

Dispelling illusions of truth: Exploring the factors that lead to inflated truth
judgements

Emma Louise HENDERSON

Faculty of Business and Social Sciences

Kingston University London

A thesis presented in partial fulfilment of the requirements of Kingston University

for the degree of

Doctor of Philosophy

June 2021

Abstract

Judging the truth of incoming information is one of the most challenging and important tasks that people face every day. How do people decide what is true and what is not? When constructing truth judgements, people use both declarative information and the subtler cues that accompany information processing. These subtle, non-content-based cues that make information feel truer are termed “truth effects”. This thesis uses trivia statements to investigate the robustness of two such non-probative truth effects driven by repetition (the *illusory truth effect*) and concrete language (the *linguistic concreteness effect*). Neither concreteness nor repetition provide substantive evidence, yet people believe repeated statements more than new ones, and concretely worded statements feel truer than their abstract counterparts. Truth effects can have direct implications in our digital world, where information may be spurious, and communicators can enlist subtle cues to persuade the addressee without detection.

Throughout the thesis I apply open methods that have the potential to increase the quality, replicability, and transparency of research. In Chapter 2, I set out to replicate and extend the linguistic concreteness effect. Across two experiments I did not observe an effect larger than the smallest effect size of interest. Therefore the remainder of the thesis focuses on the illusory truth effect. Chapter 3 uses systematic mapping to synthesise and catalogue the entire illusory truth literature in terms of methods, findings, and transparency. The results reveal a lack of standardisation in the methodology employed, and of transparency in reporting. I also find that greater diversity of stimuli and participants is required for generalisability. In Chapter 4, my final study used a longitudinal design to test whether the delay between repetitions moderates illusory truth. Contrary to previous claims, I find that across four intervals

(immediately, one day, one week, one month) the effect diminishes as delay increases. This thesis contributes to knowledge by providing an overview of the current state of truth effects research. It demonstrates that there is considerable cause to doubt the existence of a linguistic concreteness effect, and by implication, there is reason to be sceptical about other truth effects based on subtle manipulations. In contrast, this thesis establishes confidence that the illusory truth effect is robust but reduces with time. This finding has implications for the mechanisms thought to underlie truth effects. Overall, the results suggest that when truth effects research uses rigorous, transparent, and unbiased methods, it paints a different picture from that of the existing literature.

Acknowledgements

I am enormously thankful to a great number of wonderful people for their help and support throughout my PhD.

First, I would like to express my gratitude to my two supervisors, Fred Vallée-Tourangeau and Chris Chambers, for their continuous guidance. I feel truly privileged to have worked with you during the past few years. I am so grateful for your patience, insight, inspiration, and for the intellectual freedom you have given me.

I gratefully acknowledge Kingston University London for my full-time studentship, and for funding the experiment in Chapter 4. Thank you also to Prolific, PsyPAG, the European Association of Social Psychology, and Cardiff University for providing further research funding.

Thank you to John Lurquin who told a random MSc student to preregister their dissertation, and in so doing, opened my eyes to the world of open research. And to the members of Kingston's ReproducibiliTea journal club for continuing my open research education.

The replication studies in Chapter 2 would not have been possible without the help of the first author of the original study, Jochim Hansen. Thank you for supporting the replication attempt, and for discussing the minutiae of the plans.

I would also like to extend my thanks to the reviewers and editors who provided invaluable feedback on my Registered Reports. Their comments improved these studies in every way.

Huge thanks go to my collaborators Dan Simons, Sam Westwood, and Dale Barr. My research would not have been possible without their expertise and teamwork. Special thanks go to Dan, my rigorous-research compadre, for his kind, patient mentorship over these last few years. Thank you also for collecting the US data for Chapter 2.

For their continued support and motivation I would like to thank my friends and confidantes Anine Riege, Karis Moon, Natasha Roberts, Rose Turner, Kristin Hanson, Sue Cooper, and Jackie Thompson. Big thanks to Anne Scheel for her help translating materials, for encouraging me to learn R, and for the cat photos. Thank you also to my friends outside of academia (especially Milly, Meowtthew, San Bruno, and Chaslo) who allowed me to bend their ear over the last four years, and who can all now expound the benefits of Registered Reports, whether they like it or not.

Finally, to my family, this thesis is dedicated to you. To my husband, there are no words to express my appreciation for the constant love, unwavering support, and inspiration. If there were, those words would be repeated often. Thank you to my mum and sister, the strongest women I know, for always believing in me. In loving memory of my dad.

Published Chapters

The three empirical chapters in this thesis were written as Registered Reports.

The text of these chapters appears unchanged from the published version.

- Chapter 2: **Henderson, E. L.**, Vallée-Tourangeau, F., & Simons, D. J. (2019). The effect of concrete wording on truth judgements: A preregistered replication and extension of Hansen & Wänke (2010). *Collabra: Psychology*, 5, 19. DOI: <https://doi.org/10.1525/collabra.192>
- Chapter 3: **Henderson, E. L.**, Westwood, S. J., & Simons, D. J. (2021). A reproducible systematic map of research on the illusory truth effect. *Psychonomic Bulletin & Review* (2021). DOI: <https://doi.org/10.3758/s13423-021-01995-w>
- Chapter 4: **Henderson, E. L.**, Simons, D. J. & Barr, D. J. (2021). The trajectory of truth: A longitudinal study of the illusory truth effect. *Journal of Cognition*, 4(1), 29. DOI: <http://doi.org/10.5334/joc.161>

All preregistrations, data, code, and materials associated with the empirical chapters are available via the OSF links in “Online Supplemental Materials” section.

Declaration of Contributions

The contribution of all co-authors and supervisors to multi-author chapters is outlined using the CRediT – Contributor Roles Taxonomy.

Contributor Roles Defined	Ch. 2	Ch. 3	Ch. 4
Conceptualization – Ideas; formulation or evolution of overarching research goals and aims.	ELH, FVT, DJS	ELH	ELH
Data curation – Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later re-use.	ELH, DJS	ELH	ELH, DJB
Formal analysis – Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data.	ELH, DJS, CC	ELH	ELH, DJB
Funding acquisition - Acquisition of the financial support for the project leading to this publication.	ELH		ELH
Investigation – Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection.	ELH, DJS	ELH, FVT, DJS, SJW	ELH
Methodology – Development or design of methodology; creation of models.	ELH, FVT, DJS	ELH, DJS	ELH, DJB
Project administration – Management and coordination responsibility for the research activity planning and execution.	ELH	ELH	ELH
Resources – Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools.	ELH	ELH	ELH
Software – Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components.	ELH, DJS	ELH	ELH, DJB
Supervision – Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team.	ELH	ELH	ELH
Validation – Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs.	ELH, DJS	ELH	ELH, DJB
Visualization – Preparation, creation and/or presentation of the published work, specifically visualization/data presentation.		ELH	ELH, DJB
Writing – original draft – Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation).	ELH, DJS	ELH	ELH, DJB
Writing – review & editing – Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or post-publication stages.	ELH, FVT, DJS, CC	ELH, FVT, DJS, SJW, CC	ELH, FVT, DJS, DJB, CC

FVT = Frédéric Vallée-Tourangeau (1st supervisor), CC = Christopher Chambers (2nd supervisor), DJS = Daniel Simons (co-author), SJW = Samuel Westwood (co-author), DJB = Dale Barr (co-author)

Table of Contents

Chapter 1: General Introduction.....	1
1.1 A Background of Misinformation	1
1.2 The Illusory Truth Effect.....	4
1.2.1 Operationalisation of the illusory truth effect	7
1.2.2 Variations in procedure	8
1.2.3 The effect over time	9
1.3 The Linguistic Concreteness Effect.....	10
1.3.1 Concreteness and truth	11
1.3.2 The linguistic category model	12
1.3.3 Research using the linguistic category model	13
1.4 Processes Underlying Truth Effects	14
1.4.1 Frequency	15
1.4.2 Recognition and familiarity.....	15
1.4.3 Source dissociation.....	16
1.4.4 Processing fluency.....	16
1.4.5 Coherent references.....	17
1.5 Constructing Truth Judgements.....	18
1.6 Defining Truth.....	21
1.7 Truth Effects Summary	22
1.8 Methodology and Open Research	23
1.8.1 The credibility revolution.....	23
1.8.2 Replication	25
1.8.3 Power.....	28
1.8.4 Online data collection.....	30
1.8.5 Systematic maps	31
1.8.6 Registered Reports	34
1.9 Thesis Outline	35
Chapter 2: The Effect of Concrete Wording on Truth Judgements: A Preregistered Replication and Extension of Hansen & Wänke (2010).....	38
2.1 Introduction	39
2.1.1 Replication Value	40
2.1.2 The Present Experiments	41
2.2 Experiment 1a.....	42
2.2.1 Method.....	43
2.2.2 Results	51
2.3 Experiment 1b.....	56
2.3.1 Method.....	57
2.3.2 Results	58
2.4 Known Differences from the Original Study	61

2.5 Discussion.....	63
Chapter 3: A Reproducible Systematic Map of Research on the Illusory Truth Effect.....	67
3.1 Introduction.....	69
3.1.1 Research Aims	74
3.2 Method.....	75
3.2.1 Conformance with Reporting and Quality Standards	75
3.2.2 Search Term Identification and Selection.....	75
3.2.3 Search Strategy	77
3.2.4 Inclusion Criteria	81
3.2.5 Exclusion Criteria	82
3.2.6 Study Screening Procedure	83
3.2.7 Map Coding and Interrater Reliability	85
3.3 Results.....	92
3.3.1 Evidence Identification, Retrieval and Screening.....	92
3.3.2 Systematic Map Findings.....	94
3.4 Discussion.....	112
3.4.1 Key Findings.....	112
3.4.2 How to Use this Systematic Map and Database.....	114
3.4.3 Limitations of the Systematic Map	115
3.4.4 Future Research Summary	116
Chapter 4: The Trajectory of Truth: A Longitudinal Study of the Illusory Truth Effect.....	117
4.1 Introduction.....	118
4.1.1 Explanations, Predictions and Contradictions.....	119
4.1.2 The Illusory Truth Effect over Time.....	121
4.1.3 Our Experiment.....	123
4.2 Method.....	126
4.2.1 Participants.....	126
4.2.2 Design	127
4.2.3 Sampling Plan	127
4.2.4 Materials	131
4.2.5 Procedure	133
4.3 Analysis Plan	136
4.3.1 Outcome-Neutral Criteria	136
4.3.2 Analytic Reproducibility.....	138
4.4 Results.....	138
4.4.1 Exclusion Criteria	138
4.4.2 Confirmatory Analyses	141
4.4.3 Exploratory Analyses.....	148
4.4.4 Supplement	154

4.5 Discussion	155
4.5.1 Constraints on Generality (COG) Statement	157
4.5.2 Future Research	158
4.5.3 Conclusion	158
Chapter 5: General Discussion	160
5.1 Overview of Results	160
5.2 Synthesis	166
5.3 Limitations of Illusory Truth Effects Research	172
5.3.1 Statistical conclusion validity	173
5.3.2 Internal validity	174
5.3.3 External and ecological validity	175
5.3.4 Construct validity	176
5.4 Limitations of the Present Research	177
5.4.1 Real-world generalisability	177
5.4.2 Online experimental testing	178
5.5 Future Directions	180
5.5.1 Gist repetition	180
5.5.2 Avoiding the effect	181
5.5.3 Manipulating truth base rates	182
5.5.4 Metric calibration	182
5.6 Conclusion	183
References	185
Online Supplemental Materials	224
Chapter 2	224
Chapter 3	224
Chapter 4	224
Appendix A: Chapter 3 Benchmark List	225
Appendix B: Chapter 3 Bibliographic Database and Grey Literature Searches	227
Appendix C: Chapter 3 References Included in Full-text Database	235
Appendix D: Chapter 3 Summary of Statcheck Issues	241
Appendix E: Chapter 4 Supplemental Analyses	242
E.1 Analyses using ANOVA	243
E.2 Funnel debrief	244
E.3 Participants' views on future research	245
E.4 Association between statement topic and size of illusory truth effect	246
E.5 Association between statement length and size of illusory truth effect	248
Appendix F: Chapter 4 Amended Figure Showing Distribution of Participants' Age	250
Appendix G: Chapter 2 Effect Size Calculations and Explanations	251
Appendix H: Ethics	252

Chapter 1: General Introduction

Digital technologies have become entwined in our everyday lives and interactions, and with them comes an inexorable proliferation of information. From the trivial to the lifesaving, this information shapes the way we think and behave. Among the truths there are deceptive claims, political propaganda, and targeted marketing campaigns. Judging the truth of incoming information is one of the most challenging and important tasks that people face daily. How do we decide what is true and what is not?

The sheer volume of information we consume encourages frequent, rapid truth judgements rather than fully deliberative processing. When uncertainty exists regarding the veracity of a statement, judgements can be influenced by superficial characteristics, such as cognitive feelings, rather than the probative, informational content of the statement (see Schwarz, Jalbert, Noah, & Zhang, 2021). That is, signals that provide no intrinsic information about truth are used to inform truth judgements. This thesis investigates the robustness of two such non-probative “truth effects” driven by repetition (the *illusory truth effect*; Hasher, Goldstein, & Toppino, 1977) and concrete language (the *linguistic concreteness effect*; Hansen & Wänke, 2010). Neither concrete language nor repetition provide evidence of truth. If you read this paragraph repeatedly, it would not become truer. But research shows that repeated statements feel truer than new ones. Likewise, concretely worded statements feel truer than their abstract counterparts.

1.1 A Background of Misinformation

Researchers in areas ranging from law, to philosophy, psychology, politics, and marketing have spent decades considering the processes by which people

distinguish truth from falsehood. Public interest in the topic was spurred by the 2016 UK European Union membership referendum and the US presidential elections of the same year (Pennycook & Rand, 2021). Since then, the terms “fake news”, “misinformation”, and “post-truth” have become words of the year in leading dictionaries. Although misinformation is not a new phenomenon (Lazer et al., 2018), the increasing ubiquity of social media has changed the process and speed by which information is shared. Misinformation has been shown to influence attitudes towards real-world issues such as health (Iacobucci, 2019), politics (Bovet & Makse, 2019), and public policy (Bastos & Mercea, 2017). Decisions based on misinformation are likely to be at best suboptimal and at worst pernicious, both for society and for individuals. Some go as far as saying that misinformation on social media is a “global public health threat” (Larson, 2018, p. 309). And in an information-rich online world where digitally enabled falsehoods spread further and faster than facts (Vosoughi, Roy, & Aral, 2018) it is imperative that we understand the processes by which people judge truth. This thesis focuses on understanding people’s truth judgements in the domain of trivia statements. Using such objectively verifiable statements allows any effect of the experimental manipulations to be isolated from the effect of prior beliefs and opinions.

Simultaneously there is concern within the scientific community about the reliability and transparency of the findings from many disciplines that research truth judgements. The last decade has seen psychology in particular acknowledge a range of questionable practices that serve to undermine the credibility of research, including hypothesising after the results are known (HARKing; Kerr, 1998), interrogating the data until they offer up “positive” results (p-hacking; Simmons, Nelson, & Simonsohn, 2011), and results-based publishing decisions. Such practices

create neat narratives but bias the literature in ways that take us further from the truth. The prevalence of questionable research practices (John, Loewenstein, & Prelec, 2012; Simmons et al., 2011) implies that some research on truth effects could itself be misleading. This in turn could cause policy makers to enact ineffective policies, and scientists to waste resources chasing elusive effects, and to provide unreliable findings, breaking their contract with society (see Gibbons, 1999). For the truth effects literature to be a useful tool in combatting misinformation, it is imperative that the research itself is reliable and transparent. This chapter closes with a review of various methods that can circumvent questionable research practices and increase the reliability of research, and Chapter 3 includes a review of the illusory truth effects in terms of transparency. In implementing a range of open research innovations this thesis aims to ensure that the truth effects research herein does not become an accomplice to misinformation, but rather it provides reliable evidence to facilitate cumulative science.

I begin this chapter by reviewing the current illusory truth effect literature in the context of the societal impact of misinformation. This section includes a description of the basic illusory truth paradigm, a review of a previous meta-analysis, and a discussion of the conditions necessary to elicit the effect. The next section introduces the linguistic concreteness effect and the linguistic category model on which the effect is based. I then situate these truth effects within the broader context of how people assess truth, and describe the mechanisms thought to underlie both effects. Next, I move to considering some current challenges to psychological research (the “credibility revolution”). I describe the open research methods, used in this thesis, that can help overcome those challenges and improve the evidential value

of truth effects research. I conclude with a brief outline of the three empirical chapters.

1.2 The Illusory Truth Effect

The increasing proliferation of misinformation is particularly concerning when one considers that people do not form beliefs purely on information content: They also use non-content cues. The illusory truth effect, or repetition-induced truth effect is one of the most frequently studied truth effects. It describes the phenomenon whereby repeated items are judged as subjectively truer than previously unseen items. The effect is independent of the actual truth of the items. It is most often studied using trivia statements (Hasher et al., 1977) but also appears for opinions (Arkes, Hackett, & Boehm, 1989), consumer testimonials (Roggeveen & Johar, 2002), fake news headlines (Pennycook, Cannon, & Rand, 2018), and health statements (A. Sundar, Kardes, & Wright, 2015) including those relating to COVID-19 (Unkelbach & Speckmann, 2021).

The effect occurs for both true and false statements (e.g., Hasher et al., 1977), for both plausible and implausible facts (e.g., Fazio, Rand, & Pennycook, 2019b), and persists for known information (Fazio, Brashier, Payne, & Marsh, 2015). Repetition increases participants' beliefs even when statements are initially labelled as false (Garcia-Marques, Silva, Reber, & Unkelbach, 2015), or contested by fact checkers (Pennycook et al., 2018), when information about the statements veracity is available at the point of judgement (Unkelbach & Greifeneder, 2018), and when participants are explicitly warned about the effect in order to prevent it (Nadarevic & Aßfalg, 2017). Furthermore, illusory truth does not seem to be moderated by individual differences in cognitive ability (De keersmaecker et al., 2019) and preliminary work suggests it even occurs in children (Fazio & Sherry, 2020).

