1999). They are then treated accordingly, as if they knew the same information we have learned. This would make it more difficult to accurately understand the thinking of others who do *not* share knowledge that supports our political opinions. (Hence the *curse* of knowledge: it can make our understanding of other people's thinking worse than when we are ignorant of particular knowledge).

How might we understand the thinking of political opponents, when the curse of knowledge makes us implicitly assume that they have learned the same knowledge we have learned, i.e., the information that has shaped our opinion of an issue? For example, imagine two U.S. Americans in 2003 with opposing views on the invasion of Iraq. One may have supported the invasion on account of the following knowledge: claims of Iraqi weapons of mass destruction and links with al Qaeda, and past atrocities committed by the Iraqi government under Saddam Hussein. The other may have opposed the invasion, on account of knowing the same information but also having additional knowledge: Hussein's history of antipathy toward Islamic fundamentalism, UN weapons inspectors finding no WMD, and global public opinion disfavoring an invasion. If the supporter is affected by the curse of knowledge – implicitly assuming that everyone knows just what they themselves know (plus *unknown* unknowns) – what could possibly explain opposition to the war, other than opponents being careless about an existential threat at best, or "Saddam lovers" at worst? If the opponent is affected by the curse of knowledge, what could possibly explain support for the war — since their unique knowledge shows claims of WMD and al Qaeda links to be implausible — other than supporters being warmongering imperialists, motivated by the desire to control Iraq's oil?

If the curse of knowledge operates in political thinking, it may compound or exacerbate (affective) political polarization. As we implicitly assume that our political opponents know all of the facts that we know — knowledge which has helped shape our political opinions in the first place — then we may judge our opponents more harshly. That is, if we presume that others share the knowledge that has shaped our opinions, and made such opinions appear self-evidently correct to us, then our political opponents may take on a malevolent character. They are assumed to have all of the knowledge we had to arrive at the correct (our) conclusion, yet they persist in taking the wrong position; like Milton's Satan, it may seem that they have made evil their good.

This paper reports a series of studies testing whether the curse of knowledge is evident in political cognition. The results suggest that the curse of knowledge may be a contributing cause of political polarization, one of the heretofore overlooked psychological factors (Eibach, 2021) operating alongside institutional causes (Iyengar et al., 2019; Wilson et al., 2020).

1.1. Studies of the curse of knowledge

While the curse of knowledge (hereafter, CoK) has not yet been studied as it relates to and affects politics, it has been studied in a variety of other contexts. As discussed earlier, the curse of knowledge has been shown to apply in economic contexts, inhibiting marketplace actors' ability to profit from predicting the decisions of those who lack the same information (Camerer et al., 1989; Keysar et al., 1995; Loewenstein et al., 2006). The CoK can negatively affect lawyers, who may overestimate what jurors know about memory research relevant to eyewitness testimony, to the detriment of their clients (Terrell, 2014). It can impact criminal investigators and the accused alike, both of whom may overestimate what the other party knows about a crime (Granhag and Hartwig, 2008). It affects doctors, whose communications with patients can be made less effective by the CoK inflating doctors' estimates of patients' medically relevant knowledge

(Howard, 2019; Lourenco and Baird, 2020). It affects businesspeople, who may overestimate the level of knowledge widely held within a firm about that firm's organizational structure, impairing intra-firm coordination (Heath and Staudenmayer, 2000). Accountants and financial regulators may suffer CoK effects by overestimating knowledge relevant to predicting outcomes (Kennedy, 1995). Safety inspectors can suffer CoK effects, by assuming that site supervisors already know best practices they in fact do not (King, 2019). The CoK can also affect writers, making it more difficult to imagine their audience's ignorance of plot points they are intimately familiar with (Tobin, 2009), and impede communication in general by making ambiguous statements seem unambiguous to the speaker (Tobin, 2014).

The CoK is related to several other psychological biases, and was itself inspired by research on hindsight bias. Fischhoff (1975) demonstrated that we are influenced by outcome knowledge in our predictions of the likelihood of different outcome possibilities, both when we are placing ourselves in the shoes of our past ignorant selves, or the shoes of ignorant others. The CoK is also related to egocentrism; though whereas egocentrism is a difficulty in understanding perspectives other than one's own, the CoK is a specific difficulty in understanding a *less informed* perspective, not one that is better informed (Birch, 2005; Ghrear et al., 2016). Whereas egocentric bias weakens over development, with adults better able than children to inhibit their initial egocentric thinking (Epley et al., 2004), in some contexts adults exhibit greater CoK effects than children (Mitchell et al., 1996). The CoK also exhibits similarities to the false consensus effect, an overestimation of the extent to which others share our perspective on an issue (Spaulding, 2016), and pluralistic ignorance, an overestimation of the extent to which others do not share our cognition or behavior (Sargent and Newman, 2021).

Past research indicates that the CoK is persistent and difficult to eliminate. In economic contexts, monetary incentives and repeated iterations of predicting less-informed market participants' decisions reduced CoK effects, but only by half (Camerer et al., 1989). Higher education is associated with reduced CoK bias, but explicit instructions to focus attention on others' knowledge did not reduce CoK effects (Damen et al., 2018). (However, greater knowledge may actually worsen CoK effects, by hiding from one's view the areas in which one is, and others are, ignorant; Son and Kornell, 2010). Higher perspective-taking ability is also associated with reduced CoK effects, but instructing subjects to take another's perspective was not associated with lessening CoK bias (Ryskin and Brown-Schmidt, 2014; Damen et al., 2020).

We were unable to find any studies of the curse of knowledge as it relates to political cognition. Without existing research as a guide, one possibility is that the CoK has little or no effect on political thinking. Politics being an essentially allocentric domain, thinking about politics may involve greater focus on what others know and do not know, thereby overcoming the bias. Another possibility is that the CoK, by implicitly ascribing one's own knowledge to others, is a form of intellectual humility (Hannon, 2020). By reducing intellectual arrogance, this form of unconscious humility may tend toward reducing political polarization.

The possibility we thought most likely is that the CoK exacerbates affective political polarization, by masking the differences in knowledge that led to the formation of opposing opinions. The essence of the phenomenon – in Gomroki et al.'s (2023, p. 354) formulation, "When one interacts with others, one unknowingly imagines that others have the same intellectual background to understand the subject" – in the context of political disagreements, would seem to result in more negative appraisals of contrapartisans. This builds on the original definition of the CoK as an inability to inhibit one's own knowledge when imagining the thinking of others who do not share the same knowledge, shifting focus to its practical, real-world implication: that we act as if unconsciously assuming that others share the same information. This is similar to accounts of "naive realism," the result of psychological biases in which our inability to grasp that others have different knowledge informing their political opinions leads us to assume the worst about them (Ross and Ward, 1996; Friedman, 2020). Naive realism consists of the assumption that we see reality objectively, and our opinions about it are formed through an unbiased and unmediated apprehension of "the" facts. The naive realist assumes that others also see reality objectively, and will arrive at the same opinions as themselves. To explain why some people nonetheless disagree with their opinions, the naive realist has three explanatory options: the opponent may (1) be biased by ideology or self-interest, (2) be lazy, irrational, or unwilling to follow "the" evidence to its logical conclusions, or (3) not know the same information (Ross and Ward, 1996, 110–111). If the CoK affects political cognition, this third option is less likely to be taken under consideration; and the remaining options all place one's political opponents in a negative light.