The majority of the illusory truth effect literature uses verbatim repetition of trivia statements. However, the effect may not require previous exposure to the exact statement. Just repeating the general topic (Begg, Armour, & Kerr, 1985) may be enough to increase the perceived veracity of the later statements (see section 1.4.2 for further details). Similarly, research shows that participants forget elements of statements such as qualifiers that should cast doubt on the veracity of the statements (e.g., “improbable”; Stanley, Yang, & Marsh, 2018). A single repetition is enough to elicit increased truth judgements. Research using three or more repetitions is sparse but convergent indications suggest a logarithmic relationship between repetitions and judged truth (Hawkins, Hoch, & Meyers-Levy, 2001) that wanes after several repetitions (Hassan & Barber, 2021). These results imply that repetitions have an additive effect; with multiple repetitions small effects build on each other to produce a larger effect on truth ratings. However, how numbers of repetitions interact with the time between repetitions remains an open question.

There are several factors thought to reduce and even eliminate the effect. When instructions detailed the nature of the illusory truth effect and warned participants to avoid it there was no reduction in the effect after a week’s delay (Nadarevic & Aßfalg, 2017, Experiment 1). However, with strengthened instructions and with no delay between exposure and test phases, the effect was reduced (Experiment 2)¹. This finding suggests that the effect may be so automatic that it is difficult to avoid. If warnings cannot prevent the effect, it seems that context might. When statements are presented in an all repeated list, rather than the standard mix of

¹ There were several differences between the two experiments that make the results hard to interpret. Experiment 1 was in German, in the lab, and there was a one-week retention interval between the exposure and test phase. Experiment 2 was in English with US participants, conducted online, with no retention interval between the exposure and test phase.

repeated and new, the effect does not occur (Dechêne, Stahl, Hansen, & Wänke, 2010). This result implies that repetition alone may not be sufficient to induce the effect; it might be that the ease of processing associated with repeated statements (see section 1.4.4) only emerges when compared to non-repeated stimuli.

The only meta-analysis on the illusory effect was published a decade ago (Dechêne et al., 2010). It synthesised the results of 51 studies and estimated a medium effect size between $d = .39$; 95% CI: [0.30, 0.49] and $d = .50$; 95% CI: [0.43, 0.57] (random effects model) depending on how the effect was measured. A potential issue with the meta-analysis is that the effect sizes were not adjusted for bias. The authors report in the text that the funnel plot, used to assess *publication bias* (Rosenthal, 1979), appeared symmetrical (indicating no bias towards publishing significant effects in this literature) but they did not include the plot or any other formal analyses in the paper. The prevalence of publication bias elsewhere in the literature (e.g., Fanelli, 2010; Scheel, Schijen, & Lakens, 2020; Sterling, Rosenbaum, & Weinkam, 1995) suggests that it is highly unlikely that this literature contains no publication bias, and with bias correction applied the estimate would likely be a smaller effect. The meta-analysis is also somewhat dated due to the surge of new research since its publication in 2010 (see Chapter 3, Figure 3)², partly prompted by the effect's potential to explain how people come to believe misinformation (e.g., Nadarevic, Reber, Helmecke, & Köse, 2020; Pennycook et al., 2018). These and other concerns about the meta-analysis are discussed in Chapter 3, and the issue of publication bias is considered in the methodology section at the end of this chapter.

² In the systematic map in Chapter 3, 54 out of the 93 included papers had been published since 2010.

The effect's apparent robustness is concerning and has real-world implications: The consumption and sharing of information on social media and other communications technology facilitates repeated assertions over short timescales. According to the illusory truth effect, those assertions will feel truer simply because they have been repeated. And when the repetition includes spurious claims, those false or misleading statements will feel more accurate too. Moreover, information is often not randomly repeated, but strategically targeted. Such targeting can result in the reader receiving repetitions of information that reinforce their already held views (i.e., an "echo chamber"). In combination with phenomena such as confirmation bias (Nickerson, 1998), whereby people only seek out information that confirms their existing opinion, it is easy to see how patently false information can come to be believed. Furthermore, once misinformation has been accepted, it is difficult to correct (for a review, see Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012).

1.2.1 Operationalisation of the illusory truth effect

Hasher and colleagues (1977) were the first to report the illusory truth effect. Motivated by the observation that people seem willing to make truth judgements in the absence of actual knowledge, the authors sought to explore the root of such judgements. On the basis that people are extremely sensitive to frequency, they proposed that information may enter people's pool of knowledge when it is plausible and heard frequently. To investigate this hypothesis, the authors presented participants with a range of plausible and potentially verifiable but (probably) unfamiliar trivia statements. Some were true (e.g., "The thigh bone is the longest bone in the human body") while others were false (e.g., "The capybara is the largest of the marsupials"). Participants rated the truth of the statements over the course of three sessions. In the second and third sessions, half the rated statements were

repeated from the initial session, and half were new. Repeated statements were consistently rated as truer than new statements, demonstrating that repetition increases subjective truth. Furthermore, because statements that had been repeated three times were rated truer than those repeated twice, the results indicated that frequency was used as a cue for truth.

The operationalisation of the effect has remained largely unchanged since the first observation by Hasher and colleagues. During the *exposure phase* participants read or listen to a set of stimuli, most often obscure trivia statements. Additionally, they might rate the truth of the statements or perform another judgement to ensure that the statements are processed. Typically, half the statements are actually true and half false. But since it is assumed that participants will not know the answers, the actual truth of the statement should be irrelevant to subsequent truth judgements. There then follows an *intersession interval* of varying length from zero minutes to weeks. Finally in the *test phase* participants rate the perceived truth of new statements and ones repeated from the exposure phase. The illusory truth effect is calculated as the difference in truth ratings for repeated statements between exposure and test phases (within-items) or more routinely, as the difference between truth ratings for new versus repeated statements at test phase (between-items).

1.2.2 Variations in procedure

Although the basic operationalisation is consistent across studies, there are variations within the procedure. For example, most studies select an exposure task with little justification. Yet the research that directly compares the ratings given during the initial exposure phase shows that the choice of task moderates the effect. For example participants rating interest (Brashier, Eliseev, & Marsh, 2020; Calvillo & Smelter, 2020) or categorising statements (Nadarevic & Erdfelder, 2014) show the

illusory truth effect but those rating truth do not. Likewise, in the test phase participants may be told that the ratio of truth to false statements is even, random, or they may receive no information, and this might matter (Jalbert, Newman, & Schwarz, 2020). Furthermore there is no standard indicator of subjective truth with measures varying from Likert-type scales (e.g., Hasher et al., 1977), to sliding scales (e.g., Unkelbach & Greifeneder, 2018), and binary true or false choices (e.g., Fazio et al., 2019b). The issue of lack of standardisation in exposure task and intersession interval might challenge the validity of illusory truth research; I elaborate on this point in Chapter 3.

1.2.3 The effect over time

A large body of research shows that repeated information is judged truer over a range of timescales. The effect has been observed when the test phase immediately follows the exposure phase (Unkelbach & Rom, 2017), and with delays of minutes (Unkelbach & Greifeneder, 2018), days (Stanley et al., 2018), weeks (Gigerenzer, 1984), and even months (Brown & Nix, 1996). Although the effect appears after various delays almost no studies have systematically varied the delay to investigate how the illusory truth effect behaves over time. The lack of research may be at least in part because historically most illusory truth research has been conducted using university student pools, and accessing those pools on multiple occasions might have been unfeasible. However, a meta-analysis did investigate delay and found that it did not moderate the effect (Dechêne et al., 2010) but the result was primarily based on cross study comparisons, rather than studies that directly manipulated delay. Explanations of the effect (see section 1.4) - familiarity, recognition, and an associated ease of processing - all are united by memory, so it seems plausible that time would moderate the effect.

The rare studies published since the meta-analysis that directly manipulate delay suggest time may moderate the effect. However there is no consensus as to the direction of the moderation: Nadarevic and Erdfelder (2014) found no illusory truth effect with a ten-minute delay, but they did find an effect after one week. Whereas Silva, Garcia-Marques and Reber (2017) observed an effect after a few minutes delay that halved after a one-week. Based on these conflicting findings, delay might increase or decrease the illusory truth effect, or based on the meta-analysis, delay could have no effect. Therefore, the effect of time on illusory truth remains an open but important question. If, as the meta-analysis suggests, repetition has the same effect on truth after delays of minutes or weeks, that has implications both for how misinformation comes to be believed and potentially how its effect could be reduced. More precisely, if the effect does not wane with time, then misinformation interventions would need to take that into account. The question of how the illusory truth effect behaves over time is answered in Chapter 4.

1.3 The Linguistic Concreteness Effect

Here I introduce a second truth effect: the *linguistic concreteness effect*, whereby concretely worded statements are judged to be truer than the abstract equivalents (Hansen & Wänke, 2010). This effect is of interest in the context of the illusory truth effect because without changing the contents of the information communicated, the two effects could be combined and potentially create a larger truth effect. However, to anticipate Chapter 3, I did not observe the effect in two replication attempts. Therefore, in the text that follows I describe how the linguistic concreteness effect *theoretically* works. In the methodology section (1.7) that concludes this chapter, I discuss the definitions and purpose of replication, and elaborate on the value in attempting to replicate this effect.

1.3.1 Concreteness and truth

Linguistic concreteness describes the degree to which the notion represented by a word is a “perceptible entity”, rather than an abstract, unobservable concept (Brysbaert, Warriner, & Kuperman, 2014, p. 904). Concrete representations are more vivid and correspond more readily to events held in memory (Unkelbach & Rom, 2017) thus feeling more familiar and easier to process. Ease of processing, or fluency, is the meta-cognitive experience of ease or difficulty associated with processing information (for overviews, see Alter & Oppenheimer, 2009; Unkelbach & Greifeneder, 2018). A number of studies have reported such processing differences as a function of concrete language (Huang & Federmeier, 2015). Concrete words and sentences are generally responded to more quickly, and comprehended more quickly and accurately (de Groot, 1989; Schwanenflugel & Shoben, 1983), and this also holds for truth judgements (Belmore, Yates, Bellack, Jones, & Rosenquist, 1982; West & Holcomb, 2000).

Hansen and Wänke (2010) counts as one of relatively few truth effect studies that does not involve repetition (see also McGlone & Tofiqbakhsh, 2000; Reber & Schwarz, 1999). Not only do the illusory truth effect and linguistic concreteness effect share a possible underlying mechanism (i.e., processing fluency), they also share the characteristic that they influence truth judgements through a mechanism unrelated to the contents being communicated. These non-content cues provide the sense or illusion of truth rather than actual information about a statement’s truth. These cues manifest in all language, no matter the content and include tropes such as fluency, richness of detail, and vividness (Shidlovski, Schul, & Mayo, 2014). Concreteness affords such cues, and this may explain the link between truth and concreteness.

1.3.2 The linguistic category model

Whereas previous research typically manipulated concreteness by varying content and elaboration of detail, Hansen and Wänke (2010) took a more subtle linguistic perspective. They manipulated only the richness of semantic information, not by changing the content or adding detail, but by varying the description of the content. So, for example, where the concrete verb “kissing” conjures a vivid, easily verifiable image, the abstract adjective ‘loving’ requires additional interpretation to create a concrete mental image.

This characterisation of concreteness is based on the linguistic category model (LCM; Semin & Fiedler, 1988, 1991) which sets out the cognitive implications of four linguistic categories. Each category elicits inferences about stability and verifiability based on its position on a concreteness-abstractness dimension. The most concrete category is that of descriptive action verbs. These verbs are easily verifiable, require no interpretation and refer to a single, concrete, behavioural event. They preserve the perceptual properties of the event (e.g., “A punches B”). Interpretive action verbs signify a general class of behaviour and require some interpretation (e.g., “A hurts B”). Here “hurting” may be physical or mental, accidental or intentional, but it is not explicit from the verb. Moving further towards abstractness, state verbs are detached from observable behaviour and refer primarily to a psychological state rather than an event (e.g., “A hates B”). At the abstract end of the dimension are adjectives following a conjugation of the verb “to be”, abstract mediators that are potentially unobservable references to a person’s psychological being (e.g., “A is aggressive”). As demonstrated by the examples in brackets, the same event can be portrayed to a varying degree of abstraction, while still being a valid representation of the event (Semin, 2000a).

1.3.3 Research using the linguistic category model

The LCM was originally developed in the domain of social cognition and defines how the use of linguistic categories affects the way a person and their behaviour is perceived (Semin & Fiedler, 1988). Since its development, the LCM has been used to code written descriptions for abstractness of language (e.g., Fujita, Trope, Liberman, & Levin-Sagi, 2006; Gong & Medin, 2012). Schmid and Fiedler (1996) used the LCM to code transcripts from the Nuremberg trials of German Nazi generals. They found subtle strategies at work, for example prosecutors used concrete language to signpost the responsibility of the defendant (i.e., they produced the highest rate of action verbs). In this sense linguistic concreteness was used as a device to guide and direct the addressee (Semin, 2000b).

Hansen and Wänke (2010) then applied the LCM model to truth judgements of unfamiliar trivia statements. Based on the linguistic devices outlined in the model and the link between concreteness, perceived vividness and the realness of events, the authors hypothesised, and found, that statements of the same semantic content were judged to be more probably true when written using concrete language (e.g., "The poet C. Dickens *wrote* the play Miss Sara Sampson."), than those written in more abstract language (e.g., "The play Miss Sara Sampson *is by* the poet C. Dickens."; emphasis added). While the LCM has been frequently employed in the realm of person perception, Hansen and Wänke's 2010 paper was the first to apply the LCM to truth judgements.

Since then, and based on the research showing that concrete messages are perceived as more verifiable (Semin & Fiedler, 1988), and true (Hansen & Wänke, 2010), recent studies using the LCM have shown that concrete messages might serve to persuade the addressee. For example, concrete messages had a more convincing

effect on voters, and specifically concrete messages were more persuasive when the participants' political allegiance diverged from that of the messenger (Menegatti & Rubini, 2013). A second study on voting intentions manipulated pronouns ("you" or "we") and level of abstractness (using verbs in concrete statements and adjectives in the abstraction versions; Chou & Yeh, 2018). Again, even though the content of the message remained constant across the two types of message, the contextualised cues provided by concrete messages evoked more favourable voter responses. Concrete language in eyewitness statements also served to persuade mock jurors that defendants were more likely to be guilty in ambiguous criminal cases (Kurnec & Weaver, 2018, Experiment 1). As in Hansen and Wänke (2010) the semantic content of the manipulation did not differ, but concrete verbs such as "he walked" were replaced by abstract equivalents such as "he came". However, the effects were small and did not replicate in Experiment 2. In sum, when considering concreteness as characterised by the LCM, there is mixed evidence regarding the effect on persuasion but just one paper showing its effect on truth. Given this paucity of literature, replication is a useful tool to establish the reliability of the effect before attempting to extend the research (see section 1.7.2).

1.4 Processes Underlying Truth Effects

If concreteness and repetition functionally affect truth judgements, how might that be so? Next, I turn to the processes thought to underlie these two truth effects. After over 40 years of research of the illusory truth effect, there are several explanations as to why people judge repeated information as being subjectively truer than new information. All explanations are associated with memory and are closely related (for a review, see Unkelbach, Koch, Silva, & Garcia-Marques, 2019). Below I discuss the explanations, starting with the hypothesised mechanism behind the

original demonstration of the illusory truth effect and focusing on the dominant explanation – processing fluency. As mentioned previously, processing fluency is also the most likely candidate to explain the linguistic concreteness effect.

1.4.1 Frequency

In the first lab demonstration of the illusory truth effect, Hasher and colleagues reasoned that when no verifying information is available, people infer truth from the frequency of plausible statements (Hasher et al., 1977). People are highly sensitive to frequency, and this might therefore explain how information becomes knowledge.

1.4.2 Recognition and familiarity

Bacon (1979) argued that since frequency was an inference from memory, it would be inefficient to use frequency to cue truth. Instead, he showed that *perception* of repetition was critical to eliciting the effect. Whether participants recognised statements as being repeated was more predictive of increased truth ratings than whether the statements had been repeated (Bacon, 1979). That is, participants rated repeated statements as truer only when they recognised them as having been repeated. Begg, Armour and Kerr (1985) tested the familiarity hypothesis based on the idea that people remember the substantive content of statements better than the minor details. In the exposure phase, participants listened to a list of topics (e.g., “hen’s body temperature”), then during the test phase they rated the truth of full statements (e.g., “The temperature of a hen’s body is about 104 degrees Fahrenheit”). Initial exposure only to the topic made the later statements feel more familiar and truer compared to statements on unfamiliar topics. Taken together, these studies indicate that statements judged to be familiar are perceived as more likely to be true.

1.4.3 Source dissociation

Linked to familiarity is the source dissociation account. In this explanation it is not repetition per se that increases validity but the perception that the repetitions have come from independent sources, thus conveying convergent validity (Arkes, Boehm, & Xu, 1991; Arkes et al., 1989). Therefore, in situations where people remember the semantic content of a statement but not its source, they will show the illusory truth effect. However, the effect also occurs after minutes, and when people correctly remember the source of a repeated statement (i.e., within the experiment; Bacon, 1979), so source dissociation alone cannot explain the effect.

1.4.4 Processing fluency

The dominant explanation of the illusory truth effect is processing fluency (for a review, see Alter & Oppenheimer, 2009). Indeed, it likely shares this underlying mechanism with the linguistic concreteness effect. Ease of processing, or fluency, is the meta-cognitive experience of ease or difficulty associated with the processing and comprehension of information. People use this experience as a source of information to supplement the actual probative information being processed (Schwarz et al., 2021). Fluency is a unitary construct with multiple causes and implications (see Alter & Oppenheimer, 2009, Table 1). According to Alter and Oppenheimer (2009), ease of information processing may be experienced as conceptual fluency (related to the meaning of the stimulus, e.g., semantic priming; Kelley & Lindsay, 1993), perceptual fluency (related to the physical attributes of the stimulus, e.g., colour contrast; Reber & Schwarz, 1999) and linguistic fluency (concrete language; Hansen & Wänke, 2010; and rhyming; McGlone & Tofiqbakhsh, 2000). The effect of fluency is similar regardless of how it is engendered (Alter & Oppenheimer, 2009).

The implications of fluently written statements vary, but most pertinently, variables that serve to increase subjective ease of processing increase truth judgements (Marsh, Cantor, & Brashier, 2016; Reber & Schwarz, 1999). This association occurs because people use their feelings as information but are insensitive to their source (for a review, see Schwarz, 2012), and therefore mistakenly ascribe the experience of ease of processing as information about the statements' veracity. Repeated statements are more familiar, and concrete statements more vivid, and thus feel easier to process relative to new or more abstractly worded statements (Unkelbach & Rom, 2017; Unkelbach & Stahl, 2009). The relative comparison is important, because fluency is more influential when experienced as a change in fluency. That is, repeated statements feel more fluent when *compared* to their initial presentation and to non-repeated statements (for a review, see Wanke & Hansen, 2015). Note however that when the source of fluency is obvious, it is disregarded as informative and its influence is diminished or eliminated (Alter & Oppenheimer, 2009; Schwarz et al., 2021). Thus, the processing fluency account posits that when participants notice that statements are repeated, for example over short delays between exposure and repetition, they will ignore repetition as a source of veracity information. Furthermore with training the effect can be reversed so that fluency is associated with falsity rather than truth (Unkelbach, 2007).