People often attribute negative motives to others, committing the worst-motive fallacy (Reeder et al., 2005). (Hence Hanlon's razor: "never attribute to malice that which is adequately explained by stupidity" — in which "stupidity" should be replaced with "ignorance"). We expected that the CoK may contribute to political polarization, by obscuring the (highly likely) possibility that one's political opponents have arrived at contrary opinions because they do *not know* the same information that has shaped one's own opinion, and *do* know different information that has shaped their opinion. With this explanation occluded, and relevant information implicitly assumed to be universally shared, one's political opponents take on a malicious hue. For them to have arrived at an opposing opinion, after considering the same information, they must have opposing values, be "ideological" or biased by self-interest, or simply be lazy, unintelligent, or irrational. In other words, *because* people are imputing knowledge to people who do not have it, they may judge them more harshly. Lastly, if the CoK is part of the causal story behind political polarization, how might CoK effects be reduced in political contexts?

1.2. The present research

If the CoK exacerbates political polarization, the first place to look would be in the news media, the source of the basic informational building blocks that are used to form political opinions. Our first study asked partisans in Hong Kong and the U.S. about recent news stories, inquiring who was likely to know of the event or phenomenon described in the story. In this way, it lays out the direct or foundational evidence for the CoK applying in the political realm. Finding evidence of overestimating knowledge, our following studies measure what effect the CoK may have on affective polarization (and by investigating CoK effects on polarization

– from overestimating knowledge to more negative feelings toward opponents due to their overestimated knowledge – providing further, indirect evidence of the CoK operating in political cognition). In other words, the first study investigates whether partisans think opponents know about partisan news stories more than they do, while the subsequent studies measure how this over-imputation of knowledge may affect feelings toward political opponents, and how this could be debiased. Our second study investigated whether learning novel information about a political issue would lead to more negative attitudes toward one's opponents on that issue. Our third study sought to uncover whether partisans judge political opponents more favorably if they are told that they do not know the same issue-relevant facts. Finding no evidence that providing information on others' opinion-relevant knowledge or ignorance affects personal judgments, in Study 4 we asked participants to list their own most important political issue and three facts about it, and then to judge those who disagree with them on the issue — both in general, and separately for opponents who knew and did not know the factual information supporting the participants' position. This more direct method of focusing attention on knowledge gaps was associated with a moderation in judgments of less-informed others. In a final study, we attempted to debias potential CoK effects, testing a treatment condition comprising instructions to consider political opponents' lack of knowledge and how that may influence their opinion on the issue. Our overarching research question is: Does the curse of knowledge, the overestimation of knowledge shared in common, exacerbate political polarization, leading to harsher judgments of opponents (since "they should know better")? Table 1 presents the specific research questions.

In the following studies, we report sample size determination, data exclusion, measures, and manipulations where relevant. All data and research materials, including the surveys, are available on the OSF: https://osf.io/yc3tf/?view_only=525a070de6a74410aaa445f3f97cbbec. In addition, for both the sake of transparency and to inform future research, we wrote an appendix about the development of our studies and the lessons we learned, which can be found from the above link as well. Randomization for all experimental conditions was performed by Qualtrics, ethics approval was received from the relevant institution, and the studies' designs and analyses were not pre-registered. All U.S. samples were collected via Prolific among self-identified Republicans and Democrats. Besides over-representing political partisans by design, these samples contained fewer ethnic minorities, older people, men, and those with lower levels of educational attainment than the national average; median income was comparable to the nation median.

2. Study 1

During the Hong Kong protests of 2019, one of the authors realized what should have been apparent beforehand: having added people from both sides of Hong Kong's political divide ("yellow," or pro-democracy, and "blue," or pro-establishment) to his social media platforms, he began to notice that the two sets of partisans shared and commented on news stories covering entirely different events. To test whether partisans in Hong Kong were overestimating the extent to which news stories they found important were known outside of their partisan group, several questions were added to an unrelated study, and this formed the basis for a broader research proposal. This design was later adapted to the U.S. context in 2022, to test whether U.S. partisans overestimate knowledge of news stories important to their partisan group. We expected to find overestimation of knowledge shared across political divides: partisans claiming knowledge of their-side news stories at higher rates than contrapartisans, and partisans considering their-side stories to be "common knowledge" at higher rates than contrapartisans.

2.1. Method

2.1.1. Participants and design

In Hong Kong, participants were recruited via handing out flyers at pro-democracy and pro-establishment protests, for a total of 449 participants (239 women, 49 pro-establishment, M_{age} 30.59, SD_{age} 12.73). During this period, pro-establishment protests were less frequent and less attended than pro-democracy protests, resulting in the lower sample size for pro-establishment respondents. In the U.S., participants were recruited via Prolific among self-identified Republicans and Democrats, for a total of 201, due to uncertainty about whether the large effect sizes from the Hong Kong study would be found in a relatively less polarized context (98 women, 103 Democrats, M_{age} 41.23, SD_{age} 14.64). The studies were designed to present recent news stories in partisan media outlets on both sides, asking participants to identify whether they heard of the story, and whether they believed that co-partisans and/or contrapartisans had also heard of it.

2.1.2. Procedure and materials

In Hong Kong, participants responded to a longer survey on political opinions, with these questions about news stories included. In the U.S., the questions about news stories comprised the survey, plus demographic questions. In Hong Kong and the U.S., recent (late 2019 in Hong Kong, early 2022 in the U.S.) news stories were selected

TABLE 1 Studies in the current research and their respective research questions.

Study	Research questions
1	Do partisans overestimate the extent to which their political opponents know news stories/facts featured in their preferred media outlets?
2	If partisans gain new issue-relevant information – and are told that others are ignorant of it – do they nonetheless judge opponents on the issue more harshly?
3	If partisans are asked to judge one of two political opponents – the only difference between them being whether the opponent knows or does not know the same issue-related facts – does this information about a knowledge gap lead to less harsh judgments for those who do not share the same knowledge?
4	If partisans are asked to judge both political opponents who know and do not know the same issue-relevant facts, do they rate those who do not know the same facts less harshly? Is their rating of those who know the same facts closer to their rating of political opponents in general?
5	If partisans receive a simple political epistemology explanation of why people may disagree, do they judge political opponents less harshly?

from media outlets favored by the pro-democracy and pro-establishment, or Democratic and Republican, partisan groups, respectively. The Hong Kong stories were selected from among those popular on social media from established media outlets, and the U.S. stories were selected from transcripts of the popular Rachel Maddow (for Democratic stories) and Tucker Carlson (for Republican stories) shows, excluding stories that were covered on both programs. (As the most popular cable opinion shows for U.S. partisans at the time, we assumed that they would cover stories of particular interest to their partisan audiences, and that these stories would be covered by other outlets of the same partisan leaning.) An example of a "pro-establishment" story is "A police officer was burned by a Molotov cocktail thrown by protesters," and an example of a "pro-democracy" story is "A leader of the Junior Police Officers' Association used the word 'cockroaches' to describe protesters." An example of a Republican-media story is "In September [2021], Chicago experienced its deadliest month since 1992, reporting 89 homicides for the month," and an example of a Democrat-media story is "Interviews with former Trump administration staffers and associates revealed that the former president often violated the Presidential Records Act by destroying government documents."

2.1.3. Measures

Participants were presented with a sentence summarizing the news stories. They asked who they thought knew of the story, from "I do not recall ever hearing about this, or I do not think this happened" to three options starting with "I heard of it…" and ending in a progressively larger audience of others with the same knowledge: from neither partisan ingroup nor outgroup members ("…but I think most other people do not know about it") to only the partisan ingroup ("…but I do not think many [of the opposing party] know about it") to both partisan ingroup and outgroup members ("…and I think almost everyone knows it – it's common knowledge"). This provided story-aware participants epistemically sophisticated options (they heard of it, but most others may not have, or only co-partisans may have heard of it via their similar media diets), and an option representing the curse of knowledge (all others, including contrapartisans, assumed to share the individual's own knowledge). To minimize survey length, whether participants had *not* heard of the story or whether they believed it to be untrue were collapsed into the first option; the remaining options entailed knowledge of the story and belief that it was real.

2.2. Results and discussion

In Hong Kong, our pro-establishment respondents answered that they had heard of the pro-establishment stories 95.4% of the time, but our pro-democracy respondents answered that they had heard of the pro-establishment stories 71.1% of the time, $t(449) = 8.3$, $p < 0.001$, $d = 1.26$, 95% CI [0.96, 1.55]. Similarly, our pro-democracy respondents answered that they had heard of the pro-democracy stories 94% of the time, but our pro-establishment respondents answered that they had heard of the pro-democracy stories 73.2% of the time, $t(449) = 11.5$, $p < 0.001$, $d = -1.75$, 95% CI [−2.04, −1.45]. Meanwhile, pro-democracy respondents answered that the pro-democracy stories were "common knowledge" 71.3% of the time versus 38.8% for pro-establishment respondents, $t(449) = 8.3$, $p < 0.001$, $d = -1.25$, 95% CI [−1.55, −0.95], and pro-establishment respondents answered that the pro-establishment stories were "common knowledge" 69.1% of the time versus 29% for pro-democracy respondents, $t(449) = 10.6$, $p < 0.001$, $d = 1.60$, 95% CI [1.30, 1.89]. These contrast with the percentages selecting the more epistemically sophisticated answer ("I heard of it, and I'm pretty sure most people *on my side* have heard of it too"), which was selected 3.7% of the time by pro-democracy respondents for their-side stories, and 7.9% for pro-establishment respondents about their-side stories.

In the U.S., our Republican respondents answered that they had heard of the Republican-media stories 68.2% of the time, but our Democratic respondents answered that they had heard of the Republican-media stories only 38% of the time, $t(201) = 7.0$, $p < 0.001$, $d = 0.98$, 95% CI [0.70, 1.26]. Similarly, our Democratic respondents answered that they had heard of the Democratic-media stories 70.9% of the time, but our Republican respondents answered they had heard of the Democratic-media stories only 49.2% of the time, $t(201) = 5.5$, $p < 0.001$, $d = -0.77$, 95% CI [−1.05, −0.49]. Meanwhile, Democratic respondents answered that the Democratic-media stories were "common knowledge" 27.2% of the time versus 24.5% for Republican respondents, $t(201) = 0.8$, $p = 0.449$, $d = -0.11$, 95% CI [−0.38, 0.17], and Republican respondents answered that Republican-media stories were "common knowledge" 29.2% of the time versus 18.3% for Democratic respondents, $t(201) = 3.3$, $p = 0.001$, $d = 0.46$, 95% CI [0.18, 0.74]. These are similar percentages to the more epistemically sophisticated answer ("I heard about it, but I do not think many [of the opposing party] know about it"), which was selected 30.1% of the time by Democrats about Democratic-media stories, and 25.7% for Republicans about Republican-media stories.

The findings of Study 1 suggest that the CoK may be misleading some partisans to overestimate the extent to which the politically relevant information they know is widely shared. Our respondents in Hong Kong exhibited a greater degree of overestimation compared to our U.S. respondents, which may be an artifact of the particularly charged environment at the time. But in both contexts, either majorities or sizable minorities mistook their own knowledge of political news for common knowledge, when that knowledge was not actually shared in common. Democrats did not evince this overestimation, while Republicans did; but to a lesser degree than both groups of partisans in Hong Kong. Likewise, U.S. partisans were more likely to select the epistemically sophisticated option, acknowledging knowledge gaps between partisan groups – possibly the result of wider awareness of political polarization and media bias. However, simply overestimating the degree to which partisan knowledge is shared might not exacerbate polarization on its own. Contrariwise, the greater likelihood of selecting the epistemically sophisticated option in the U.S. might not reduce polarization, if such considerations do not come to mind in real-world contexts, without prompting. If the CoK contributes to polarization, such knowledge would be overestimated *and* political opponents would be judged more harshly on account of having this imputed knowledge, but persisting in their opposition regardless.

3. Study 2

Overestimating the extent to which politically relevant knowledge is widely shared would be of little consequence, if such overestimation did not result in harsher judgments of those whose actual knowledge leads them to take an opposing opinion. In this study, we explored whether receiving information about a new, fabricated political issue would lead toward harsher judgments of those disagreeing with the opinion such information would tend to support — despite being instructed that effectively no one else had been informed about it. We expected that treatment-group participants would make harsher judgments of opponents on the issue compared to those in the control group, overlooking the fact that their opponents had not received the same information.

3.1. Participants and design

Expecting a small effect size but without examples from the literature, we used G*Power to calculate the sample needed for a range of possible lower-end effect sizes; 1,048 participants were recruited via Prolific among self-identified Republicans and Democrats in the U.S.; 942 passed the attention check (a question testing factual memory of the treatment or control texts) and were included in the final analysis (487 women, 463 Republicans, M_{age} 37.6, SD_{age} 14.14).