1.4.5 Coherent references

Most recently, Unkelbach and Rom (2017) integrated previous explanations into a “referential theory” of repetition-induced truth effects. The theory argues that the first presentation of a sentence links previously unlinked references (e.g, “vaccinations cause autism”). Repetition reactivates the same links and references in memory (i.e, “vaccinations” and “autism”). Statements with a richer set of links will

benefit from more fluent processing and feel more familiar (Unkelbach & Rom, 2017).

1.5 Constructing Truth Judgements

The effects of concrete language and repetition can be compounded by people's tendency to default to truth. In most situations people proceed on the tacit assumption that incoming information is relevant and honest (Grice, 1975) unless a salient cue suggests otherwise. Where the prevalence of honest (i.e., lacking deceptive purpose, though not necessarily completely accurate) communication is relatively higher than deceptive messages, as in much of daily life, this bias promotes efficient communication and co-operation, and may therefore be considered adaptive (Brashier & Marsh, 2019; Levine, 2014; Reber & Unkelbach, 2010). However, in situations where the number of deceptive claims outweigh honest messages, such a bias is maladaptive and would result in inaccurate truth judgements. When deception does occur, people are poor at detecting it, doing so at levels barely above chance (Bond & DePaulo, 2008; Street, 2015). People are also poor at picking up on errors (i.e., knowledge neglect; Marsh & Umanath, 2013). These factors converge, making the acceptance of a statement more likely than its critical scrutiny, which suggests that when it occurs, people are vulnerable to deceit (Levine, 2014).

Defaulting to truth is consistent with Gilbert's Spinozian-inspired model of belief which contends that the comprehension and acceptance of a proposition occur simultaneously and precede disbelief (Gilbert, Krull, & Malone, 1990; Gilbert, 1991). That is, belief is the default, and is required to comprehend a statement. Rejecting a claim, and labelling it as false, occurs as a second, more cognitively effortful step. In contrast, the Cartesian account asserts that a proposition must be comprehended before it is assessed and labelled true or false (Gilbert, 1991). There is

therefore an initial period of indecision, and if a person has insufficient information to tag the proposition, it remains in an unlabelled limbo. Recent work suggests that when forced to make binary truth/false judgements, people show a Spinozian truth bias, but appear more Cartesian when allowed to indicate uncertainty (Street & Richardson, 2015). A comprehensive review of the Spinozian and Cartesian models of belief is beyond the scope of this thesis; for further discussion, see Asp et al. (2020), Levine (2014), Nadarevic and Erdfelder (2013), Street and Kingston (2017), and Street and Richardson (2015).

Truth judgements are constructed based on inferences from various cues (Brashier & Marsh, 2019). Schwarz (2015) outlines five criteria people may consider when evaluating the truth of a statement: 1) Is the claim compatible with my existing knowledge? 2) Is the claim coherent? 3) Is there evidence for the claim? 4) Does the claim come from a credible source? 5) Do others agree with the claim? Crucially, information from both declarative and meta-cognitive experiential sources (i.e., ease of processing) are always available and used in the assessment of these criteria (Schwarz et al., 2021; Schwarz, 2015)³. For example, declarative assessments of credibility would include evaluation of the communicator's expertise and motive, whereas experience based evaluations may be based on feelings of fluency associated with an easy-to-pronounce name (Silva, Chrobot, Newman, Schwarz, & Topolinski, 2017). Similarly, compatibility could be checked analytically in a relatively slow and effortful process of comparing incoming information to one's knowledge, or it could be founded on the ease of processing associated with having read the information before.

³ Brashier and Marsh's (2019) three-part framework similarly includes declarative sources (memories, e.g., source, knowledge), and experiential sources (feelings, e.g., processing fluency), and adds inferences from base rates (i.e., most claims are true) as a third component in truth judgements.

Across all five criteria fluent processing signals truth (e.g., statements feel more consistent, better supported) where disfluent processing signals uncertainty and the need for further scrutiny (Schwarz, 2015). This indicates that variables that facilitate fluent processing, such as concrete language and repetition, increase subjective truth judgements regardless of the assessment criterion(s) used.

Declarative information is more likely to be used when people have the cognitive ability and time to engage in more intense processing. Conversely, people might be more reliant on the faster route of using feelings as information when ability, motivation, and time are scarce (for a review, see Greifeneder, Bless, & Pham, 2011). This suggests that the way people often consume online information, such as skimming social media, encourages a reliance on feelings as information. And the feelings associated with processing information are always available, where probative information might not be.

In this sense, fluency or feelings-based truth judgements can be categorised within a broader set of judgements originating from the heuristics and bias programme (e.g., Kahneman, 2003). Heuristics are mental shortcuts that speed up or reduce the cognitive load of decision making (for a review, see Keren & Teigen, 2004). While heuristics often lead to acceptable estimates, they are imprecise and can result in systematic suboptimal judgements. For example, the availability heuristic, whereby the likelihood or frequency of an event is evaluated based on the ease with which relevant examples can be brought to mind (Tversky & Kahneman, 1973). The original definition of availability is compatible with the concept of retrieval fluency, one of several types of processing fluency (see Reber & Greifeneder, 2017; Riege & Reber, 2022). The formative work by Tversky and Kahneman propagated a new way of conceptualising the decision-making process. While the original heuristics and

biases primarily pertained to judgements of frequency or probabilities, the concept has broadened to include many other types of judgements. In the case of truth effects, processing fluency likely provides a useful heuristic to inform decisions about truth or falsity.

1.6 Defining Truth

The definition of truth has been, and will continue to be, the topic of much philosophical and semantic debate. In this thesis I define truth in the sense that knowledge is possible: Since truth should be objective, our knowledge of true propositions must be about real things. This circumspect view has parallels with the correspondence theory of truth which posits that truth matches or accurately describes something real. When I refer to research bringing us closer to or further from “the truth”, I am referring to an absolute truth - the objective truth to which science aspires.

In the experimental Chapters 2 and 4, I use verifiable claims that refer to concepts and people external to the believer (e.g., claims about geography). Thus while truth effects are likely influential in higher stakes settings, and on topics about which people hold dearly held beliefs, I selected trivia statements as stimuli specifically because they comprise topics about which people would not have previously constructed their own truth. Furthermore, trivia statements can be classified as either “true” or “false” based on them being objective, externally verifiable facts about the world.

However in this thesis I also refer to subjective truth judgements. In the present work, the truth value of a given statement is the tool used to study people’s subjective judgements of absolute truths. Therefore, when I refer to subjective truth judgements this is how close to objective truth the participant feels a statement is -

which is, by definition, placing the idea of truth on a spectrum. When participants feel that statements are truer, or increase their belief, they are attributing a greater level of truth to that statement.

1.7 Truth Effects Summary

People *construct* truth judgements based on a range of cues. In the current social and political climate an understanding of the factors that affect truth judgements is particularly pertinent. If repetition and concreteness increase truth judgements independent of the message content this is important to understand, not only because these mechanisms can be used insidiously, but because they could be used to increase the cognitive congeniality of facts. The truth by its very nature is constrained by the facts. Those facts may be complex, boring or both. In contrast, misinformation can enlist intriguing titbits, novel and emotional content, tailor-made to persuade the reader. Theoretically, concrete language could be used to shape important socio-political messages ensuring that they are communicated in a clear and accessible manner, and repetition used at appropriate intervals as necessary to reinforce the message and help counter misinformation.

While the illusory truth effect appears robust, much of the foundational research was conducted prior to the 2010s when issues of replicability came to the forefront (see section 1.7.1). Additionally, there is a paucity of literature regarding how the time between repetitions interacts with illusory truth. In contrast to the illusory truth effect, the linguistic concreteness effect has only been shown in one paper. The paper has been frequently and widely cited as evidence that concrete wording increases truth judgements (e.g., Beukeboom, Tanis, & Vermeulen, 2013; Elliott, Rennekamp, & White, 2015). To date no research has considered how these two potentially complementary truth effects might work when combined.

1.8 Methodology and Open Research

This section introduces the methodological approaches adopted in this thesis. First, I describe the credibility revolution currently occurring in psychology and beyond, in order to situate the work in the broader context of scientific change. I go on to discuss the open methods applied in this research - replication, systematic mapping, and Registered Reports - that establish credibility, synthesise the literature, and reduce bias respectively. I also outline the related design considerations of power and the research setting.

1.8.1 The credibility revolution

Reliable, transparent truth effects research is the basis of theory advancement and the development of real-world interventions and applications. However, the last decade has seen psychology, along with other scientific disciplines, undergo a credibility⁴ or replicability “crisis” (Giner-Sorolla, 2019) or “revolution” (for a more detailed discussion, see Spellman, 2015). It has been a period of introspection and improvement driven by a series of events relating to reproducibility and replicability. The most consequential of those events include first, the publication of an article revealing how easily researchers can exploit questionable research practices (QRPs; such as selectively reporting variables) to obtain statistically significant findings for non-existent effects (Simmons et al., 2011), and many psychologists admitting to engaging in those practices (John et al., 2012). Second, the *Journal of Personality and Social Psychology* published a controversial study purporting to show extrasensory perception (Bem, 2011). Third, several high-profile examples of scientific fraud have come to light (Funder, 2014). Last, there was the observation that 91.5% of articles published in psychology claimed support for their first

⁴ The term “credibility” refers to how believable a claim is based on the available evidence.

hypothesis (the highest of any science in the study; Fanelli, 2010). In sum, these events highlighted that during the last few decades “it was impossible to distinguish between findings that are true and replicable and those that are false and not replicable” (Nelson, Simmons, & Simonsohn, 2018, p. 512).

There followed several high-profile multi-lab replication projects. Replication refers to the uncommon practice of repeating an experimental procedure and obtaining results in the same direction as the original study⁵. Notably, the Open Science Collaboration attempted to replicate 100 studies from three top psychology journals. Just 36% yielded significant findings and the effect sizes were around half the size of those in the original studies (Open Science Collaboration, 2015; see Gilbert et al., 2016 for a commentary). Similarly, Multi-Labs 2 attempted to replicate 28 classic and contemporary papers, testing how the effects varied as a function of samples and settings. Despite well-powered designs, only 54% of studies replicated (Klein et al., 2018). These and other unsuccessful replication attempts suggest in part, that there may be numerous false positives in the psychology literature driven by factors including QPRs (Simmons et al., 2011) and low power (Pashler & Harris, 2012). Low powered studies might be prevalent because small studies are more feasible to run than larger ones. Such studies are prone to producing large effect sizes, and the publication process is biased towards “positive” results (Fabrigar, Wegener, & Petty, 2020) while non-significant effects languish in file drawers.

The credibility revolution highlighted the flaws in the research and publication process, primarily caused by practices obscured by the current publishing and incentive system. The renewed focus on these weaknesses catalysed new

⁵ Replicability is distinct from (computational) reproducibility. Computational reproducibility refers to attempts to reproduce the original results using the original raw data and analyses.

initiatives and calls to improve the transparency and rigour of research using preregistration, Registered Reports, open data, code, and materials (Chambers, 2013; Munafò et al., 2017; Nosek et al., 2015; Nosek, Ebersole, DeHaven, & Mellor, 2018), by reporting all results and data exclusions (Vazire, 2018), and by normalising replications (Everett & Earp, 2015; Zwaan, Etz, Lucas, & Donnellan, 2018). One central theme of these initiatives was to facilitate the identification of false positives in the current literature, and minimise their existence in future research (Fabrigar et al., 2020).

1.8.2 Replication

If the goal of science is to advance knowledge, then we must ensure that knowledge is built on a foundation of trustworthy claims. Differentiating what is replicable from what is not is a necessary condition of such knowledge building, and fundamental to the scientific process (Zwaan et al., 2018). If a claim is true, it should replicate under specifiable conditions and can be used to make predictions about the future. If the claim does not replicate, it cannot. When the likelihood is that many published findings are false, then replication is an efficient way to prevent researchers from building on unreliable effects (Coles, Tiokhin, Scheel, Isager, & Lakens, 2018) and avoid wasting resources pursuing research based on false positives (Giner-Sorolla et al., 2019). I intended to build on the linguistic concreteness effect but since I could not check the reproducibility of the original results, or adequately evaluate the strength of evidence (Vazire, 2017), I planned an independent replication attempt.⁶

⁶ During the planning stages of the replication the first author did share the data but could not locate the code.

Replications are typically termed *close* or *conceptual* (but see Nosek & Errington, 2020). The decision to pursue one over the other hinges on the intended function of the study. Replication studies are considered “close” or “direct” when they are conducted using procedures and materials that match the original as closely as possible. If an effect is robust⁷, it should be observable under the conditions of a close replication (Simons, 2014). Close replications are therefore used to help determine whether the original finding was credible. When a close replication attempt is rigorous, well documented, adequately powered, and conducted under the conditions originally specified, obtaining results that differ from the original raises concerns about the reliability of the original results.

On the other hand, conceptual replications test the original hypothesis or result but varying some aspect(s) of the design (Schmidt, 2009). Thus, a result similar to the original study is informative about the generalisability of the effect. That is, the effect still occurs under the new conditions. However conceptual replications are less effective at falsifying chance findings. When the results differ from the original it is not possible to establish whether the difference is due to the features the replicator intentionally varied, or that the original study was invalid (e.g., due to sampling or measurement error).

1.8.2.1 Trust but verify. Chapter 2 presents the results of two replication attempts of the linguistic concreteness effect reported in Hansen and Wänke (2010). My primary reason for replicating was that because the effect had only been reported in one paper, before attempting to build on it, I wanted to confirm that it was not a chance finding (i.e., control for sampling error). That is, I sought evidence that the

⁷ Robust refers to the stability of experimental findings to variations in experimental procedures and analytic strategies.

original findings reflected an effect that could be separated from the specific circumstances of the original experiment (time, place, participants, etc.). If an effect is highly context-dependent, it is not solid ground on which to build new knowledge. The appropriate mechanism for establishing the robustness of the previous finding is close replication.

Close replications should follow the recipe laid out in the original paper, and keep all elements of the experimental design as faithful as possible to the original (Brandt et al., 2014; Schmidt, 2009). This proposition presents two considerable challenges: First, following the recipe from the original paper. The traditional focus on results along with article word limits mean that the methods sections of papers are often lacking. Below I describe how employing Registered Reports helps these issues. In this case I was fortunate that the first author of the original study was forthcoming and provided the missing details of the procedure and the original materials. Second, because no replication is exact, a close replication will necessarily have changes compared to the original. What is critical is that those changes are not theoretically relevant, and that they are transparently reported so that others can assess the replication's utility. In Chapter 2, differences are reported in a section entitled "Known Differences from the Original Study". Moreover, when planning my replication attempts, I discussed the differences with the first author of the original study, and they believed them to be irrelevant in obtaining evidence in the same direction as the original.

1.8.2.2 Replication Value. Isager and colleagues (2020, p.1) define replication value as "...the maximum expected utility we could gain by conducting a replication of the claim, and is a function of (1) the value of being certain about the claim, and (2) uncertainty about the claim based on current evidence.", In relation to

parameter 1, at the time of the replication attempt, Hansen and Wänke (2010) had been cited over a hundred times (Google Scholar, August 27, 2018, approximately 10 times the mean for 2010). The paper has also received media attention indicating that this topic has broad interest from both the academic community and the public, and as noted, has potential practical implications.

In respect to the second parameter, there are several areas of uncertainty. First, despite its theoretical and applied importance, the linguistic concreteness effect has not been replicated within truth judgements. Second, we do not have the tools to scrutinise the original study (i.e., raw data, code, preregistration) and this brings uncertainty. Third, the original authors themselves seem surprised by the effect: “Although the effects are small, it is still remarkable that such a slight manipulation as the subtle linguistic variation can account for any variance in subjective truth at all” (Hansen & Wänke, 2010, p. 1585). Fourth, the authors incorrectly halved a p-value resulting in the claim of a significant effect: When comparing truth ratings for concrete versus abstract statements aggregated across participants, the authors state that the ANOVA is “one-tailed” and report a p-value of .041. The correct p-value for that test is .082. Finally, the sample size is relatively small ($N = 46$), and in their second experiment there is evidence of the Proteus phenomenon (Ioannidis & Trikalinos, 2005) - the larger sample produced a smaller effect. Considering the potential effect of this subtle manipulation, and the uncertainty of the original results, the benefits of replication outweighed the costs of building on unstable ground.

1.8.3 Power

The reproducibility crisis has highlighted, among other things, the need to collect larger sample sizes to achieve the statistical power necessary to detect true effects. Adequate power is critical for maximising the informativeness of a study’s

results (Ledgerwood, Soderberg, & Sparks, 2017) and is a key component of a replication study (Brandt et al., 2014). Appropriately sized samples reduce both false positive and false negative rates (see Button et al., 2013) thus increasing replicability and cumulativeness respectively. Whereas running studies with weak sample sizes is futile: They jeopardise statistical conclusions and contribute to a lack of reproducibility. Low power inflates the risk of false positives: Although the rate of false positives is fixed, when power drops there are fewer true positives, hence the proportion of true positives decreases (Button et al., 2013). In the context of replication, an underpowered study is not informative because it cannot distinguish between a false negative in the replication study, and a false positive in the original study.

The challenge then is determining the effect size for the power calculation. The presence of publication bias in the literature suggests that many unimpressive results are unpublished, and the effect sizes in the published literature are likely overestimates (Giner-Sorolla et al., 2019). One could then consider powering to detect a smallest effect size of interest (SESOI) that we should theoretically or practically care about (Lakens, Scheel, & Isager, 2018). However, theories in psychology are often underspecified and do not often stipulate a SESOI. I therefore chose to power to small effect sizes relative to previous work. In Chapter 2 both studies had greater than 95% power to detect an effect half the size of the original. For the lab study the sample size ($N = 253$) was achieved via collaboration facilitated by the StudySwap platform (StudySwap, 2018): Half the data were collected in the US and half in the UK.