3.2. Procedure and materials

Participants were randomly sorted into control and treatment groups. In the control, participants read a description of the executive branch of the federal government, focusing on the 15 federal executive departments. In the treatment group, participants read a fabricated announcement about a senior Department of Homeland Security official accused of accepting illegal bribes by a DHS whistleblower, who had just shared this accusation alongside incriminating evidence on the OpenSecrets website. The announcement noted further that the web page listing the accusation had received under 200 "hits" or visitors since it went public, and no media outlets had yet reported on the story, hence "it is safe to say that almost no one (beside you) has heard about it yet." To ensure that participants read and understood the materials, they were given a multiple-choice question about the content, and were asked to briefly explain the reasons for their rating.

3.3. Measures

Participants were given an 11-point feeling thermometer to rate their feelings toward "those Americans who think that federal prosecutors should *not* focus more effort on investigating possible corruption in government agencies." (Please see Appendix A for a discussion on the feeling thermometer, how it seemed to sometimes be misinterpreted, and what we did to clarify the interpretation of it.) A 0 represented "how you feel about your worst enemy," and a 10 represented "how you feel toward the person you love most in the world."

3.4. Results and discussion

We expected that participants in the treatment group would fail to account for the ignorance of those who might not see a need for federal prosecutors to divert their attention away from other concerns toward corruption in federal agencies, and judge people holding this opinion more harshly than those in the control group. We found that treatment-group participants did judge opponents on this issue more harshly ($M = 2.27$; $\text{SD} = 2.03$) than those in the control group ($M = 2.53$; $\text{SD} = 1.93$), $t(942) = 2.0$, $p = 0.045$, $d = 0.13$, 95% CI [0.01, 0.26].

This was a small difference, as would be expected in our theoretical model of how the CoK exacerbates polarization: We provided only a small piece of information at one point in time, whereas politically engaged partisans absorb large amounts of information over their lifetimes. As more information is learned cumulatively, the CoK would attribute more information to others who have not actually acquired it, making opposition to the opinions such information supports harder to explain other than by invidious motives.

However, this result might also be explained as an effect of priming: that treatment-group participants were primed to think of government corruption in general, and with this problem at the forefront of their minds, made less charitable judgments of those who disagreed that prosecutors should focus more on rooting out government corruption. In our next study, we investigated whether by focusing attention on what political opponents *do* and *do not* know about an issue, judgments of less knowledgeable opponents would be moderated.

4. Study 3

The CoK may contribute to polarization by obscuring differences in knowledge that resulted in differences of opinion. In this experiment, we tested whether judgments of political opponents would be moderated for those opponents who were described as being ignorant of the issue-relevant knowledge participants knew, compared to opponents who were described as sharing the same issue-relevant knowledge. We expected participants to judge political opponents less negatively if they were informed that they do *not* share the same knowledge of the issue — *if* this information overcomes potential CoK effects, and is interpreted to suggest that the opponent's opinion may have been produced by the absence of issue-relevant knowledge known to participants — compared with participants who were informed that an opponent *did* share the same issue-relevant knowledge.

4.1. Participants and design

Without effect sizes from existing research, we tentatively expected a small effect size, as prior research has demonstrated the CoK to be robust against instructions to consider others' knowledge or take others' perspectives. We recruited 600 self-identified Republicans and Democrats in the U.S. via Prolific (295 women, 238 Republicans; M_{age} 39.21, SD_{age} 14.12). For a two-tailed *t*-test of mean differences with, this sample would have an 80% chance of finding a true effect of slightly over 0.2; but it would be underpowered to detect smaller effect sizes.

4.2. Procedure and materials

Participants were asked to name a political issue important to them, and to provide three facts they knew about the issue that support their opinion on it. Then they were instructed that text-mining software would search through notes from a previous interview-based study and match them with an interviewee to rate. Participants were randomly selected into two conditions. In both conditions, participants were presented with excerpts from interview notes; in the ignorant condition, the interview notes did not mention any of the facts the participant provided, and in the knowledgeable condition, the interview notes indicated that the interviewee did know the facts the participant wrote about. In both conditions, participants were informed that the interviewee expressed opposition to the opinion expressed by the participant.

4.3. Measures

Participants were given a 10-point scale with happy to angry faces as graphic references.

4.4. Results and discussion

No difference was found between the ratings of the interviewee who knew the same facts ($M = 4.53$; SD = 2.35) and one who did not know the same facts ($M = 4.48$; SD = 2.00), $t(585) = 0.275$, $p = 0.784$, $d = 0.02$, 95% CI [−0.14, 0.19]. Excluding participants whose stated issues and facts were independently judged as indicating inattention or misunderstanding by both authors did not affect results. This null result is consistent with the explanation that the CoK may *not* inflame polarization by harshening judgments of others via wrongly assuming them to know the same issue-relevant information. However, it is also consistent with previous research, which has established the robustness of CoK effects in the face of instructions to consider others' knowledge and to take another's perspective (Ryskin and Brown-Schmidt, 2014; Damen et al., 2018, 2020); here too, providing only evidence of what another knows and does not know did not affect judgments. Our next study sought to distinguish between these two explanations.

5. Study 4

Study 3 randomly provided either an example of a political opponent who knew, or did not know, the same issue-related facts as participants. In this study, we made knowledge gaps more visible by instructing participants to separately judge those who did and did not know the same issue-related facts. In this way, we hypothesized that if the CoK were harshening judgments of political opponents by occluding epistemology, participants asked to separately rate knowledgeable and ignorant opponents would be forced to grapple with political epistemology, considering how knowledge gaps might affect the development of an opposing opinion – and would judge less-informed opponents less harshly. We furthermore expected that ratings of knowledgeable opponents would be closer to the initial rating of all opponents, compared to ratings of ignorant opponents. That is, we expected participants to judge opponents who are *ignorant* of issue-relevant knowledge less harshly than opponents who were knowledgeable; and that ratings of knowledgeable opponents would be closer to ratings of opponents in general, evincing CoK bias. Alternatively, if the CoK were not influencing political judgments according to our theoretical expectations, ratings of more knowledgeable opponents might be the same or higher than ignorant opponents, owing simply to the positive quality of being knowledgeable.

5.1. Method

5.1.1. Participants and design

Self-identified Republicans and Democrats in the U.S. ($N = 152$; without an expected effect size, funding limitations necessitated a small sample) were recruited through Prolific (70 women, 66 Republicans, M_{age} 35.34, SD_{age} 14.71).

5.1.2. Procedure and materials

Participants were asked to name a political issue important to them and to list three relevant facts that back up their opinion on the issue. An example issue was provided: whether to create a new state park, along with three example facts that support a favorable opinion on the issue. After naming their issue and writing down three related facts, participants were asked to rate how they feel about people who disagree with them on this issue. In the next step, they were asked to separate those who disagree into two groups — first, opponents who know the facts they listed, and then those who do not — and to separately rate how they feel towards these two groups. Both authors independently examined the provided facts to verify good-faith effort and understanding of the instructions. Any differences in the coding were discussed and resolved. Excluding participants who failed these checks did not affect results, so all data are reported below.

5.1.3. Measures

Participants were given 11-point scales with happy to angry faces as graphic references, with higher ratings indicating harsher judgments.