Another solution for increasing power is to move from the lab and recruit participants online using platforms such as MTurk or Prolific, where adequate

sample sizes can be collected with relative ease and speed. The results of online and lab-based studies are often comparable (e.g., Gosling, Vazire, Srivastava, & John, 2004; Schnoebelen & Kuperman, 2010; Weigold, Weigold, & Russell, 2013). In the longitudinal study in Chapter 4, I powered to the bottom of the distribution of effects reported in the literature. As this meant recruiting over 600 participants, I chose to collect data online using Prolific. Further details of the power calculations for all experiments are available in the sampling plan sections of their respective chapters.

1.8.4 Online data collection

Beyond increasing sample sizes, there are multiple benefits to collecting data online. Such research may have more external validity for two reasons: First, online experiments are run in the familiar conditions of participants' homes on devices that they use every day, meaning that the environmental conditions are more ecologically valid than lab experiments. Second, online samples are generally more diverse (e.g., in terms of age, socio-economic status) than the convenience samples recruited from university participant pools often used for lab-based experiments (Gosling et al., 2004). Admittedly, online samples will be biased towards those who are sufficiently wealthy and educated enough to have access to the Internet (Gosling & Mason, 2015), but less so than studies using university students. Additionally, Prolific allows researchers to pre-screen based on eligibility criteria, providing accurate targeting while maintaining heterogeneity in non-target criteria (Paolacci, Chandler, & Ipeirotis, 2010). There is also the ethical consideration that participants from online platforms have voluntarily signed up to research, but students may feel pressured to participate for course credits. Furthermore, demand and experimenter effects are likely to be lower (Reips, 2002). From a practical perspective, online platforms

facilitate the collection of longitudinal data (such as that in Chapter 4) that might otherwise be logistically difficult.

Along with the benefits that come with conducting research online, there are also various possible disadvantages. One potential concern is a lack of participant naivety (i.e., “professional survey takers”). Prolific compares favourably to MTurk in terms of higher levels of participant naivety and lower levels of dishonest behaviours (Peer, Samat, Brandimarte, & Acquisti, 2017). Another drawback is the lack of control over the experimental setting in terms of distraction, appropriate lighting etc. Without the supervision of a researcher, there are more opportunities for participants to lie about their eligibility for an experiment, to cheat, and to respond randomly or multiple times. Various measures can be put in place to minimise the effect of these limitations. The measures I enlisted are detailed in the Methods sections of relevant chapters and summarised in Chapter 5 (section 5.4.2).

1.8.5 Systematic maps

As the number of published studies rapidly increases year by year, summarising the literature has become essential practise for those wishing to draw conclusions about the state of a particular research area (i.e., cumulative science). Arguably, meta-analysis, a statistical technique for combining multiple studies, has been considered at the top of hierarchy of evidence, precisely because their conclusions rely on a body of evidence rather than a potentially fallible single study (Siddaway, Wood, & Hedges, 2019). However, as discussed, that body of evidence is likely subject to biases including publication bias and selective reporting bias (e.g., Franco, Malhotra, & Simonovits, 2014; John et al., 2012; Simmons et al., 2011), and these biases have the potential to undermine meta-analytic conclusions (Corker, 2018). An empirical test of the reliability of meta-analyses compared 15

preregistered, multi-lab replications with 15 meta-analyses on the same topic (Kvarven, Strømmland, & Johannesson, 2020). On average, the meta-analyses inflated the true effect size by a factor of three, and generally applying bias adjustment methods⁸ did not improve the results. Given these and other issues with meta-analysis (for a more detailed discussion, see Corker, 2018) other synthesis methods can be considered.

A possible alternative to meta-analysis is the traditional narrative review that aims to describe some aspect or subset of the available literature (Gunnell, Poitras, & Tod, 2020). Narrative reviews do not follow a standardised or reproducible methodology and may therefore be susceptible to selection bias by way of “cherry-picking” appropriate studies, resulting in incorrect or misleading results (Corker, 2018; Haddaway et al., 2020). Alternatively, systematic approaches comprehensively synthesise the literature and represent a more rigorous and less biased method. One example of the latter is systematic mapping. A systematic map is a form of synthesis that uses pre-planned, transparent methods to catalogue the available evidence on a topic, demonstrating what research has been done and where (Haddaway et al., 2019; James, Randall, & Haddaway, 2016). Unlike a systematic review, it does not aim to answer a specific research question, and different to meta-analysis it does not attempt to quantify the effectiveness of a particular intervention (see Wolffe, Whaley, Halsall, Rooney, & Walker, 2019, Table 2 for a comparison of systematic reviews and maps).

Central to the philosophy of systematic approaches is that the research protocol, including methods for searching, screening, and data extraction, is

⁸ Trim-and-fill, 3PSM, and PET-PEESE.

registered (and in some cases published) *a priori*. This process both avoids mission creep and improves the transparency and replicability of the review (Haddaway et al., 2020). Reporting guidelines such as ROSES (Haddaway, Macura, Whaley, & Pullin, 2018b), NIRO-SR (Topor et al., 2020), and PRISMA (Page et al., 2021) also help ensure that all relevant methodological information has been considered and reported.

The output from a systematic map is first and foremost an open, searchable database of relevant studies. This filterable library can be used by researchers as the basis of other targeted research synthesis or interrogated to identify interesting patterns and help pinpoint fruitful lines of study. The process also identifies knowledge gaps – areas where studies might reasonably be expected but are missing, and knowledge clusters – groups of similar studies that warrant further synthesis through systematic review or meta-analysis (Haddaway et al., 2019). Systematic maps typically make recommendations for best practice in methods for future research.

Given that the only existing meta-analysis on the illusory truth effect is outdated, I had originally intended to conduct an update. However, piloting revealed substantial underreporting of the information necessary to calculate effect sizes. I instead chose to synthesise the literature using systematic mapping. The motivation for systematic mapping stems from curiosity about the state of knowledge on a particular topic (i.e., “what evidence exists on the effect of repetition on truth judgement?”). Maps are therefore suitable for heterogeneous areas of research, such as the illusory truth literature, that originate from disciplines including social and cognitive psychology, consumer research, and education. Chapter 3 further details the process that led to the decision to undertake a systematic map. The map includes

a review of the current levels of open research practices used in the illusory truth effects literature.

1.8.6 Registered Reports

Science is naturally messy. Yet current research culture incentivises tidy narratives in which researchers “predicted” the positive results in a linear hypothetico-deductive narrative. Although aesthetically pleasing, these stories are not what research looks like. Furthermore, novel results are prized over replications and extensions (Makel, Plucker, & Hegarty, 2012; Schmidt, 2009). The pressure to produce clean narratives, in order to obtain a job/promotion/grant/tenure, incentivises QRPs, distorting and threatening the validity of research (“Tell it like it is,” 2020). Preregistration is being increasingly adopted as means to tackle selective reporting, QRPs, and publication bias (Chambers, 2013; Nosek & Lakens, 2014; van ’t Veer & Giner-Sorolla, 2016; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Preregistration entails specifying and publicly registering the study protocol *a priori* so that others may transparently evaluate the evidential value of the research.

Taking this concept further, Registered Reports were conceived both to reduce bias in research, and to free authors from the pressures of producing novel, positive findings, and instead reward transparent, rigorous work (Chambers & Tzavella, 2020; Nosek & Lakens, 2014). In this model initial peer review of the study protocol (including rationale, methods, and analysis plan) occurs before the study is run. The publication decision is based on the importance of the research question, and the rigour of the methods. By making the decision results blind, Registered Reports take the focus off outcome-based decisions and provide a powerful antidote to publication bias. Furthermore, authors are required to describe their methodology with sufficient detail and transparency that another researcher

could replicate it (i.e., provide a replication recipe), and assess the credibility of the scientific claims made.

While Registered Reports are not a cure-all, preliminary evidence suggests that they offer clear benefits over traditional publishing formats. A recent study showed that when assessed by peer reviewers, Registered Reports outperformed standard papers on all 19 criteria, including creativity, the rigour of the methods and analysis, and the importance of the outcomes (Soderberg et al., 2020). The results from Registered Reports are also more reproducible than those from standard reports, though there is room for improvement (Obels, Lakens, Coles, & Gottfried, 2019). The format appears to be working as intended with regard to safeguarding against publication bias. The number of supported hypotheses are much lower in Registered Reports – 44% in psychology (Scheel et al., 2020) and 39% in various fields (Allen & Mehler, 2019), compared to standard articles – 96% (Scheel et al., 2020). These results suggest that research conducted using the Registered Reports format is not subject to publication bias in the same way as the standard literature.

1.9 Thesis Outline

Inspired by psychology's credibility revolution, this thesis applies many methods that have potential to increase the quality, replicability, and transparency of research. All three empirical chapters were written as Registered Reports (Chapters 2 and 4 have been published, Chapter 3 is under Stage 2 review) with open protocols, materials, data, and code. These methods can improve the credibility of truth effects research and provide a more faithful representation of real research. This will ultimately accelerate future scientific progress and better inform academics, the public, and potential misinformation interventions. In this thesis I aim to answer three central questions relating to truth effects by replicating, cataloguing, and

extending previous research. Overall this thesis asks “Now we understand that our old methods produced findings that are not diagnostic of truth (Nelson et al., 2018), what does truth effects research look like when it’s done using transparent approaches?”.

In Chapter 2, I aim to establish the credibility and stability of the linguistic concreteness effect by attempting to replicate the effect as originally reported in Hansen and Wänke (2010) and since cited over 192 times (Google Scholar, 06 June 2021). I use the original materials and procedures in two studies: classroom (as per the original experiment), and online. Both experiments were well powered to detect an effect half the size of the original but I did not detect an effect in either study. I interpret these results as likely casting considerable doubt on the robustness of the original claims, and I therefore do not pursue this line of research.

Chapter 3 is a systematic map that catalogues the illusory truth effect spanning over 40 years of research from the original paper in 1977 to those published in 2020. Along with collaborators I coded papers at study level for 74 variables under the broad categories of methodological information (study design, stimuli, and subjects), number of repetitions and test phases, outcome, and adherence to transparent, open research practices. The chapter describes areas of research that are well developed and identifies gaps where further research is necessary, as well as providing a snapshot of reporting practices from the entirety of a single research area. The output from the map is a comprehensive, open database that can be used by the research community to plan future work.

Chapter 4 is a longitudinal study designed to answer a fundamental question about whether the illusory truth effect is dependent on time. The research extends

previous work, most pertinently it directly tests a claim from the 2010 meta-analysis, that the effect is independent of time. That is, that the size of the illusory truth effect remains the same regardless of whether the repetition occurs after minutes or weeks. I report the results of a well-powered, within-subjects, focused exploration of the effect over two short (within-session, one day) and two longer (one week, one month) intersession intervals. I also consider factors that may moderate the effect.

The thesis therefore provides an overview of the current state of truth effects research, with a particular focus on the illusory truth effect. It takes initial steps towards testing the basis of such work and provides suggestions on how further to improve the credibility of the research area, while embracing reforms in transparency and rigour. More specifically, this research programme contributes to knowledge by 1) establishing the robustness (or not) of the linguistic concreteness effect, 2) synthesising the illusory truth effects literature, and 3) providing evidence for the role of time as a moderator of the illusory truth effect.

Chapter 2: The Effect of Concrete Wording on Truth Judgements: A Preregistered Replication and Extension of Hansen & Wänke (2010)

Abstract

When you lack the facts, how do you decide what is true and what is not? In the absence of knowledge, we sometimes rely on non-probative information. For example, participants judge concretely worded trivia items as more likely to be true than abstractly worded ones (the *linguistic concreteness effect*; Hansen & Wänke, 2010). If minor language differences affect truth judgements, ultimately they could influence more consequential political, legal, health, and interpersonal choices. This Registered Report includes two high-powered replication attempts of Experiment 1 from Hansen and Wänke (2010). Experiment 1a was a dual-site, in-person replication of the linguistic concreteness effect in the original paper-and-pencil format ($n = 253$, $n = 246$ in analyses). Experiment 1b replicated the study with an online sample ($n = 237$, $n = 220$ in analyses). In Experiment 1a, the effect of concreteness on judgements of truth (Cohen's $d_z = 0.08$; 95% CI: [-0.03, 0.18]) was smaller than that of the original study. Similarly, in Experiment 1b the effect (Cohen's $d_z = 0.11$; 95% CI [-0.01, 0.22]) was smaller than that of the original study. Collectively, the pattern of results is inconsistent with that of the original study.

Keywords: replication, truth judgements, truth effect, concreteness, language, Registered Report

2.1 Introduction

The perceived truth of a statement can be influenced by factors other than its probative, informational content (Koriat & Adiv, 2012), including the source of the information, the context in which it is presented, and characteristics of the statement itself (Dechêne et al., 2010). This paper examines an effect of the statement wording: Participants judge concretely worded trivia items as more likely to be true than abstractly worded versions of the same content (the linguistic concreteness effect; Hansen & Wänke, 2010). For instance, the statement, “The poet C. Dickens wrote the play Miss Sara Sampson,” was judged more likely to be true than the more abstract equivalent, “The play Miss Sara Sampson is by the poet C. Dickens.” Across all statements, more concrete versions were judged as more probably true than their abstract equivalents (Cohen’s $d_z = 0.48$).

This manipulation is based on the linguistic category model (Semin & Fiedler, 1988, 1991) which posits that a concrete verb (“wrote”) conjures a vivid, reliable, and easily verifiable image, but an abstract one (“is by”) does not (Semin & Fiedler, 1988). The model was originally designed to assess descriptions of people’s behaviour, and it has also been applied to analyses of persuasion and influence. For example, prosecutors in the Nuremberg trials used concrete language to signpost the responsibility of Nazi generals (Schmid & Fiedler, 1996).

According to the model, descriptive action verbs, such as “wrote” or “punch” require no interpretation; they refer to a single, concrete, behavioural event and convey the perceptual properties of that event (e.g., “A punches B”). All of the concrete statements used by Hansen and Wänke (2010) contained such descriptive action verbs. In contrast, their abstract statements described the same event but required more interpretation (e.g., “A hurts B”). Although their abstract statements

were guided by the linguistic category model, they did not fully implement it. Some of their abstract statements contained no state verbs or adjectives, the two categories classified as abstract in the model. Those statements that lacked state verbs or adjectives “map the criteria of the LCM of abstractness (e.g., high stability, low situational dependency)” (J. Hansen, personal communication, January 25, 2018) and rely on characteristics associated with abstract word categories rather than always containing the word categories themselves.

2.1.1 Replication Value

Understanding how and when belief in the truth of a statement is influenced by its superficial characteristics rather than its substance is of great practical and theoretical importance. That a statement’s truthiness can influence judgements is well established (Fazio et al., 2015; Newman, Garry, Bernstein, Kantner, & Lindsay, 2012; Newman et al., 2015). Most studies examining the factors that influence truth judgements, other than the statement’s substance, have focused on the *illusory truth effect* (Hasher, Goldstein, & Toppino, 1977); repeated statements are believed more than new statements. A Google Scholar search for “illusory truth effect” revealed 247 results, compared to 2 for “linguistic concreteness effect” (Google Scholar, August 27, 2018). Yet, Hansen and Wänke’s (2010) experiment underlies research on the persuasiveness of concrete language in political communication (Menegatti & Rubini, 2013), voting intentions (Chou & Yeh, 2018), and eyewitness testimony (Kurnec & Weaver, 2018). Given the ease with which this concrete/abstract manipulation can be applied in practice and the estimated effect from the original study ($d_z = 0.48$), the experimental manipulation merits further investigation and a more precise estimate of the effect size. In practice, manipulation of beliefs via linguistic

concreteness might be easier to do and harder for readers to notice. If robust, the effects of linguistic concreteness could potentially be combined with the illusory truth effect (or other such effects) to yield even greater effects on beliefs.

Despite its theoretical and practical implications, the effect of linguistic concreteness on truth judgements has not been independently replicated, either conceptually or closely. In light of this paucity of research, combined with the practical implications of the effect if it proves robust, we designed two high-powered replications using sample sizes substantially bigger than the original study. We undertook this research for three further reasons. First, Hansen and Wänke's (2010) experiment has been heavily cited (102 citations according to Google Scholar, August 27, 2018, approximately 10 times the mean for 2010) and used to motivate research on topics ranging from political persuasion to eyewitness testimony. It has also been discussed in the media as a technique for increasing trustworthiness (e.g., Stott, 2011). Second, the relatively subtle manipulation of concreteness yielded an effect size of Cohen's $d_z = 0.48$, but the sample size ($n=46$) means that the estimate was not precise (95% CI [0.19, 0.78]). A direct replication using the same materials will verify the effect and estimate its size more precisely. Finally our second experiment will directly replicate the original study using the same design, but with a different source of participants (online) to determine whether the effect is equally robust.

2.1.2 The Present Experiments

With guidance from the original authors, we designed a high-powered, pre-registered replication of Experiment 1 from Hansen and Wänke (2010). We aimed to match, as closely as possible, the conditions and methods of the original paper with an implementation that addressed those factors that the original authors believe are

necessary for obtaining the effect. Like the original study, we tested the prediction that participants would judge concretely worded trivia items as more probably true than abstractly worded versions (H1 - confirmatory hypothesis). We also added several enhancements and extensions. First, to test whether the effect would generalise beyond the originally sampled population, we tested participants in the United Kingdom and the United States, both with in-person samples and online. Second, to ensure that our primary hypothesis tests were adequately powered to detect the original effect and to enable a more precise measure of the effect size, we tested approximately five times as many participants as the original experiment. Third, in their fourth study, Hansen and Wänke (2010) inferred that some participants already knew answers to some of the trivia items (i.e., their objective truth value). Consequently, we added a check for prior knowledge of the answers to the trivia questions. Finally, at the suggestion of the first author of the original study, we used an expanded stimulus set to test the exploratory hypothesis that the perceived psychological distance of the statement content would interact with the concreteness of the wording (H2 - exploratory hypothesis).

For both experiments, we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures (Simmons, Nelson, & Simonsohn, 2012). Our preregistration, materials, and data are available on the Open Science Framework at <https://osf.io/s2389/>. Our stage 1 manuscript, and a supplement outlining the changes between our stage 1 and stage 2 manuscripts, can also be found there.

2.2 Experiment 1a

Experiment 1a was designed to replicate the linguistic concreteness effect (Hansen & Wänke, 2010, Experiment 1). Participants judged the truth of trivia items

and we assessed whether, in the absence of self-reported knowledge of the correct answer, their judgements were influenced by the concreteness of the wording.

2.2.1 Method

Our replication follows the procedures of the original paper and uses the original materials provided by the authors (translated from the original German wording). In consultation with J. Hansen (personal communication, January 25, March 01, April 09, and April 16, 2018), we further adapted those materials to our participant populations in order to test the same hypotheses as the original (see below). Differences between this experiment and the original are outlined in the “Known Differences from the Original Study” section below. The experimental procedures were approved by both the Kingston University Research Ethics Committee and the University of Illinois Institutional Review Board. Participants provided informed consent before participating.

2.2.1.1 Sampling plan. There is no clear theoretical lowest effect size of interest for the linguistic concreteness effect that we can use as the basis of a power analysis. As an alternative, we could use the effect size from the original study for power analysis, but that effect size might not reflect the “true” effect due to chance variation, sampling error, and the possibility of publication bias. Consequently, we conducted a sensitivity power analysis using G*Power 3.1.7 (Faul, Erdfelder, Lang, & Buchner, 2007) to determine the smallest effect that we would have high statistical power (95%) to detect given pragmatic constraints on our total sample size. Our preregistered plan was to collect usable data from 210 participants (five-times the original sample size), which would give our sample 95% power to detect

an effect of $d_z = 0.23$ at $\alpha = 0.5$ (one-sided)⁹. Hansen and Wänke (2010) reported effect sizes (η^2_p) of 0.19 and 0.08 (for Experiments 1 and 2 respectively), which correspond to Cohen's d_z of 0.48 and 0.29 (Lakens, 2013). Given that both reported effects are larger than $d_z = 0.23$, our planned sample had greater than 95% power to detect the originally reported effects as well (with our sample size, we have greater than 95% power to detect an effect that is 50% the size of the original Experiment 1).

2.2.1.2 Participants. Undergraduate students (and some masters students in the UK) participated in the study in exchange for course credit or a chance to win one of three £50 prizes. These incentives were used in Hansen and Wänke's (2010) Experiment 1 and Experiment 2, respectively. Participants were recruited via a dual-site collaboration enabled by StudySwap (StudySwap, 2018); approximately half the participants were from Kingston University, UK and half from the University of Illinois at Champaign-Urbana, USA. For recruiting purposes, and in line with the original study, the experiment was described as a “study on truth judgements.” Our final sample was larger than our target sample due to higher signup rates and lower no-show rates than anticipated during scheduling (UK: $n = 130$, $M_{age} = 24.7$; USA: $n = 123$, $M_{age} = 19.3$).

2.2.1.3 Materials. Two native German speakers translated the original 52 trivia statements from German to English. These items cover a myriad of general knowledge topics including history, geography, and science. Half of the statements are true and half are false. All statements are plausible but describe facts that few

⁹ The design used to conduct the power analysis was a paired t-test (called “Means: Difference between two dependent means (matched pairs)” in G*Power). This clarification is thesis-specific and does not appear in the published version. Additionally, effect sizes that were reported to three decimal places in the published version, have been reduced to two decimal places in the thesis.

participants know. Each trivia item has both an abstract and a concrete version, with concreteness determined using linguistic category model criteria (Semin & Fiedler, 1988, 1991). For example, in the first statement in Table 1, “wrote” is more concrete (i.e., a descriptive action verb) than “is by.” To maximise the chance of observing the linguistic concreteness effect, we took care to ensure that the English translations complied with the description of the original items. For each statement: 1) the concrete version contains a descriptive action verb; 2) both versions were approximately the same length; and 3) the abstract and concrete versions used equally common language because any unusual words were common to both versions of each statement (e.g., words like “bandoneon” were core to the content of the statement). The translation was checked by the first author of the original paper.

Table 1

Examples of Trivia Statements with Abstract and Concrete Phrasings

Statement	Concrete	Abstract
1	The poet C. Dickens wrote the play Miss Sara Sampson.	The play Miss Sara Sampson is by the poet C. Dickens.
2	The Roe River flows into the Missouri River.	The Roe River is a tributary of the Missouri River.
3	People nicknamed the Cuban composer Esteban Salas y Castro the "Santiago Angel".	The Cuban composer Esteban Salas y Castro was also known as the "Santiago Angel".

Note: Statements 1 and 3 are false: The author of Miss Sara Sampson is Gotthold Ephraim Lessing. Esteban Salas y Castro had no well-known nickname.

2.2.1.3.1 Updated statements. Hansen and Wänke (2010) argued that the match between concreteness and psychological/physical distance also influences truth judgements. In their Experiment 4, concretely worded items presented in the foreground of a landscape photograph (i.e., close) were judged to be more true than those presented in the background. Similarly, abstract items presented in the background were judged to be more true than those presented in the foreground.

These effects presumably result from a match in the participant's mindset: both physical proximity and linguistic concreteness activate a more "concrete" mindset which increases judged truth values. A mismatch in those factors reduces truth judgements. In reviewing our replication plan, Hansen suggested that the content of some original items might induce a similar "distance" effect. In the original experiment, some of the statements related to culture and history local to Switzerland, and those statements might be more psychologically distant for a Briton or an American. That distance might interact with the linguistic concreteness effect.¹⁰

The original experiment was conducted at a university in Switzerland. The first author coded each statement as being either spatially close, distant, or neutral from Switzerland, the UK, and the USA, and these judgements were checked by the first author of the original study. We then generated additional trivia items (modelled on the originals) for those deemed close for Swiss participants but far for Britons (8 items) or Americans (18 items).¹¹ Thus, participants in the UK judged a total of 60 items and USA participants judged 70 items (see Table 2). The new statements were modified versions of the original items created by swapping words that conveyed spatial distance for our participants for equivalent spatially close words while maintaining the concreteness/abstractness of the original item. For example, we changed "In Hamburg, one can count the largest number of bridges in Europe" to "In London you can count the largest number of surveillance cameras in Europe." We did not change the actual truth of the new statements (i.e., if the original statement was true the replacement was also true). The statements, modifications, and plans for

¹⁰ We analysed the original data from Hansen and Wänke's (2010) Experiment 1 and did not observe the predicted effect of distance.

¹¹ One statement (about Swiss Cantons) that was likely not understandable for UK and USA participants was amended and remains in the original 52.

confirmatory analyses were discussed with the first author of the original paper. Confirmatory analyses were carried out on both the 52 original statements, and the updated version containing 52 statements in which distant statements have been removed and replaced with close statements. Planned secondary analyses explored whether the linguistic concreteness effect differed for the matched subsets of original and replacement items (8 for Britons and 18 for Americans).

Table 2

Psychological distance of the original items in each country (with numbers of replaced items)

Stimuli	Spatial psychological distance		
	Close	Distant	Neutral
Swiss Original Study			
Original items	20	12	20
UK Replication			
Original items	12	20	20
Changed	8	-8	0
USA Replication			
Original items	3	30	19
Changed	17	-18	1

Note. Original = original 52 statements; Changed = number of added or removed statements needed to match the proportions in original set.

2.2.1.3.2 Statement verification. Before conducting the study, we followed the same procedures used by Hansen and Wänke (2010) to ensure that the concrete versions of the statements were seen as more concrete than were the abstract ones. We combined all trivia items into a single set of 78 (52 original + 18 USA-specific items + 8 Britain-specific items), and then created two sets of 78 items (set A and set B) so that the concrete and abstract version of each item appeared in different sets. Four student raters (2 for set A and 2 for set B), who were blind to the experimental hypothesis and who were briefly trained on the pertinent aspects of the linguistic category model (see <https://osf.io/s2389/> for complete training instructions) then independently coded each item on a 1 (*most concrete*) to 4 (*most abstract*) scale. For

set A, the correlation between raters was $r = .77$; 95% CI: [0.66, 0.85)]¹². For set B, the correlation between raters was $r = .81$; 95% CI: [0.72, 0.88)]¹³. As in the original experiment, concrete versions were consistently coded as more concrete than their corresponding abstract versions (see Table 3).

Table 3

Coder concreteness ratings for concrete and abstract statements, their difference, and the confidence interval around that difference

Statements	n	Concrete M(SD)	Abstract M(SD)	Diff	95% CI
Overall	78	1.67 (0.69)	3.27 (0.59)	-1.60	[-1.81, -1.40]
Set A	78	1.75 (0.75)	3.28 (0.60)	-1.53	[-1.85, -1.22]
Set B	78	1.59 (0.62)	3.26 (0.60)	-1.67	[-1.96, -1.39]
True	39	1.77 (0.81)	3.36 (0.57)	-1.59	[-1.91, -1.28]
False	39	1.57 (0.53)	3.19 (0.61)	-1.61	[-1.89, -1.34]
Old	52	1.78 (0.72)	3.27 (0.62)	-1.49	[-1.74, -1.23]
New	26	1.45 (0.57)	3.28 (0.56)	-1.83	[-2.19, -1.48]

In the experiment, the statements were presented in the same two sets (A and B), and in same order as in the original experiment, with the new items randomly interspersed among them (we used <https://www.randomizer.org/> to allocate positions). If a new version of a statement was assigned to a position within five places of the corresponding original statement, it was re-randomised. In each set, half of the statements are actually true and half are false. Each trivia item appears only once in each set, in either its abstract or concrete form; statements presented as

¹² CIs for the correlations have been added to this thesis and do not appear in the published version.

¹³ The use of Pearson's correlation to quantify the reliability of non-continuous data has been criticised. However, in recent work using simulated ordinal data and two raters, the conclusions drawn from Pearson's correlation were similar to those drawn from intraclass correlation, Spearman's rho, Kendall's tau-b, and quadratic kappa (de Raadt, Warrens, Bosker, & Kiers, 2021). This clarification is thesis-specific and does not appear in the published version.

concrete in set A were presented as abstract in set B, and vice versa. The concreteness and actual truth of the statements were fully crossed.¹⁴

In the original study the statements were presented across four pages, with the following number of statements on each page: 15 (including instructions), 17, 17, 3. We standardised the number of statements presented across the paper-and-pencil (Experiment 1a) and online formats (Experiment 1b). The first page presented the instructions and four statements; each page thereafter contained six statements (except that the last page in the UK set contained two statements). The UK set consisted of 11 pages and the USA set consisted of 12 pages.¹⁵

2.2.1.4 Procedure. The experiment followed the procedure used by Hansen and Wänke (2010), including directly translated instructions. It was administered as a paper-and-pencil questionnaire study to students enrolled in introductory psychology and other undergraduate and masters psychology classes. The experiment was conducted in classrooms. Participants were given one of the two versions of the questionnaire (set A or set B) containing 60 (UK) or 70 (USA) statements in a fixed random order. Questionnaire packs were distributed to participants in each sample in alternating order to ensure that approximately equal numbers of participants received each set. In each set, half the statements were actually true and half were false, and for each actual truth value, half the statements were abstract and half concrete. Items that were concrete in Set A were abstract in Set B, and vice versa. Participants were asked to judge the truth of each statement

¹⁴ Note that in the original study, sets A and B had unequal numbers of concrete and abstract versions of the items (the design was not fully crossed between truth value and concreteness). To fully cross the factors in the replication, we swapped the concrete and abstract versions of two of the original items between set A and set B.

¹⁵ Due to a copy/paste error in creating the printed packets for the USA versions, the item numbering was out of sequence (... 38, 39, 40, 44, 45, 46, 41, 42, 43, 47, 48 ...). We noticed the error after testing had already started, so we did not change it for the remaining participants. The sequence was correct for the USA online version and for both laboratory and online UK versions.

on a scale ranging from 1 (*definitely false*) to 6 (*definitely true*; Hansen & Wänke, 2010, p. 1579). In short, English-speaking participants at each testing site were randomly assigned to a 2 (concreteness of statements: concrete vs. abstract) x 2 (actual truth: true vs. false) x 2 (statement set: set A vs. set B) mixed design with the first two factors varied within participants and the last factor varied between participants.

In Experiment 4 of Hansen and Wänke (2010), which used a subset of these statements, the authors inferred from the pattern of responses that a few participants knew the answers to some items. We added a check for prior knowledge to ensure that ratings were of items with unknown truth value. After completing all truth judgements, participants viewed the list of items again, and indicated next to each item if they knew the answer to that item. After completing the trivia items and the knowledge check, participants reported their age, gender, nationality, the number of years they had lived in the UK/USA, and whether they had used any sources to find out answers to any of the items. Finally, participants were thanked and debriefed. The experimental tasks were self-paced and took approximately 10-20 minutes to complete. The experimenter remained in the room for the duration of the experiment. Given that successful recruiting from the subject pool in the USA required a longer testing session (approximately 40 minutes was needed to receive a full credit), most participants in the USA completed an additional packet of questionnaires following completion of the tasks for this study (see online supplement for more information).

The experimental data were entered into spreadsheets. The UK data files were verified by re-entering all numbers and cross-checking discrepancies. The USA data files were verified by reading aloud the entered numbers from the spreadsheet while

an assistant verified that they matched the responses in the packets. Any entirely ambiguous responses (e.g., two numbers marked) were coded as missing. These verified data files are stored on OSF along with the data from Experiment 1b.

2.2.2 Results

Analysis scripts were generated from pilot data that was created by having the first author repeatedly complete each survey herself (varying her responses to questions to allow tests of various exclusion rules). All analyses were written using R (R Core Team, 2018) and the following packages: tidyverse (Wickham, 2017), janitor (Firke, 2018), datatable (Dowle & Srinivasan, 2018), varhandle (Mahmoudian, 2018), ez (Lawrence, 2016), BayesFactor (Morey & Rouder, 2018), summarytools (Comtois, 2018), and bootES (Gerlanc & Kirby, 2015). This manuscript was written in RMarkdown (Allaire et al., 2018) and formatted using papaja (Aust & Barth, 2018), knitr (Xie, 2015), kableExtra (Zhu, 2018), and xtable (Dahl, 2016). The RMarkdown file includes the full analysis script and results are analysed and inserted into the manuscript without human intervention. The scripts, data, and RMarkdown files are available at <https://osf.io/s2389/>. Unless explicitly noted otherwise, all exclusion rules and analyses followed the pre-registered plan specified in our stage 1 manuscript.

For the primary analyses, data were pooled across country (UK and USA) and across set (A and B). The original study excluded no participants. We excluded responses to any items that were already known by a participant (as indicated by checking the box next to that item in the knowledge check), regardless of whether their actual answer was correct or incorrect. We excluded data from any participant who elected to end their participation prior to completing the study ($n=4$), who self-reported using technological aids to answer questions ($n=2$), or who responded

uniformly (e.g., always answer 1) to all statements in either the original 52 items or the new set of 52 items ($n=0$).

In addition to the preregistered exclusion criteria, we excluded participants who reported knowing 59 or more items ($n=1$) because they could not be included in the primary analyses after excluding “known” items (see Table 4). Finally, we did not enter data from one additional USA participant who the experimenter observed marking responses in a pattern (1-2-3-4-5-6-5-4-3-2-1, etc.) without reading the items.

Table 4

Participants recruited, excluded, and analysed, separated by country, set, and gender for Experiment 1a

Group	N recruited	N excluded	N analysed
UK Set A			
Male	8	0	8
Female	55	2	53
Gender Variant	0	0	0
Not Reported	1	1	0
UK Set B			
Male	11	1	10
Female	52	0	52
Gender Variant	1	0	1
Not Reported	2	2	0
USA Set A			
Male	17	1	16
Female	43	0	43
Gender Variant	1	0	1
Not Reported	1	0	1
USA Set B			
Male	30	0	30
Female	30	0	30
Gender Variant	0	0	0
Not Reported	1	0	1
Total	253	7	246

Note. Recruited includes all participants who started the study, even if they did not complete it.

For both Experiment 1a and 1b, as in the original paper, our primary, confirmatory analyses examined the effect of concreteness of language on the perceived truth of trivia statements, with the six-point Likert ratings as the

dependent measure. The linguistic concreteness effect predicts that Likert scores should be higher for concretely worded statements than for more abstractly worded statements. We separately computed each participant's mean rating across items falling into each combination of the truth of the statement (true/false) and the concreteness of the statement (concrete/abstract). Our confirmatory hypothesis tests were based on the data after exclusions and after removing any items that participants reported having known previously, and the online supplement presents further exploratory analyses including items that participants reported knowing already.

2.2.2.1 Primary confirmatory analyses. The original study used a mixed-design ANOVA to analyse the effects of concreteness, actual truth, and set. Given that we had no a-priori hypotheses about actual truth or set, we did not use an ANOVA for our confirmatory hypothesis test. For completeness, we report the results of a comparable ANOVA (adding country as a factor) in the online supplementary materials at <https://osf.io/s2389/>.

As a test of the linguistic concreteness effect, we directly compared the average responses to concrete and abstract statements in a paired, one-sided t-test for the original 52 items (H1). Average ratings for concrete items ($M = 3.57$, $SD = 0.41$) were about the same as those for abstract items ($M = 3.54$, $SD = 0.41$), $t(245) = 1.21$, $p = .115$, $BF_{10} = 0.29$ (The Bayes Factor used $r_{scale} = 0.34$, the d_r effect size for the original study, as an informed alternative hypothesis)¹⁶. The Bayes Factor shows that our observed difference is 3.45 times¹⁷ more consistent with the null

¹⁶ The calculation for the d_r effect size and an explanation of the scale parameter can be found in Appendix G. This footnote and appendix are thesis-specific and do not appear in the published version.

¹⁷ $BF_{01} = 1/BF_{10}$ (i.e., $1/0.29$). This clarification is thesis-specific and does not appear in the published version.

hypothesis of no difference or a negative effect than with a distribution centred at the original effect size.

Given that the t-test was not statistically significant, we compared the upper confidence bound around the observed effect (observed effect: Cohen's $d_z = 0.08$; 95% CI: [-0.03, 0.18]) to the criterion value from our sensitivity power analysis (Cohen's $d_z = 0.23$) to determine whether the observed effect was “inferior” to that planned minimum effect. Because the upper bound of the confidence interval was smaller than 0.23, the observed difference between truth ratings for the concrete and abstract statements in the revised set of items was statistically inferior to a positive effect of Cohen's $d_z = 0.23$.

The same analysis conducted on the revised set of 52 items – replacing items that were close for the Swiss participants in the original study with new items that were close for the UK or USA participants (H1) – revealed a pattern that was similar to that for the original 52 items: Average ratings for concrete items ($M = 3.58$, $SD = 0.40$) were again about the same as those for abstract items ($M = 3.55$, $SD = 0.40$), $t(245) = 1.60$, $p = .056$, BF_{10} (with $r_{scale} = 0.34$) = 0.49. The Bayes Factor shows that our observed difference is 2.06 times more consistent with the null hypothesis of no difference or a negative effect than with a distribution centred at the original effect size.