5.2. Results and discussion

Participants rated those who shared the same issue-relevant knowledge yet disagreed with their opinion ($M = 7.92$; SD = 2.40) significantly more negatively than those who disagreed with them but were unaware of the same facts ($M = 5.72$; SD = 2.21), $t(151) = 10.846$, $p < 0.001$, $d = 0.88$, 95% CI [0.69, 1.07]. When participants were directed to separately consider their feelings about those who *do* and *do not* know the same issue-related facts, they were more forgiving of opponents who lacked the knowledge participants deemed important to understanding the issue. The rating difference between all opponents and opponents who lacked the same knowledge ($M = 1.82$; SD = 2.55) was greater than the rating difference between all opponents and those who knew the same facts ($M = -0.38$; SD = 2.03), $t(151) = -10.846$, $p < 0.001$, $d = -0.88$, 95% CI [−1.07, −0.69]. This indicates that when people think of political opponents in general, they judge them in much the same way as they judge opponents who know what they know about an issue — a curse of knowledge effect (i.e., imputing one's own knowledge to all others). When thinking separately about opponents who do not share the same knowledge, instead of punishing them for their ignorance, they were judged more charitably: opponents' ignorance of the knowledge that supports one's opinion was treated as a mitigating factor. The final study uses an experimental design to look for CoK effects by testing an attempt to debias the curse of knowledge.

5.3. Study 5

Study 3 found that providing information on what political opponents know or do not know about an issue did not affect judgments. But Study 4 found that by focusing attention solely on the difference between opponents who know the same facts as the participant — the facts that shaped their position on the issue — and opponents who were ignorant of those facts, participants rated the ignorant more favorably, and opponents who knew the same facts more harshly. To look for clear evidence of a CoK effect in political judgments of others, we designed a final experimental study testing an attempt to debias the CoK. If the CoK were influencing affective polarization – and such polarization were not exclusively caused by other factors – participants receiving a simple political epistemology explanation of why people may disagree should make more moderate judgments. We expected that judgments of political opponents who lacked participants' issue-relevant knowledge would be moderated by an epistemological treatment instructing participants to consider how this ignorance may influence an opponent's opinion. Alternatively, if the CoK were not negatively influencing judgments of political opponents, this attempt to debias a nonexistent influence should have no effect on judgments.

5.4. Participants and design

Tentatively expecting a mid-range effect size, 200 participants were recruited via Prolific among self-identified Republicans and Democrats in the U.S. (100 women, 102 Republicans; M_{age} 38.11, SD_{age} 14.75). This sample would have 0.95 power to detect a true effect size of at least 0.4 in a two-tailed *t*-test.

5.5. Procedure and materials

As in Studies 3 and 4, participants were asked to name a political issue of importance, and list three facts supporting their position on the issue, with the same example provided. They were instructed that text-mining software would search through notes from a previous interview-based study and match them with an interviewee to rate. Participants were randomly selected into treatment and control conditions. In the control, participants were presented with excerpts from interview notes, presenting "Jessica" as "very knowledgeable" in general, but "when we asked Jessica about <participant's issue>, she did not seem to know as much about this issue as the other issues we discussed; she explained that this is an issue she has not yet learned much about." The specific facts participants had written were presented, alongside a low "text-mining similarity score" of 5% indicating that "Jessica *does not know* the same facts that support your opinion, and she takes the opposite position on this issue." The debiasing treatment condition was the same, except this information was followed by an explanation that political disagreements are sometimes caused by a lack of knowledge held in common — since what we know and do not know about an issue influences the opinion we develop — and other times by different value judgments. The example provided pre-treatment was then used to illustrate how sometimes learning more about an issue may change one's opinion, but such new knowledge might also leave one's opinion unchanged if it is rooted in conflicting values or beliefs, or differing interpretations of the same information.

5.6. Measures

Participants were asked to rate "Jessica" on a feeling thermometer, from 1 "Strongly dislike" to 10 "Strongly like."

5.7. Results and discussion

We expected that the debiasing treatment would moderate judgments by reducing CoK effects, and found that participants in the debiasing treatment rated their political opponent more favorably ($M = 5.22$; SD = 2.08) than those in the control ($M = 4.40$; SD = 1.89), $t(200) = 2.9$, $p = 0.004$, $d = 0.42$, 95% CI [0.14, 0.69]. There was no significant difference in the results when eliminating validity check failures, so results from the full sample are reported.

If the CoK were not negatively influencing judgments of political opponents, the treatment focusing attention on the epistemology of political disagreement should have made little difference. Negative judgments based on a "know *or should know*" standard, not a CoK overestimation of knowledge, would unlikely be affected by this treatment. But here, as in Study 4, focusing participants' attention on the role of knowledge and ignorance in the formation of political opinions – alongside the alternative possibility that differences in values and beliefs may make knowledge gaps irrelevant – resulted in more favorable, less harsh judgments of a political opponent. This provides additional evidence that the CoK, by occluding epistemology and exaggerating similarities in knowledge, makes an independent contribution to political polarization.

6. General discussion

If the curse of knowledge affects political cognition, one likely effect is exacerbating political polarization. Partisans would make the CoK error of unconsciously assuming that their political opponents know the same information that they themselves have learned, and which led them to form the opinion rejected by their opponent. With the possibility occluded that one's opponent has not learned the key information that led to the formation of one's own opinion, how is one to explain the opponent's position? Ignorance aside, the remaining options — laziness, self-interest, ideological bias, malice — all paint the opponent in a negative light. This is less likely to occur regarding casual or ambivalent opinions, without much personal investment or about which the partisan merely leans to one of several known, well-supported sides. But for strongly held opinions, where "the other side" seems self-evidently wrong or immoral given what the partisan knows (and does not know), the CoK would tend to make opposing opinions unfathomable – except as motivated by discreditable intentions.

To investigate, Study 1 first collected evidence that partisans overestimate the extent to which news stories in their preferred media outlets were also known by their political opponents. This is a clear CoK effect, related to pluralistic ignorance and the false consensus effect, but on its own might not contribute to political polarization. If the CoK does tend toward worsening political polarization, partisans would gather information leading them to form an opinion on an issue, unthinkingly assume that such information is universally held, and then, blind to the fact that others might not have learned the same information, judge those with an opposing opinion more harshly. Study 2 provided evidence of this process: participants were given unique information about an issue, formed an opinion on it, and despite being told that most people have not learned the same information, tended to judge those with a differing opinion on that issue more harshly. Study 3 used a more ecologically valid design, asking participants to make judgments of someone they met online with an opposing opinion on an issue. With only this person's level of knowledge about the issue experimentally manipulated, most participants did not evince counter-CoK thinking. For example, they did not take the other's ignorance and knowledge into account and temper their judgment with the charitable interpretation that what the other does not know might prevent her from forming the same opinion. As in Damen et al.'s (2018) study, this indirect attempt to get participants to focus on another's knowledge did not succeed. However, these results are also consistent with an absence of CoK effects in judging political opponents. Contrariwise, assuming that the information provided about opponents' ignorance did reduce CoK bias but the true effect size was small, our study was underpowered to detect it.