Given that the t-test was not statistically significant, we compared the upper confidence bound around the observed effect (observed effect: Cohen's $d_z = 0.10$; 95% CI: [0.00, 0.20]) to the criterion value from our sensitivity power analysis (Cohen's $d_z = 0.23$) to determine whether the observed effect was “inferior” to that planned minimum effect. Because the upper bound of the confidence interval was smaller than 0.23, the observed difference between truth ratings for the concrete

and abstract statements in the revised set of items was statistically inferior to a positive effect of Cohen's $d_z = 0.23$.

2.2.2.2 Secondary exploratory analyses. Hansen and Wänke (2010) found that physical distance moderated the linguistic concreteness effect (Experiment 4). In their study, items were displayed against a photographic background so that they appeared either near or far. Concrete items were judged to be more true when they were close and abstract items were judged to be more true when they were far. In consulting with Hansen about the design of our replication, he suggested a conceptual replication of that effect based on the geographic proximity of the item contents to our participants. That suggestion motivated the addition of the new items, but it also permits a conceptual replication of the proximity effect. We compared truth ratings for the original “distant” versions of statements (those judged to be geographically “close” for Swiss participants but remote for participants in the UK or USA) with the new replacements for those items (8 original and updated items for the UK, and 17 for the USA; in the USA, one additional close item was replaced by a neutral item to ensure a fully crossed design with a total of 18 new items) that were intended to be “close” for our participants (see Table 5). For close items, the difference between concrete and abstract should be positive, because of the conceptual “match” between concrete and close and the mismatch between abstract and close. In contrast, for distant items, the difference between concrete and abstract should be negative, because of the conceptual “mismatch” between concrete and distant and the match between abstract and distant. Consequently, we compared difference scores (*Concrete – Abstract*) between the original (distant) and replacement (close) items with a one-sided t-test (H2).

Partially consistent with the prediction that a match between proximity and concreteness would increase truth judgements, the difference between concrete and abstract was positive for the close items ($M = 0.06$), but it was also positive for the distant items ($M = 0.02$), and near zero in both cases, $t(245) = 0.67, p = .253$.

2.3 Experiment 1b

The research reported in Hansen and Wänke (2010) tested undergraduate participants in person using paper-and-pencil materials. This extension attempted to replicate the linguistic concreteness effect using the same materials as Experiment 1a but in an online setting.

A growing literature suggests that people process online material more superficially, relying on heuristics to judge message credibility (Metzger & Flanagin, 2013; S. S. Sundar, Knobloch-Westerwick, & Hastall, 2007) and believability (Sungur, Hartmann, & van Koningsbruggen, 2016). If so, we might expect to observe a larger linguistic concreteness effect online. Conversely, a recent meta-analysis of studies of the illusory truth effect (Dechêne et al., 2010) showed a reduction in effect size online; when judgements of a set of repeated statements were compared to judgements of new statements (between-items), the effect size was reduced from $d = 0.59$ using paper-and-pencil to $d = 0.30$ on the computer. The reasons for this reduction are unclear, but the authors suggested it might be due to differences in presentation time (i.e., constrained intervals or participant paced) or presentation appearance (i.e., how many statements are presented at once). Given that Experiment 1b samples from a different population using a different medium, differences in absolute performance levels and the size of the concreteness effect could differ between Experiments 1a and 1b for many reasons. Hence, rather than

directly comparing the effect sizes in the two studies, we report whether the linguistic concreteness effect emerges in each study relative to the same standard set by our sensitivity analysis.

2.3.1 Method

2.3.1.1 Participants. As for Experiment 1a, our plan was to continue recruiting participants until we had usable data from 210 participants, with approximately half from the USA and half from the UK. Participants were recruited and tested online using the Prolific platform and Qualtrics. We used Prolific's pre-screening to ensure that participants were between 18 and 65 years of age, listed English as their first language, and had a "participation on Prolific" approval rating of 98% or higher (Final sample: UK: $n = 120$, $M_{age} = 34.3$; USA: $n = 117$, $M_{age} = 33.2$). The experimental procedure was approved by the Kingston University Research Ethics Committee, and participants provided informed consent before completing the study. Each participant was randomly assigned to set A or set B, and as in Experiment 1a, they completed equal numbers of items in each cell of a design that fully crossed concrete/abstract and true/false. Upon completion of the experiment, participants received £2.18 as compensation.

2.3.1.2 Materials and procedure. Except as noted, the materials and procedure matched those used in Experiment 1a. To ensure that the formatting, font size, and number of statements on each page were the same between Experiments 1a and 1b, we created the Qualtrics survey used in Experiment 1b first and produced the paper-and-pencil version from that version. To promote consistency in the appearance of the items, we constrained the study to allow participation only via a desktop or laptop computer (rather than a handheld device). At the end of the experiment, participants reported the type of device they used to complete the

survey and whether or not they used any technology to aid their responses. The UK survey can be viewed at <https://bit.ly/2NrUKmc>, and the USA survey can be viewed at <https://bit.ly/2PLgrPF>.

Table 5

Means and SDs for new items (close) and distant items (replaced original) for Experiments 1a and 1b

	Item type	Concrete M(SD)	Abstract M(SD)	Diff	Correlation (r)
Experiment 1a					
	Close	3.64 (0.62)	3.58 (0.56)	0.06 (0.72)	0.26
	Distant	3.57 (0.56)	3.55 (0.54)	0.02 (0.64)	0.32
Experiment 1b					
	Close	3.59 (0.59)	3.55 (0.62)	0.05 (0.79)	0.16
	Distant	3.60 (0.60)	3.54 (0.57)	0.06 (0.61)	0.46

2.3.2 Results

The planned data analysis and exclusion rules were identical to those of Experiment 1a, with an added criterion to account for overly fast or slow completion of the study in the absence of an experimenter observing data collection in person. We set the “maximum time allowed” to 45 minutes within the Prolific settings, and we also excluded participants who completed the study in less than 3 minutes.¹⁸ We excluded data from any participant who elected to end their participation prior to completing the study ($n=0$), who self-reported using technological aids to answer questions ($n=9$), who responded uniformly (e.g., always answer 1) to all statements in either the original 52 items or the new set of 52 items ($n=1$), or who reported knowing 59 or more items ($n=1$) because they could not be included in the primary analyses after excluding “known” items.

¹⁸ The first author was able to complete the survey in approximately 2 minutes when responding randomly to all items and neglecting to read the instructions. In pilot testing of the online version of the study (Experiment 1b), no participant completed the study in less than 5 minutes.

Given that online participants could cheat by looking up the answers, and that we could not identify overly long response times to individual questions using Qualtrics, we used the data from Experiment 1a to establish a plausible accuracy level (because participants in Experiment 1a could not easily cheat in answering questions). We calculated the mean number of questions that each participant correctly answered in Experiment 1a, where we operationally defined a correct answer as a response of 1 (*definitely false*) when the statement was false and 6 (*definitely true*) when the statement was true. We excluded any participant in Experiment 1b whose percentage correct according to that same standard was more than three standard deviations above the mean from Study 1a (Experiment 1a $M = 0.08$, $3SD$ cutoff = 0.34; total excluded $n=9$; note, though, that 3 of those participants had already been excluded for self-reported use of technological aids).

Given that we anticipated needing to replace some excluded participants, we initially collected data from 240 participants, with the plan to test additional batches of 20 participants as needed until we achieved final sample with usable data from at least 210 participants (see Table 6).

Table 6
Participants recruited, excluded, and analysed, separated by country, set, and gender for Experiment 1b

Group	N recruited	N excluded	N analysed
UK Set A			
Male	22	0	22
Female	37	1	36
Gender Variant	1	0	1
Not Reported	0	0	0
UK Set B			
Male	23	1	22
Female	37	5	32
Gender Variant	0	0	0
Not Reported	0	0	0
USA Set A			
Male	19	4	15
Female	37	2	35
Gender Variant	1	0	1

Not Reported	0	0	0
USA Set B			
Male	33	2	31
Female	27	2	25
Gender Variant	0	0	0
Not Reported	0	0	0
Total	237	17	220

Note. Recruited includes all participants who started the study, even if they did not complete it.

2.3.2.1 Primary confirmatory analyses. As in Experiment 1a, we compared the average responses to concrete and abstract statements in a paired, one-sided t-test for the original 52 items (H1). Average ratings for concrete items ($M = 3.66$, $SD = 0.42$) were about the same as those for abstract items ($M = 3.63$, $SD = 0.38$), $t(219) = 1.61$, $p = .055$, BF_{10} (with $r_{scale} = 0.34$) = 0.51. The Bayes Factor shows that our observed difference is roughly equally consistent with the null hypothesis of no difference as with a distribution centred at the original effect size; it does not favour either hypothesis over the other by more than a 2:1 ratio (although it is 1.95 times more consistent with the null than the alternative).

Given that the t-test was not statistically significant, we compared the upper confidence bound around the observed effect (observed effect: Cohen's $d_z = 0.11$; 95% CI: [-0.01, 0.22]) to the criterion value from our sensitivity power analysis (Cohen's $d_z = 0.23$) to determine whether the observed effect was “inferior” to that planned minimum effect (Lakens et al., 2018). Because the upper bound of the confidence interval was smaller than 0.23, the observed difference between truth ratings for the concrete and abstract statements was statistically inferior to a positive effect of Cohen's $d_z = 0.23$.

The same analysis conducted on the revised set of 52 items – replacing items that were close for Swiss participants with new items that were close for the UK or USA participants (H1) — revealed a pattern that was similar to that for the original 52

items: Average ratings for concrete items ($M = 3.65$, $SD = 0.42$) were again about the same as those for abstract items ($M = 3.63$, $SD = 0.39$), $t(219) = 0.95$, $p = .170$, BF_{10} (with $r_{scale} = 0.34$) = 0.24. The Bayes Factor shows that our observed difference is 4.23 times more consistent with the null hypothesis of no difference than with a distribution centred at the original effect size.

Given that the t-test was not statistically significant, we again compared the upper confidence bound around the observed effect (observed effect: Cohen's $d_z = 0.06$; 95% CI: [-0.05, 0.17]) to the criterion value from our sensitivity power analysis (Cohen's $d_z = 0.23$) to determine whether the observed effect was “inferior” to that planned minimum effect. Because the upper bound of the confidence interval was smaller than 0.23, the observed difference between truth ratings for the concrete and abstract statements in the revised set of items was statistically inferior to a positive effect of Cohen's $d_z = 0.23$.

2.3.2.2 Secondary exploratory analyses. As in Experiment 1a, we tested whether a match between proximity and concreteness increased truth ratings by comparing difference scores (*Concrete – Abstract*) between the original (distant) and replacement (close) items (H2). Partially consistent with the prediction that a match between proximity and concreteness would increase truth judgements, the difference between concrete and abstract was positive for the close items ($M = 0.05$), but it was also positive for the distant items ($M = 0.06$), and near zero in both cases, $t(219) = -0.26$, $p = .603$.

2.4 Known Differences from the Original Study

The instructions, measures, and procedures were adapted directly from those of the original study. The original study was conducted in German at the University

of Basel in Switzerland, whereas our study was conducted in English at universities in the UK and USA. The first author of the original study reviewed the translated statements and agreed that the procedures should work with our populations. Upon realising that truth value and concreteness were not fully crossed in the original study design, we exchanged the concrete and abstract versions of two items across sets A and B to ensure that each set had equal number of items for each combination of true/false and concrete/abstract. Our primary analysis combined across sets, and there is no theoretical reason to expect this change to affect the outcome. Participants in the original study were all undergraduate psychology students who received course credit. Our sample in the USA also consisted of undergraduate psychology students who received course credit or extra credit for their participation. Our sample in the UK was composed of undergraduates from psychology and also included some masters students. For the UK sample, participants had a chance to win one of three £50 prizes rather than receiving course credit. This compensation was commensurate with that used by the original authors in their Experiment 2 which tested the same hypothesis and used the same materials as Experiment 1 (Hansen & Wänke, 2010, p. 1580). We added a check to ensure that participants did not actually know the answers to any questions (see Procedure section).

We included additional, culturally-aligned trivia items to the study (see Materials section). Our participants therefore completed 60 (UK) or 70 (USA) statements rather than 52 in the original study. The Qualtrics platform constrained the presentation format of the statements resulting in more white space between statements than in the original questionnaire. The number of statements presented on each page was identical for our paper-and-pencil and online formats, and

differed from the original study (see Materials section). We discussed these changes in advance with the first author of the original paper, and neither we nor they expected these changes to affect the outcome.

In experiment 1b, data collection occurred online rather than using the paper-and-pencil format of the original study.

2.5 Discussion

In Experiment 1a we attempted to replicate the linguistic concreteness effect from Experiment 1 of Hansen and Wänke (2010) in which participants judged concretely worded trivia items as more probably true than abstractly worded versions (H1). Concrete items were not rated as significantly truer than abstract items for either the original items or the revised set of items, which is inconsistent with the original study. The Bayes Factor for the original set favoured the null - a distribution centred at no effect — over a distribution centred at the original effect size by a 3.45:1 ratio. For the revised set, it favoured the null by a ratio of 2.06:1. For the original items, the upper bound of the confidence interval around the effect was smaller than our smallest effect of interest, and therefore also smaller than the original effect size, meaning that the data were inconsistent with the original finding. Similarly, for the revised items, the data were inconsistent with the original finding. Collectively, these results do not provide evidence for a linguistic concreteness effect on truth judgements.

In Experiment 1b, we extended our test of the linguistic concreteness effect to an online sample. Inconsistent with the original study, concrete items were not rated as significantly truer than abstract items for either the original items or the revised set of items. The Bayes Factor for the original set favoured the null over a distribution

centred at the original effect size by a 1.95:1 ratio. For the revised set, it favoured the null by a ratio of 4.23:1. For the original items, the upper bound of the confidence interval around the effect was smaller than our smallest effect of interest, and therefore also smaller than the original effect size, meaning that the data were inconsistent with the original finding. Similarly, for the revised items, the data were inconsistent with the original finding. Collectively, these results do not provide evidence for a linguistic concreteness effect on truth judgements.

In designing these replications, we consulted the first author of the original study to ensure that our replication matched the procedures necessary to test the original hypothesis and to verify that any changes were consistent with the original conceptualization of the hypothesis. Still, by necessity, some aspects of the design differed between the original study and our replication attempt, and those differences might contribute to the different outcome.

First, our study used English rather than German materials. Although the change in language might contribute, neither we nor Hansen suggested theoretical reasons why translated materials would be ineffective in producing the effect. Indeed Hansen and Wänke's (2010) manipulation was based on the linguistic category model (Semin & Fiedler, 1988, 1991) which was developed based on experiments with English-speaking participants.

Second, in developing the protocol, Hansen suggested the possibility that perceived psychological distance might interact with the experimental manipulation (H2). Consequently, we added additional trivia items intended to match the "distances" of those items for our participants to the distance of the items for the original Swiss participants. Our study showed no effect of this distance manipulation; the close-distant effect was close to zero and numerically in the

opposite direction to the prediction. Although it is possible that adding more trivia items to the original set of 52 might dampen the effect, we saw no difference in the pattern of results for the UK participants (60 items) and USA participants (70 items) in either study. If testing language, perceived proximity, or number of items explain the different patterns of results between our studies and the original study, then the effect might be specific to theoretically uninteresting aspects of the testing context.

Our use of a 6-item response scale maximized the chances of observing an effect because it lacked a neutral mid-point; participants were forced to lean toward true or false for each statement. Consequently, even a small linguistic concreteness effect should nudge participants to make the appropriate directional response, leading to a measurable difference. Using a scale with a neutral midpoint (e.g., 4 on a 1-7 scale) would allow participants to ignore a slight sense of truth or falsity.¹⁹ Future research could consider using a scale with a neutral midpoint. Future studies would also need a substantially larger sample size in order to have adequate sensitivity to measure a much smaller effect.

The aim of the present studies was to accumulate evidence for the reliability of the linguistic concreteness effect and provide a robust estimate of its size for use in subsequent studies. Our experimental design and analyses were planned to optimise the chances of observing the effect: In Experiment 1a we collected data in a setting comparable to that of the original study and used paper/pencil materials matched as closely as possible to the original study. Experiment 1b adopted those

¹⁹ Hansen and Wänke (2010) reported no difference in average ratings for true and false items. Across our studies and conditions, a post-hoc analysis showed that true statements were rated slightly higher than false statements (less than 0.20 rating points on average), regardless of whether or not we excluded items that participants claimed to have known. This small difference is difficult to interpret, but it is consistent with a slight bias to respond on the larger end of the scale (toward true) coupled with some limited sense about the truth or falsity of items even when participants did not know the answer.

materials for online testing with a broader population using Prolific. Each study had greater than 95% power to detect an effect half the size of the original, and each produced evidence more consistent with the absence of an effect than with the original effect. Across these two studies, our analysed sample (466) was approximately ten times the size of the original study ($n=46$). Although no single study is definitive about the existence of an effect, our studies raise doubt about the reliability of using concrete/abstract language as a way to manipulate the judged truth of trivia statements.

Chapter 3: A Reproducible Systematic Map of Research on the Illusory Truth Effect

Abstract

Background: People believe information more if they have encountered it before, a finding known as the *illusory truth effect*. But what is the evidence for the generality and pervasiveness of the illusory truth effect? Our preregistered systematic map describes the existing knowledge base and objectively assesses the quality, completeness, and interpretability of the evidence provided by empirical studies in the literature.

Methods: A systematic search of 16 bibliographic and grey literature databases identified 93 reports with a total of 181 eligible studies.

Results: All studies were conducted at Western universities, and most used convenience samples. Most studies used verbatim repetition of trivia statements in a single testing session with a minimal delay between exposure and test. The exposure tasks, filler tasks, and truth measures varied substantially across studies, with no standardisation of materials or procedures. Many reports lacked transparency, both in terms of open science practices and reporting of descriptive statistics and exclusions.

Conclusions: Systematic mapping resulted in a searchable database of illusory truth effect studies (<https://osf.io/37xma/>). Key limitations of the current literature include the need for greater diversity of materials as stimuli (e.g., political or health contents), more participants from non-Western countries, studies examining effects of multiple repetitions and longer intersession intervals, and closer examination of the dependency of effects on the choice of exposure task and truth measure. These gaps could be investigated using carefully designed multi-lab studies. With a lack of

external replications, preregistrations, data and code, verifying replicability and robustness is only possible for a small number of studies.

Keywords: illusory truth effect, repetition, truth judgement, systematic map, transparency, Registered Report

3.1 Introduction

“Sixty-two thousand four hundred repetitions make one truth.”

-- Aldous Huxley, *Brave New World* (p. 46)

With this satirical statement, Huxley highlights the power of repetition to manipulate belief. Repetition can increase subjective truth judgements, a phenomenon known as the “illusory truth effect”. The effect of repetition on belief occurs for both true and false statements (Brown & Nix, 1996), for both plausible and implausible ones (Fazio, Rand, & Pennycook, 2019), and for both known and unknown information (Fazio, Brashier, Payne, & Marsh, 2015). It appears with only minutes between repetitions (Unkelbach & Warrens, 2018), and with delays of weeks (Gigerenzer, 1984) and even months (Brown & Nix, 1996). Although most studies use sets of trivia statements, it apparently works for consumer testimonials (Roggeveen & Johar, 2002), statements of opinion (Arkes, Hackett, & Boehm, 1989), and false news stories (Polage, 2012). If the illusory truth effect truly generalizes beyond the lab, it might help explain the use of repetition to override facts in propaganda campaigns (Lewandowsky, Stritzke, Oberauer, & Morales, 2005; Paul & Matthew, 2016; Pennycook, Cannon, & Rand, 2018). By the same token, it seems that information can enter the public lexicon through repetition rather than accuracy. Familiarity can apparently trump rationality. But what is the evidence for the generality and pervasiveness of the illusory truth effect?

Over the past few years, awareness of the illusory truth effect has grown, with articles in *Vox*, *The Atlantic*, and *Wired* (Dreyfuss, 2017; Paschal, 2018; Resnick, 2017) linking it to “fake news”, “truthiness”, and President Trump’s communication style. Yet the only meta-analytic review of this literature appeared in 2010 (Dechêne,

Stahl, Hansen, & Wänke, 2010). It combined the results of 51 studies conducted before 2008, and it estimated a medium effect size: ($d = .39$; 95% CI: [0.30, 0.49]) within-items, $d = .50$; 95% CI: [0.43, 0.57]) between-items, random effects model). The meta-analysis is somewhat dated, both because new studies have been published and because it was completed prior to recent advances in techniques used to address publication bias.

Publication bias is prevalent in psychology. Approximately 95%²⁰ of published articles contain statistically significant confirmation of the stated hypothesis (Fanelli, 2010; Sterling, Rosenbaum, & Weinkam, 1995). Synthesizing the results from a biased pool of research, dominated by significant, “positive” findings, threatens the validity and interpretation of results, and in meta-analyses it also makes the overestimation of effect sizes likely (Renkewitz & Keiner, 2019). Although Dechêne et al. (2010) note that a funnel plot for the analysed studies appeared symmetrical, their article did not include the funnel plot or any formal analyses of it, and it is possible that other bias correction approaches would estimate a smaller effect.

We originally preregistered a plan to conduct an updated meta-analysis of the illusory truth effect (<https://osf.io/j6fmr/>). As part of the pilot testing in that plan, intended as a first stage to help develop an appropriate coding scheme, the first and third authors, along with an additional coder each independently coded a random selection of papers from those included in the 2010 meta-analysis. It quickly became apparent that these papers did not report sufficient information to estimate the observed effect size for the illusory truth effect without making strong, questionable

²⁰ This figure likely also reflects HARKing, p-hacking, and other questionable research practices that can occur prior to article submission.

assumptions. For example, the selected papers did not consistently report inferential statistics for the main effect of repetition (the illusory truth effect), included no variance estimates, and/or obscured the effect of interest by combining groups into a more complex analysis²¹. Dechêne et al. (2010) encountered the same issues of underreporting and described the assumptions they made in order to address them in their meta-analysis:

“Twenty-one studies provided standard deviations for the reported means; seven studies reported a range of standard deviations. In the latter case, we computed the pooled standard deviations from the range. Where no standard deviations were provided [23 studies, 45% of the sample of studies], we chose to impute the pooled standard deviation from an overall estimate that was obtained from those studies in which standard deviations were reported or could be extracted”. (Dechêne et al., 2010, p.243; text within brackets added)

The extent of the issue was unclear, though, because the paper did not specify the number of effects that required imputed variance estimates.

In our view, these assumptions cloud conclusions about the overall strength and consistency of the evidence for the illusory truth effect. Imputing estimates of variance when computing standardized effect sizes is suboptimal for at least two reasons: First, it is possible that the subset of studies that do report information about variance differ systematically from those that do not. For example, the studies that report variance might have been more rigorous and precise in their measurement practices, leading to smaller variance estimates and larger standardized effects. If so,

²¹ The results of the pilot coding are available at <https://osf.io/jd72s/>.

using their variance estimates for other studies would yield inflated overall effect estimates. Second, studies with different designs may not have similar variance estimates. For example, variance estimates will differ with the number and type of experimental items and the breadth of the scale used to measure truth ratings (e.g., dichotomous, 1-6, or a continuous response slider). Unfortunately, Dechêne and colleagues could not provide us with the coded data that were used to produce their 2010 meta-analytic estimates, and many of the studies included are old enough (~30 years) that the original data are unavailable. Based on our coding attempt, the lack of available data, and the need to make overly strong assumptions in order to estimate effects for many of the published papers, we concluded that a valid meta-analysis is not possible for the entirety of this literature.

Given these challenges, we chose instead to create a systematic map; a method of evidence synthesis designed to assess the nature of a literature base (Haddaway et al., 2019). The primary objective of a systematic map is to locate and catalogue the breadth of evidence on a particular topic using predetermined, transparent, and reproducible methods (Haddaway, 2018). Systematic maps can thereby answer questions such as: how many studies have been conducted? Which methods were used? What is the mean sample size used? The output from a systematic map is an accessible, searchable database(s) that can then be used by the research community. Specifically the database can be used to highlight knowledge clusters, knowledge gaps, areas with limited or weak evidence (Corker, 2018), or investigations of particular combinations of variables. Systematic maps differ from systematic reviews or meta-analyses in that they do not attempt to answer specific questions about the effectiveness of an intervention, the truth or falsity of a hypothesis, or to estimate effect sizes. Rather, systematic maps have an open framing

that allows a wider range of evidence to be summarised in the database (James, Randall, & Haddaway, 2016). For a comparison of systematic maps and systematic reviews see James et al. (2016) Table 1. Systematic mapping is particularly useful for domains with a wide range of experimental manipulations (e.g., different delays, different types of items) tested in a wide range of contexts and with different measures (James et al., 2016). We created two interrelated databases: an abstract-level database that includes relevant articles where the full text could not be obtained, and an extensively coded full-text database.

In addition to producing a traditional systematic map, we assessed the transparency and reproducibility of the empirical studies identified by the map. Transparency and reproducibility are the cornerstones of the scientific method and knowledge generation. Recent concerns about poor transparency and low reproducibility have catalysed open practices and reforms designed to enable more transparent science (Munafò et al., 2017; Nosek et al., 2015). Meta-research has begun to evaluate adoption of reforms across broad areas, for example in social sciences research (Hardwicke et al., 2020). Here we assess the statistical, methodological, and reporting practices that may impact the robustness of conclusions that can be drawn from the entirety of a single research area. We coded a number of indicators of transparency and reproducibility. For example, the availability of raw data, the provision of which allows computational reproducibility. We also coded whether the main effect of repetition was reported as observed/significant marginally significant/non-significant by the authors as a proxy measure for publication bias; a published literature without bias towards significant results should be characterised by a mix of both significant and non-significant results. A full list of the variables coded is detailed in Table 2.

3.1.1 Research Aims

Illusory truth effect research typically follows a standard paradigm: Participants first read or hear a number of statements, normally trivia statements, during an exposure phase. At test, participants judge the truth of a set of statements generally comprised of half old statements (repeated from the exposure phase) and half new statements (previously unseen). However, these studies can vary in a number of ways. For example, they might measure the truth effect as the difference in truth ratings from exposure to test phase (within-items), or as the difference in truth ratings between new and repeated statements at test (between-items). At exposure stage, they might ask participants to simply read the statements (Unkelbach & Rom, 2017), or rate them for familiarity (Garcia-Marques, Silva, & Mello, 2017). They might test clinical or non-clinical populations, use one or multiple repetitions, or introduce no delay or a long delay between repetitions.

The primary aim of this research was to systematically identify and map published and unpublished research examining the relationship between repetition of statements and subjective truth ratings with the following objectives:

1. Describe the current nature and extent of the literature on the topic
2. Assess the transparency and reproducibility of the literature (using the objective measures described below)
3. Collate and highlight any well-represented subtopics (e.g., studies that use trivia statements as stimuli) that might benefit from more detailed secondary research (knowledge clusters)
4. Identify knowledge gaps in the evidence base (i.e., areas that have not been frequently studied)

5. Provide direction for novel research or single/multi-lab replication studies in which outstanding questions can be empirically tested
6. Produce a systematic map that is transparent, reproducible, and open so that it may be used and updated by others (Lakens, Hilgard, & Staaks, 2016)²²

3.2 Method

3.2.1 Conformance with Reporting and Quality Standards

In preparing the systematic map protocol, we adhered to the Preferred Reporting Items for Systematic reviews and Meta-Analyses Protocols (PRISMA-P; Moher et al., 2015; Shamseer et al., 2015) and the RepOrting standards for Systematic Evidence Syntheses (ROSES; Haddaway et al., 2018b).²³ The completed ROSES form for systematic review protocols is available at <https://osf.io/ux2vz/>. In our reporting of systematic searches, we followed the PRISMA-S extension for the reporting of systematic review searches (Rethlefsen et al., 2021). The meta-data is open with fully reproducible analyses; coded data files, analysis scripts, and supplementary materials available at <https://osf.io/dm9yx/>.

3.2.2 Search Term Identification and Selection

All steps in this search term identification and selection section were completed prior to submitting the stage 1 Registered Report. We defined our search terms with the assistance of the R package litsearchr (Grames, Stillman, Tingley, & Elphick, 2019). Litsearchr reduces bias in keyword selection by partially automating the selection process (Grames, Stillman, Tingley, & Elphick, 2019). Litsearchr uses a

²² Although this paper is concerned with meta-analyses, the recommendations remain relevant for systematic mapping.

²³ ROSES was developed for systematic reviews and maps in the field of conservation and environmental management but can be applied in the current context without the need for adaptation.

keyword extraction algorithm to locate potential keywords from a sample of papers and combines them with author and database tagged keywords to create a list of potential keywords. Important keywords are identified from their predominance in a keyword co-occurrence network.

3.2.2.1 Scoping search. First, a scoping search was conducted using Scopus and Web of Science and the below search string. Searches were conducted on 14 June, 2019 with no date restrictions. The number of hits were as follows: Scopus (156), Web of Science (63).

“illusory truth” OR “illusory truth effect” OR “illusions of truth” OR “reiteration effect” OR “repetition induced truth effect” OR “repetition based truth effect” OR “truth effect” OR “truth judgment”)

3.2.2.2 Litsearchr. The results of the scoping search were imported into R. N-grams that occurred at least three times in the dataset and in a minimum of three studies were extracted and coded as relevant/irrelevant to the search. The same process was followed for similar terms. The litsearchr code and resulting files can be found at <https://osf.io/hdtgb/>. We incorporated the additional terms identified by the litsearchr package, along with relevant unigrams into the search string.

3.2.2.3 Testing the comprehensiveness of the search. To estimate the comprehensiveness of the search, we compiled a set of 20 papers of known relevance to the review to serve as a benchmark list (see Appendix A). We conducted a scoping search, using Web of Science and Scopus, to ensure that all 20 papers were indexed and captured by the search terms. For any papers that were not initially found, we identified the reasons why they were missed, adjusted the search string accordingly, and checked that the string now captured those papers. The search string below is shown as formatted for Web of Science (exact search strings by database are documented in Appendix B):

((("illusory truth" OR "illusion* of truth" OR "induced truth effect" OR "reiteration effect" OR "tainted truth effect" OR "repetition based truth effect" OR "repetition induced increases" OR repeat OR repeated OR repeating OR repetition OR "prior exposure") AND (true* OR truth OR "truth effect*" OR belief) AND (statement* OR items OR stimulus OR stimuli OR claim* OR judgment* OR judgement* OR rating* OR "subjective truth" OR "truth value" OR "judged validity" OR "validity ratings" OR "processing fluency" OR "fluency effect*" OR "perceptual fluency"))))

3.2.3 Search Strategy

A summary of the workflow for our search strategy can be found at <https://osf.io/f9462/>. Using the predefined search string we carried out an extensive literature search that aimed to minimise the effect of publication bias on our map. Considerable effort was devoted to searching for both published and unpublished studies, as well as replications. We consulted an academic librarian for advice on the details of our scoping search terms and search strategy. The electronic searches were conducted by the first author on the 4th and 6th of February 2020 without any limits or restrictions. Any articles published after that date were not be included.²⁴ We preregistered that if the review took more than two years to complete, we would update the searches. All searches and outcomes were recorded in a Search Record Appendix (<https://osf.io/xsnhm/>). Table 1 outlines further details of the fields used for each search.

3.2.3 1 Electronic bibliographic database searches. First, a comprehensive computerized search of illusory truth studies was performed using the above search string in eight bibliographic databases/platforms. This selection of databases includes

²⁴ We hope that others will continue to update the database.

all seven of those used in the Dechêne et al. (2010) meta-analysis (see Table 1 and Appendix B).

3.2.3 2 Grey literature searches. Furthermore, we included grey literature by searching for items such as doctoral theses, conference papers, preprints, and replication attempts in eight databases (see Table 1). Google Scholar has been identified as effective in retrieving grey literature (Haddaway, Collins, Coughlin, & Kirk, 2015) and was used to supplement the other search methods. To increase reproducibility we used Publish or Perish (Harzing, 2007) to carry out searches and export the results. Because Google Scholar allows only basic Boolean operators in search strings, the search string was reduced to the key components detailed in Appendix B. Search strings for all other grey literature sources are also documented in Appendix B.

Table 1

List of bibliographic and grey literature databases/platforms searched along with the search fields used

	Type	Database	Field	Comments
1	Bibliographic	Business Source Premier (EBSCOHost)	“Abstract or author-supplied abstract”	Using “Advanced Search”
2	Bibliographic	EconLit (EBSCOHost)	“Abstract”	Using “Advanced Search”
3	Bibliographic	ERIC (EBSCOHost)	“Abstract”	Using “Advanced Search”
4	Bibliographic + Grey	Google Scholar	“The phrase”	Accessed via Publish or Perish
5	Bibliographic	PsycINFO (Ovid)	“Abstracts”	Using the “Advanced Search”
6	Bibliographic	PubMed (NCBI)	“Title/Abstract”	Using “Advanced Search Builder”

7	Bibliographic	Scopus (Elsevier)	“Article title, Abstract, Keywords”	Using “Advanced Search”
8	Bibliographic	Web of Science	“Topic”	Using “Basic Search”
9	Theses & conference papers	OpenGrey		
10	Preprints	PsyArXiv (OSF Preprints)		
11	Replications	Curate Science		
12	Replications	PsychFileDrawer		
13	Theses	DART-Europe		
14	Theses	EthOS (British Library)		
15	Theses	ProQuest Dissertation & Theses Global (ProQuest)		
16	Theses	Thesis Commons (OSF Preprints)		

Note. The interface or platform through which the database was searched is in parentheses. The Web of Science platform was used to search the following collections: Web of Science Core Collection, BIOSIS Citation Index, BIOSIS Previews (until 2008 only), KCI-Korean Journal Database, MEDLINE, Russian Science Citation Index, SciELO Citation Index.

3.2.3 3 Researcher-to-researcher channels. Upon completing the electronic searches, we issued calls for unpublished studies through the Listservs of the Society for Personality and Social Psychology (SPSP), the Society for Judgment and Decision Making (JDM), Psychonomic Society, Cognitive Science Society (CSS), European Association of Social Psychology (EASP), and the Society for Consumer Psychology (SCP)²⁵. We issued one call per society. If the calls led to direct correspondence with a researcher, we asked them to send us any (other) unpublished studies directly. Simultaneously, we posted notices on Twitter (twice each week, for 3 weeks) and included a link to a public Google document to allow researchers to suggest additional citations.

²⁵ We preregistered issuing a call through the Academy of Marketing Science (AMS). We attempted to contact AMS three times but received no response.

Finally, once eligible papers from the database and grey literature searches had been identified through full-text screening, we contacted corresponding authors for any preprints, or unpublished studies/papers that they were aware of and any published studies we might have missed. We used the email address provided in the paper. If the email was returned undelivered, we searched online for a current email address. If none could be found, we tried to reach the other authors. If authors did not respond to the initial email, 2 weeks later a second email offered the chance to provide unpublished studies anonymously using a file transfer service. We did not send further request emails. A record of the correspondence (who was contacted, on which date, the general nature of the response) was retained. We kept this record private but report the response rates. The wording of emails and the Listserv message can be found at <https://osf.io/52c4q/>.

After initiating an email correspondence with a researcher, either as a corresponding author who might have unpublished studies or as a response to a Listserv contact, we allowed 10 weeks (from the date of the first email) to receive studies from them. Even where relevant studies were received after 10 weeks, we were able to include them in the map.

As a result of our calls, three authors contacted us via Twitter and four authors responded to Listserv messages. Of the remaining 46 first authors, we were able to contact 32, and 23 responded. From this correspondence, eight authors offered potentially relevant papers, resulting in 21 additional papers that eventually were included in the map.

3.2.3 4 Manual searches. Once relevant meta-analyses and review articles had been identified during title/abstract screening, their reference lists were manually screened for supplementary papers (i.e., *backward search*). Upon completion of full-

text screening, we also manually reviewed the bibliographies of the eligible papers for any additional studies that had not been captured by the database searches.

Additional papers identified via manual searches were screened at the full-text level.

Reproducibility of unpublished studies. Unpublished studies pose a threat in terms of reproducibility and the cumulative updating of a systematic map. For any unpublished studies we received, we asked the author's permission to share the unpublished report/data/summary. In all cases, authors agreed either to share the whole report or a summary.