Study 4 provided a more direct intervention, asking participants first to judge those who oppose them on an issue of personal importance, and then asking them to rate separately those who know the same factual information relevant to the issue, and those who are ignorant of such information. With political epistemology brought to the fore of their minds by separating opponents into those who do and do not share the participant's issue-relevant knowledge, participants made more charitable judgments of their opponents who were unaware of the information shaping participants' opinions. Meanwhile, their judgments of opponents who shared the same information were as hostile as the judgments they made before considering the role of knowledge gaps. In Study 5, a similarly externally valid setting as Study 3 was used: again, making judgments of strangers online based on their political opinions. While past attempts to mute CoK effects have proven largely ineffective, this attempt was at least partially successful. When instructed to consider the basics of political epistemology — that opinions are formed on the basis of the information one has acquired, plus values and beliefs which affect the interpretation of that evidence, such that some may arrive at opposing opinions simply because they do not know the same information — participants judged a political opponent, on a self-selected issue of personal importance, less harshly. Taken together, these studies indicate that the curse of knowledge is one of several psychological contributors to political polarization, and that engaging in epistemological thinking may reduce its effects.

6.1. Limitations

The limitations of these studies include recruiting using an online platform, which limits the sample to those with access to the internet and basic computer literacy. The samples included only political partisans, and were non-representative on several demographic categories; representative samples may reveal differences between demographic groups. We took ecological validity into account in designing our studies, e.g., by presenting a person who they might meet online. However, the example person's characteristics could influence the results; providing a representative array of example people would have remedied this problem but was not feasible. Another limitation lies in depending on self-reported awareness of different news stories. Varying degrees of social desirability bias and humility in admitting what one does not know could have influenced the results. Studies 4 and 5 were limited by relatively small sample sizes. Our manipulation checks may have eliminated participants who were paying attention to the experimental materials, but whose attention lapsed only during the attention-check question, or whose written answers were incorrectly judged as indicating misunderstanding or inattention.

With regard to the possible confounding effect of priming in Study 2, future research could present the control with an old news story about the same treatment topic, i.e., an example of corruption, that received sufficiently ample media coverage as to be nearly universally known. In this way, the topic would be salient in both the treatment and control conditions, eliminating the potential priming effect. Another possibility would be to give the control condition the same story but tell the control group that everyone has heard of it, or omit information about who knows it. However, this design has interpretation difficulties: if the experimental group judged opponents less harshly than this control, it could be that the instructions alerting participants to the lack of media attention (only the participant is likely to have heard of it) debiased default CoK effects. At the same time, if there were no difference in ratings between the two groups, this could be the result of the experimental group instructions being overwhelmed by the CoK bias, as has occurred in prior research. In other words, experimental-group participants could have defaulted to CoK over-imputation of knowledge to others, making their ratings equivalent to those in the control group. Another possibility would be to use a three-group design: (1) the original experimental treatment, (2) a condition told that widespread media coverage means that nearly everyone has heard of it, and (3) a control with no information about who has heard of the allegations. All else being equal, those in the original experimental treatment would be expected to have the least harsh judgments of opponents, as their instructions should at least somewhat reduce CoK effects by focusing attention on widespread ignorance of the story. No difference would be expected between the group told that everyone has heard of it and the control group in which no information about others' knowledge was provided. A manipulation check would be needed to ensure that those in the group which was told that essentially everyone had heard of the story, actually believed that there was widespread media coverage sufficient to ensure that effectively everyone would know of it.

For greater ecological validity, studies building on the successful debiasing procedure in Study 5 should try introducing participants to others whose level of knowledge is *not* stated, to explore how depolarization efforts can be best designed for most real-world situations in which political opponents' knowledge and ignorance is unknown. Furthermore, our U.S. story selections relied on editorial decisions made by producers at the most popular cable opinion shows; a better method of selecting those stories of greatest interest to contrasting partisans may be to exploit engagement data from social media companies, where available. Future research could also try to exploit any existing measures of how often a news story is covered by media sources on one side versus the other. Lastly, the difference in the salience of polarized political debates during the 2019 Hong Kong protests and the U.S. of early 2022 may be a contributing cause of the lesser overestimation of contrapartisan knowledge in our U.S. data. Collecting similar data during a presidential campaign season in the U.S. may result in more similar levels of overestimation. Alternatively, the greater degree of collectivism in Hong Kong compared to the U.S. may have affected the likelihood of respondents to select the "common knowledge" response.

6.2. Theoretical implications

This study is the first to demonstrate that politics is another domain in which the CoK affects cognition, and provides evidence that its effect is to exacerbate political polarization. Political polarization has been increasing over recent decades, but psychological biases, the CoK included, have likely remained unchanged over the period. A constant being unable to explain a variable, clearly the CoK cannot be *the* cause of increasing political polarization in many countries. Rather, the CoK is likely an adjunct or accelerant to the central causes of increasing political polarization.

For instance, changes in the U.S. media system are a central cause of political polarization there (Prior, 2007). Before the rise of cable and then the internet, broadcast television news was more widely watched and influential; and to attract the largest possible audience, political news tended to be presented in a down-the-middle, nonpartisan manner. The introduction of cable television vastly expanded the number of options for viewers, and helped create a niche for news channels with a decidedly partisan bent. Buoyed by the market success of partisan news outlets, and with social media algorithms facilitating ideologically homogeneous communication networks, the U.S. media system became more populated with content designed to appeal to opposing partisan groups (Taibbi, 2020). With separate media diets providing contrasting perspectives on political issues, as well as covering different stories entirely (e.g., Radtke, 2017), not only were partisans entitled to their own opinions — they were presented with their own sets of facts. As partisan groups accumulate differing sets of politically relevant knowledge, they become more susceptible to CoK effects. For instance, Republicans absorbing copious information from their partisan media diets about problems attributed to immigration would wonder why Democrats seem unconcerned about a problem they have learned has caused tremendous suffering to U.S. citizens. Blinded by the CoK to the explanation that Democrats' media diets do not include so many stories about victims of immigrant criminals and public services overwhelmed by newly arrived migrants, other explanations must be found (e.g., "Democrats tend to be more privileged, do not face these problems in their own lives, and so do not care about working class Americans who have to live where such problems are most acute"). Democrats absorbing information from their partisan media diets about the existential threat of climate change would wonder why Republicans seem so unconcerned about it. Blinded by the CoK to the explanation that Republicans' media diets do not include so many stories explaining the danger of climate change or linking destructive weather events to it, other explanations must be found (e.g., "Republicans are anti-science, and they care more for oil companies than life on earth").

Whereas in economic contexts the CoK may produce socially beneficial effects by making information asymmetries more difficult to exploit, in political contexts, as in many others, the CoK is more likely to contribute to social harms, like an increasingly polarized society.

6.3. Practical implications

If the curse of knowledge is merely an accelerant or partial cause of political polarization, then muting its effects is unlikely to solve the problem entirely, but it would ameliorate it. The results of Study 5 suggest that thinking about political epistemology, if not eliminating the CoK, may reduce its negative effects on judging one's political opponents. Political epistemology, or how people come to know or believe what they know or believe about politics, involves many factors: what one learned from one's parents, peers, teachers, media diet, life experiences, and other sources of politically relevant information, along with psychological traits that draw one toward some ideas and away from others (Beattie, 2019). By drawing attention to the fundamental arbitrariness of the process by which we accumulated some knowledge and not other knowledge, and realizing that what we have and have not learned affected the development of our political opinions, we may be humbled (if not humiliated). But so too are our political opponents: their opinions were also formed through a fundamentally arbitrary process of learning some things we likely have not, and not learning other things we have. Focusing the attention of partisan disputants on knowledge gaps between them may make attributions to malice less likely, and attributions to ignorance more likely. And if the apparent solution to an opponent motivated by malice is combat, the solution to opposition rooted in ignorance should be dialogue.

In commercialized media systems, where media outlets compete for advertisers and subscribers, educating audiences in political epistemology is unlikely to occur unless such efforts result in greater revenues. If audiences reward such efforts, they would likely spread across the media system; but if audiences prefer partisan animosity from their media diets, political epistemology is unlikely to be featured. However, educators could teach basic political epistemology alongside media literacy in schools. Students would be taught to critically analyze the news media, considering (among others) potential sources of bias and the adequacy of evidence provided to support an argument or explanation, and also to think about how what they and others learn (and do not learn) influences opinion formation. For such educational interventions to succeed at reducing polarization — or at least the portion of polarization produced by the curse of knowledge — additional research is needed.

7. Conclusion

The present studies extend research on the curse of knowledge to the domain of political cognition, by demonstrating that overestimating political opponents' knowledge is linked to more negative appraisals. When partisans commit the CoK error of assuming that political opponents share the same knowledge as themselves, opponents take on the malevolent character of one who knows why a differing opinion is correct, yet persists in opposing it.

This theoretical understanding led us to develop and successfully test an intervention to debias CoK effects: prompting partisans to think like political epistemologists. By engaging in thinking about how differences in knowledge affect opinion formation, partisans may find their opponents less implacable, and their character less that of an enemy and more that of one who could be made an ally through dialogue. Indeed, the opponent must always be explained; but if the *first* explanation that we look for is that he sees a different set of facts, political polarization may be reduced.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/yc3tf/?view_only=525a070de6a74410aaa445f3f97cbbec.

Ethics statement

The studies involving human participants were reviewed and approved by Survey and Behavioural Research Ethics, CUHK. The patients/participants provided their written informed consent to participate in this study.

Author contributions

PB did the original research in Hong Kong and the initial literature review. PB and MB collaborated on theoretical development, research design, and data analysis and wrote and revised the manuscript. MB performed the additional literature review. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by a generous grant from CUHK's MGPE Research Fund.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1200627/full#supplementary-material

References

Beattie, P. (2019). *Social evolution, political psychology, and the media in democracy: the invisible hand in the U.S. marketplace of ideas.* Cham, Switzerland: Palgrave Macmillan.

Birch, S. A. (2005). When knowledge is a curse: Children's and adults' reasoning about mental states. *Curr. Dir. Psychol. Sci.* 14, 25–29. doi: 10.1111/j.0963-7214.2005.00328.x

Birch, S. A., and Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychol. Sci.* 18, 382–386. doi: 10.1111/j.1467-9280.2007.01909.x

Birch, S. A., Brosseau-Liard, P. E., Haddock, T., and 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. doi: 10.1016/j.cognition.2017.04.015

Camerer, C., Loewenstein, G., and Weber, M. (1989). The curse of knowledge in economic settings: an experimental analysis. *J. Polit. Econ.* 97, 1232–1254. doi: 10.1086/261651

Damen, D., van der Wijst, P., van Amelsvoort, M., and Krahmer, E. (2018). The curse of knowing: the influence of explicit perspective-awareness instructions on perceivers' perspective-taking. In T. T. Rogers, M. Rau, X. Zhu and C. W. Kalish (Eds.), Proceedings of the 40th annual conference of the cognitive science society, 1578–1583. Cognitive Science Society

Damen, D., van der Wijst, P., van Amelsvoort, M., and Krahmer, E. (2020). Can the curse of knowing be lifted? The influence of explicit perspective-focus instructions on readers' perspective-taking. *J. Exp. Psychol. Learn. Mem. Cogn.* 46, 1407–1423. doi: 10.1037/xlm0000830

Dębska, A., and Komorowska, K. (2013). Limitations in reasoning about false beliefs in adults: the effect of priming or the curse of knowledge? *Psychol. Lang. Commun.* 17, 269–278. doi: 10.2478/plc-2013-0017

Drayton, L. A., and Santos, L. R. (2018). What do monkeys know about others' knowledge? *Cognition* 170, 201–208. doi: 10.1016/j.cognition.2017.10.004

Eibach, R. (2021). Ideological Polarization and Social Psychology. *Oxford Research Encyclopedia of Psychology*. Available at: https://oxfordre.com/psychology/view/10.1093/acrefore/9780190236557.001.0001/acrefore-9780190236557-e-240 (Retrieved June 24, 2023).

Epley, N., Morewedge, C. K., and Keysar, B. (2004). Perspective taking in children and adults: equivalent egocentrism but differential correction. *J. Exp. Soc. Psychol.* 40, 760–768. doi: 10.1016/j.jesp.2004.02.002

Fischhoff, B. (1975). Hindsight is not equal to foresight: the effect of outcome knowledge on judgment under uncertainty. *J. Exp. Psychol. Hum. Percept. Perform.* 12, 288–299. doi: 10.1136/qhc.12.4.304

Fisher, M., Goddu, M. K., and Keil, F. C. (2015). Searching for explanations: how the internet inflates estimates of internal knowledge. *J. Exp. Psychol. Gen.* 144, 674–687. doi: 10.1037/xge0000070

Friedman, J. (2020). *Power without knowledge: a critique of technocracy*. Oxford: Oxford University Press, *18*, 937–938

Ghrear, S. E., Birch, S. A., and Bernstein, D. M. (2016). Outcome knowledge and false belief. *Front. Psychol.* 7:118. doi: 10.3389/fpsyg.2016.00118

Gomroki, G., Behzadi, H., Fattahi, R., and Salehi Fadardi, J. (2023). Identifying effective cognitive biases in information retrieval. *J. Inf. Sci.* 49, 348–358. doi: 10.1177/01655515211001777

Granhag, P. A., and Hartwig, M. (2008). A new theoretical perspective on deception detection: on the psychology of instrumental mind-reading. *Psychol. Crime Law* 14, 189–200. doi: 10.1080/10683160701645181

Hannon, M. (2020). "Intellectual humility and the curse of knowledge" in *Polarisation, arrogance, and dogmatism*. eds. A. Tanesini and M. Lynch (Oxfordshire: Routledge), 104–119.

Heath, C., and Staudenmayer, N. (2000). Coordination neglect: how lay theories of organizing complicate coordination in organizations. *Res. Organ. Behav.* 22, 153–191. doi: 10.1016/S0191-3085(00)22005-4

Howard, J. (2019). "Curse of knowledge," in *Cognitive errors and diagnostic mistakes* (Cham: Springer), 89–91.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annu. Rev. Polit. Sci.* 22, 129–146. doi: 10.1146/annurev-polisci-051117-073034

Kennedy, J. (1995). Debiasing the curse of knowledge in audit judgment. *Account. Rev.* 70, 249–273.

Keysar, B., Ginzel, L. E., and Bazerman, M. H. (1995). States of affairs and states of mind: the effect of knowledge of beliefs. *Organ. Behav. Hum. Decis. Process.* 64, 283–293. doi: 10.1006/obhd.1995.1106

King, T. (2019). The curse of knowledge. *Prof. Saf.* 64:61.

Loewenstein, G., Moore, D. A., and Weber, R. A. (2006). Misperceiving the value of information in predicting the performance of others. *Exp. Econ.* 9, 281–295. doi: 10.1007/s10683-006-9128-y

Lourenco, A. P., and Baird, G. L. (2020). Optimizing radiology reports for patients and referring physicians: mitigating the curse of knowledge. *Acad. Radiol.* 27, 436–439. doi: 10.1016/j.acra.2019.03.026

Mitchell, P., Robinson, E. J., Isaacs, J. E., and Nye, R. M. (1996). Contamination in reasoning about false belief: an instance of realist bias in adults but not children. *Cognition* 59, 1–21. doi: 10.1016/0010-0277(95)00683-4

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: imputing one's own knowledge to others. *Psychol Bull* 125, 737–759. doi: 10.1037/0033-2909.125.6.737

Prior, M. (2007). *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge: Cambridge University Press.

Rabb, N., Fernbach, P. M., and Sloman, S. A. (2019). Individual representation in a community of knowledge. *Trends Cogn. Sci.* 23, 891–902. doi: 10.1016/j.tics.2019.07.011

Radtke, D. (2017) Study: Fox news covered immigration five times as much as CNN and MSNBC combined. Media Matters for America. Available at: https://www.mediamatters.org/breitbart-news/study-fox-news-covered-immigration-five-times-much-cnn-and-msnbc-combined (Accessed October 27, 2022).

Reeder, G. D., Pryor, J. B., Wohl, M. J., and Griswell, M. L. (2005). On attributing negative motives to others who disagree with our opinions. *Personal. Soc. Psychol. Bull.* 31, 1498–1510. doi: 10.1177/0146167205277093

Ross, L., and Ward, A. (1996). "Naive realism in everyday life: implications for social conflict and misunderstanding," in *Values and knowledge*. eds. E. S. Reed, E. Turiel and T. Brown (Mahwah, NJ: Lawrence Erlbaum Associates), 103–135.

Ryskin, R. A., and Brown-Schmidt, S. (2014). Do adults show a curse of knowledge in false-belief reasoning? A robust estimate of the true effect size. *PLoS One* 9:e92406. doi: 10.1371/journal.pone.0092406

Sargent, R. H., and Newman, L. S. (2021). Pluralistic ignorance research in psychology: a scoping review of topic and method variation and directions for future research. *Rev. Gen. Psychol.* 25, 163–184. doi: 10.1177/1089268021995168

Sloman, S. A., and Rabb, N. (2016). Your understanding is my understanding: evidence for a community of knowledge. *Psychol. Sci.* 27, 1451–1460. doi: 10.1177/0956797616662271

Son, L. K., and Kornell, N. (2010). The virtues of ignorance. *Behav. Process.* 83, 207–212. doi: 10.1016/j.beproc.2009.12.005

Spaulding, S. (2016). Mind misreading. *Philos Issue* 26, 422–440. doi: 10.1111/phis.12070

Taibbi, M. (2020). *Hate Inc: Why Today's media makes us despise one another* New York: OR Books.

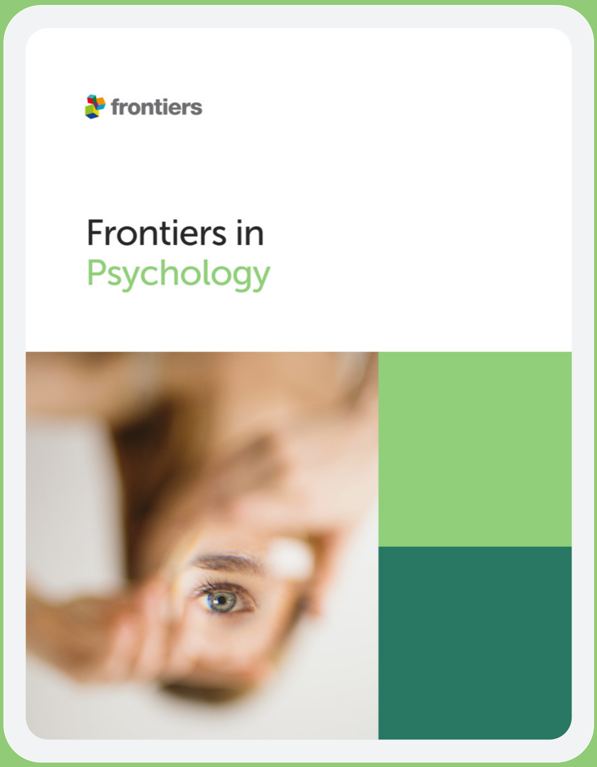
Terrell, J. T. (2014). The curse of knowledge in estimating jurors' understanding of memory: attorneys know more about memory than the general population. *Appl Psychol Crim Just* 10, 98–105.

Tobin, V. (2009). Cognitive bias and the poetics of surprise. *Lang. Lit.* 18, 155–172. doi: 10.1177/0963947009105342

Tobin, V. (2014). "Where do cognitive biases fit into cognitive linguistics? An example from the curse of knowledge," in *Language and the Creative Mind*. eds. B. Dancygier, M. Borkent and J. Hinnell (Stanford: CSLI Publications), 347–363.

Wilson, E. O. (2012). *The social conquest of earth*. New York: WW Norton & Company.

Wilson, A. E., Parker, V. A., and Feinberg, M. (2020). Polarization in the contemporary political and media landscape. *Curr. Opin. Behav. Sci.* 34, 223–228. doi: 10.1016/j.cobeha.2020.07.005



Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us
+41 (0)21 510 17 00
frontiersin.org/about/contact