3.2.4 Inclusion Criteria

Since systematic maps are designed to give an overview of the topic area, they adopt broad inclusion criteria. We included articles that adhered to all of the following criteria:

1. **Population:** human populations of any age, including those from clinical groups
2. **Intervention:** verbatim or gist repetition of multiple statements (e.g., trivia, political, marketing) presented visually or aurally
3. **Comparator:** within-subjects (repeated vs. non-repeated statements), or between-subjects (non-repetition control vs. repetition group)
4. **Outcome:** numerical (Likert-type scale, slider, or similar) or binary (true/false) measures of subjective truth judgements, either comparing truth ratings made before and after repetition (within-items), truth ratings for new vs. repeated items (between-items), or non-repetition control vs. repetition group (between-subjects)
5. **Study type:** empirical quantitative studies
6. **Time frame:** no constraints

3.2.5 Exclusion Criteria

At the title and abstract screening stage, excluded papers were simply marked as “no”. At full-text screening stage a list of excluded articles is reported along with a specific reason from the list below. The list illustrates the sequence in which exclusion criteria were applied. Therefore, if an article could have been excluded for multiple reasons, we required that only one reason be given (i.e., the first criterion at which it fails). We excluded studies for the following predefined reasons:

1. **Population:** non-human population
2. **Study type:** review paper ²⁶
3. **Study type:** no quantitative data
4. **Intervention:** the study did not use repetition as a manipulation to increase subjective truth judgements
5. **Outcome:** the study did not measure subjective truth judgements
6. **Comparator:** the study did not compare ratings for repeated vs. non-repeated statements, or ratings from a non-repetition control group with those from a repetition group
7. **Other:** entirely superseded by a later paper. Multiple reports of the same study were-collated into a superset and coded as one unit. Papers were only excluded where it was clear that the earlier version contained no additional information. Specifically this refers to cases in which a study described in a preprint, dissertation, stage 1 Registered Report, or conference abstract/presentation was fully reported in a later paper. In cases of partial overlap (e.g., a paper that reports only 3 of the 4 studies included in a

²⁶ Review papers were searched as detailed under “Manual searches”.

dissertation), the reports were connected in the database to ensure that all studies were coded

8. **Other:** any dataset that was not accompanied by descriptive meta-data detailing the methods used to test the illusory truth effect (e.g., unpublished data received via contact with authors) were excluded from the full-text database because the information needed to code the study was missing. However, it was included in the abstract-level database
9. **Other:** the paper was written in a language other than English or French and a translator could not be recruited

In addition to the above preregistered exclusion criteria, if an abstract was incomplete during the title/abstract screening stage and we subsequently retrieved the full abstract, we first reviewed that complete abstract during full-text screening and if it was excluded, we coded it using the additional criterion, “screened abstract - not relevant.”

3.2.6 Study Screening Procedure

The publications returned from the electronic searches were imported into Zotero. Duplicate references were identified and removed using Zotero’s “duplicate items” feature based on title, DOI, and ISBN fields. In cases of dual publication (e.g., a conference paper or PhD thesis later published in a peer reviewed journal), we extracted the superset of studies in case each had content that the other did not. For the purpose of maintaining records, we kept a comprehensive list of all references before duplicates were removed.

The deduplicated records were then imported into *Covidence*, Cochrane’s online systematic review tool that facilitates collaborative screening. We followed a two-stage screening process: Initially two coders independently screened the titles

and then the abstracts using Covidence and the predefined inclusion and exclusion criteria. Studies were coded as 1) yes, 2) no, or 3) maybe. A paper was coded “maybe” if insufficient information was available to enable an eligibility decision or if there was doubt about the presence of an inclusion criterion. In this case, the paper was retained and a decision made at the full-text stage. Screening decisions were compared using Cohen’s Kappa. Scores of 0.64 (ELH and DJS) and 0.60 (ELH and FVT) were obtained, indicating substantial agreement. Covidence highlights any discrepancies in a section called “resolve conflicts”. Any conflicts were reviewed by the first author and resolved by discussion with the relevant coder.

We then retrieved the full text of each paper. Each article was downloaded in PDF format from whatever source was available (e.g., journal website, interlibrary loan, author website, email to the corresponding author, British Library). If the full text was unavailable, the article was still coded, but in the abstract-level database only. Once full texts had been retrieved, coders independently used Covidence to apply the inclusion and exclusion criteria based on a brief evaluation of the full text. The Cohen’s Kappas for full-text screening were 0.94 (ELH and DJS), and 0.61 (ELH and FVT). Any disagreements about either the inclusion/exclusion decision or the reason for exclusion were discussed between the two coders, and any remaining disagreements were adjudicated by the remaining coder. A record of full-text evaluations is available at <https://osf.io/xsnhm/>. Once full-text eligibility screening was complete, we carried out the additional manual searches of bibliographies and contacted corresponding authors, as detailed in the search strategy section.

We used the ROSES flow diagram for systematic maps (Haddaway, Macura, Whaley, & Pullin, 2018a) to report the flow of articles through all stages of the process from searching to synthesis for the systematic map (Figure 1).

3.2.7 Map Coding and Interrater Reliability

Two interrelated databases were created in Excel files. The abstract-level database includes articles that appear to be relevant but where the full text could not be obtained. These articles were coded for bibliographic information only. To produce the full-text database, we extracted data from full-text articles using the coding scheme outlined below (see Table 2). If multiple studies were reported within one article, each study was coded on a separate line. Studies included only in appendices or described as pilot data were coded and flagged when enough information was provided to do so.

The coding scheme was split into article-level (Table 2, codes 1 – 33) and study-level codes (Table 2, codes 34 – 74). Data entered at the article-level included information such as citation count, study language, and the reporting of open research practices. None of the article-level codes required a judgement call, and the first author single-coded them.

At study level, initially we independently double-coded 30 papers. Each author coded 10 papers with each other coder, resulting in 20 papers coded by each author. Papers were randomly chosen by executing the below commands in R:

```
set.seed(123)
sample(112, size = 30, replace=FALSE)
```

The first 10 of these papers were coded by DJS and ELH, the next ten by DJS and SJW²⁷, and the remaining 10 by ELH and SJW. After the coding was complete, we identified all disagreements and jointly evaluated whether they resulted from ambiguities in the coding instructions or from coding errors. For any cases of

²⁷ The original second author of the Stage 1 Registered Report (FVT) withdrew from the project and was replaced by SJW. Thus, the second author was an external, independent coder who joined after Stage 1 IPA had been received.

ambiguity, we reviewed the coding *instructions* and adjusted them. Each pair of authors then coded those previously ambiguous variables using the adjusted instructions for an additional set of five randomly selected papers. Where disagreements on interpretation remained, we repeated this process and coded a new set of five papers. This process iterated until the authors reached 100% agreement that the coding instructions were unambiguous and that they led to consistent coding (i.e., codes for dropdown menus exactly matched and codes for free text variables other than “notes” columns semantically matched. The changes to the coding instructions during this iterative process were documented and are reported at <https://osf.io/a9mfq/>).

Once 100% agreement was reached on the final set of coding instructions, the second author coded 20 additional papers, and all of the remaining papers were coded by the first author. By reducing ambiguity we aimed to make our coding scheme as reproducible as possible. Even so, no coding scheme is perfect for every paper, and cases that the coder felt were ambiguous were discussed with either of the other authors, depending on availability at the time (such cases are documented in the coding file).

Coders highlighted the text for each coded variable in the article PDF files.²⁸ Highlighted electronic copies of the extracted articles (PDF) have been made as publicly available as possible given copyright restrictions.²⁹ Following data extraction (at stage 2) we approached the publishers (and authors for unpublished work) of all extracted articles to seek permission to archive the highlighted PDFs publicly, on the OSF. Two publishers (Instituto Superior de Psicologia Aplicada and

²⁸ We did not compare highlighting when evaluating the reliability of the coding instructions.

²⁹ Highlighted by the primary coder.

University of Illinois Press) approved the request, but the majority of publishers declined (APA, Elsevier, MIT Press, Oxford University Press, Springer, Taylor & Francis, and Wiley). We received no response from three other publishers (Chicago Press, Guildford Press, Sage). We therefore placed the annotated PDFs in a password protected zip archive which is stored at <https://osf.io/3hzmf/>. The password will be provided upon request.

Table 2 summarises the study characteristics we extracted and coded. We did not contact authors for additional information. The planned coding scheme is detailed in the “codingScheme_stage1RR_2ndrevision” Excel file <https://osf.io/h2e5g/>. We piloted the coding scheme by coding randomly selected papers from the reference section of Dêchene et al. (2010) and iteratively adapting the coding scheme. The pilot was preregistered on the OSF (<https://osf.io/d7tb5>). Additionally, the coding scheme was updated based on reviewer input during review of the stage 1 Registered Report submission. The final coding scheme is available at <https://osf.io/a9mfq/>.

Where possible, our predefined coding scheme used dropdown menus to constrain data entry. For variables that we expected to be idiosyncratic (e.g., retention interval between exposure and test sessions) we entered data as free text. The free text variables are highlighted in Table 2. Once coding was complete, we merged any codes that used different terms for the same content to ensure consistent labelling. We then reviewed the free-text coding to determine whether meaningful clusters could be grouped for simplification. Such groupings are reported in the results section, and in files at <https://osf.io/ebnm5/>.

The following broad categories of data were extracted for coding at either article or study level:

1. Bibliographic information
2. Methodological information about the study design, stimuli, and subjects
3. Information about the number of repetitions and delay between exposure and test phases
4. Study outcome
5. Level of adherence to transparent data reporting and open science practices

Table 2

Summary of study characteristics extracted and coded

Variable	Details/examples	Variable described in text
General Information (article level)		
1-5	Bibliographic information APA citation, Author, Year, Title, Journal	Y (partially)
6	Google Scholar link	N
7	Document type Journal article, PhD thesis, MSc dissertation, conference paper, poster, book chapter, unpublished article, unpublished data, unpublished preprint	Y
8	Publication status Was the study published in a peer-reviewed journal?	Y
9-10	Citation count ELH coded the citation count on a single day using Web of Science and Google Scholar	N
11	Source How was the study or these data first located?	N
12	Subject area What is the broad subject area?	Y
13	Evidence synthesis Has the study been included in a previous evidence synthesis?	Y
14	Retraction Has the paper been retracted? (http://retractionwatch.com)	Y
15	Language In which language is the article written?	Y
16	Number/name of coders Report who coded the study	N
17	Full text ^b Is the full text of the article available?	N
18	Study country Which country is the corresponding author based in according to their affiliation?	Y

19	Number of studies	How many studies does the article report?	N
20	Number of illusory truth effect studies	How many of the studies relate to the illusory truth effect?	N
<hr/>			
Open Research Practices (article level)			
<hr/>			
21	Replication	Does the article claim to report a replication study?	Y
22	Preregistration	Does the article report a study (or some aspect of a study) that was preregistered?	Y
23	Preregistration located	Where does the article indicate the study was preregistered?	N
24	Open data	Does the article state whether or not data are openly available?	Y
25	Raw data	Can you access, download, and open the raw data files?	Y
26	Open analysis scripts	Does the article state whether or not analysis scripts are available?	Y
27	Open materials	Does the article state whether or not materials are available?	Y
28	OSF	Were any additional data files or materials shared on the OSF?	Y
29	Article access	Is the article available open access (using https://openaccessbutton.org/)?	Y
30	statcheck	Can statcheck (http://statcheck.io/) read the PDF?	Y
31	statcheck checked	Report number of statistics checked by statcheck	N
32	statcheck issues _c	Report number of issues highlighted by statcheck	supplement
33	Links	Links to preregistrations, open data, code, or materials	N
<hr/>			
Study Design (study level)			
<hr/>			
34	Experimental aim ^d	Describe the main aim/purpose of the study	N
35	Goal vary ITE	Did the abstract state that the primary goal of the study was to vary the magnitude of the overall ITE effect by varying some factor (moderation/mediation)?	Y
36	Results vary ITE	Did the abstract report finding evidence that the magnitude of the overall ITE varied as a function of a manipulated variable?	Y
37	Overall test ITE	In the abstract, do the authors describe the outcome of the overall test of what they define as the illusory truth effect?	Y

38	Sample size tested ^d	Number of participants tested	Y
39	Sample population	Which population made up the study sample?	Y
40	Study design	Was repetition manipulated within- or between-subjects?	N
41	Design ^d	Describe the overall factorial design of the study	N
42	Within-subjects factors ^d	Describe the within-subjects factors and groups	N
43	Between-subjects factors ^d	Describe the between-subjects factors and groups	N
44	Stimuli type	Type of experimental stimuli	Y
45	Study setting	In which setting was the study conducted (e.g., lab, online)?	Y
Exposure Session(s) (study level)			
46	Stimuli presentation exposure	How were the stimuli presented during exposure phase (e.g., auditory, visual)?	N
47	Repetitions manipulated exposure	Were the number of repetitions manipulated during exposure phase?	Y
48	Number of repetitions exposure ^d	Number of times participants are exposed to statements during exposure phase(s)	Y
49	Tasks exposure ^d	List all tasks completed with the critical items during exposure phase(s)	Y
Retention Interval (study level)			
50	Retention interval ^d	Time between exposure and (each) test phase(s)	Y
51	Filler task ^d	List any task(s) completed during retention interval	Y
Test Session(s) (study level)			
52	Repetition type	Were the statements repeated verbatim or gist?	Y
53	Stimuli presentation test	How were the stimuli presented during test phase (e.g., auditory, visual)?	N
54	Statement mix	At test were all statements repeated, or a mix of old and new?	N
55	Number of test sessions ^d	Number of test sessions (excluding exposure phase(s))	Y

56	Number of repetitions test ^d	Total number of exposures across all test phases	Y
57	Truth measure	Type of truth measure used as the dependent measure	Y
58	Prior knowledge	Does the study test whether participants already knew the answers to test items prior to the study?	Y
Results (study level)			
59	Overall test reported	Do the authors report a single overall test of what they define as the illusory truth effect?	Y
60	Measurement design	How was the overall illusory truth effect measured (i.e., between/within-items)?	N
61	Test statistic ^{d e}	Report the test statistic for the overall effect of illusory truth	N
62	Degrees of freedom ^{d e}	Report degrees of freedom for main effect of illusory truth	N
63	Reported p-value ^{d e}	Report p-value for main effect of illusory truth	N
64	Calculated p-value ^{d e}	Report calculated p-value from statcheck	N
65	Direction of test ^e	Report whether the statistical test was specified as one-sided or two-sided	N
66	Effect size ^{d e}	Report the type and value of the effect size for main effect of illusory truth	Y
67	Confidence interval ^{d e}	Report the confidence/credible interval for the effect size	N
68	Overall test significant ^e	Do the authors report in their prose in the results section that the overall test of illusory truth effect was observed/statistically significant marginally significant/non-significant?	Y
Sample Size & Transparent Data Reporting (study level)			
69	Sample size justification	Does the study report a justification for the choice of sample size?	Y
70	Statistical sampling plan	Does the study report a formal power analysis or Bayesian sampling plan?	Y
71	Exclusions reported ^e	Does the study report where participants, or data within participants, were excluded from analysis?	Y
72	Exclusions number reported ^{d e}	How many participants does the study report as being excluded?	N
73	Means ^{e f}	Does the study report means for critical conditions?	Y

74	Measures of variance ^{e,f}	Does the study report the variance (or SDs) for the means of critical conditions?	Y
----	-------------------------------------	---	---

Note. Y means that the variable is reported in the text of this paper. N means that the variable can be found in the systematic map database. ^b Articles in the abstract-level database were not coded beyond this variable. ^c We reported the statcheck results without further evaluation. Where statcheck was able to read the PDF, summary reports are available on the OSF. ^d Indicates variables coded using free-text rather than dropdown options. ^e Indicates variables that were not be coded if the study did not report a focused test of new vs. repeated statements (i.e., a main effect for repetition). ^f Where inferential statistics were reported. For changes between the Stage 1 approved coding scheme and the final coding scheme please see <https://osf.io/a9mfq/>.

3.3 Results

The analysis script was written in R Markdown (Allaire et al., 2020).

Analyses used R version 4.0.3 (R Core Team, 2019) with the packages `plyr` 1.8.6 (Wickham, 2011) for recoding variables, and `tidyverse` 1.3.0 (Wickham et al., 2019) for data wrangling and visualisation. See the “Data & Analysis” component on the OSF.

3.3.1 Evidence Identification, Retrieval and Screening

The ROSES diagram (Figure 1) summarises the steps involved in this systematic map and the number of articles added or excluded at each stage. The 5,336 potentially relevant results from bibliographic and grey literature searches (4th and 6th of February 2020) resulted in 3,290 results after de-duplication (1,958 detected automatically with Zotero, 60 manually identified, and 28 identified via Covidence³⁰). Of those, 3,104 (94%) were excluded via title and abstract screening. If the abstract had only been partially available during title and abstract screening, then the complete abstract was added prior to the full-text review stage. An additional 25 were excluded based on these full abstracts.

³⁰ We did not preregister the use of manual or Covidence’s deduplication. However both methods picked up duplicates that Zotero missed.

Of the 186 (6%) papers that merited full-text review, 10 were irretrievable and they were coded for bibliographic information only and not incorporated into the results below (they are included in the abstract-level database). Following full text review, 109 (62%) of the remaining 176 papers were excluded (see Figure 1 for reasons), leaving 67 (38%) unique results from the bibliographic search.

The first author then manually reviewed the references cited by those 67 articles as well as by any on-topic review papers that had been excluded. This “backward” search identified 5 additional results, all of which were included. An additional 21 included articles were added from researcher-to-researcher channels (eight from emails to authors, seven from Twitter posts, and six from Listserv posts). After adding these 26 additional results to the 67 identified via bibliographic search, the final full-text systematic map included 93 articles (Appendix C) documenting a total of 181 studies. Researcher-to-research channels yielded two additional references for the abstract-level database, for a total of 12; see <https://osf.io/37xma/>).

The only pre-existing research synthesis (Dechêne et al., 2010) included 25 results, 22 of which were among the 93 articles we had already identified. We were unable to obtain the three additional results that were based on unpublished data.

All 58 of the published articles included in the final map were written in English. Of the 35 unpublished references, two were undergraduate theses written in Spanish and one was a PhD thesis in German. The abstract, methods and results sections of the Spanish theses were translated, and the German PhD included three manuscripts prepared for submission in English.

The full-text database for the 93 articles included in our review is available at <https://osf.io/37xma/> and includes Google Scholar links for each article and citations

counts from both Web of Science and Google Scholar (completed on 17 November 2020). As of 02 October 2020, none of the articles had been retracted.

All 93 articles were run through statcheck (Rife, Nuijten, & Epskamp, 2016) to check for errors in statistical reporting. Statcheck recomputes p-values and compares them to those reported in the text. Inconsistent p-values are recorded as an “error.” If the reported result was significant and the recomputed result was not, or vice versa, the result was recorded as a “decision error.” Of the 57 PDFs that were readable, only 31 had no issues, 26 contained errors, and four of those were decision errors (for a summary see Appendix D; for complete statcheck reports see <https://osf.io/r3cwg/>). However, no errors related to the p-values for the overall effect of illusory truth: For all studies, the value that statcheck recalculated for the critical test matched the one reported in the paper. As preregistered, we did not further evaluate the statcheck results.

3.3.2 Systematic Map Findings

The Stage 1 manuscript was preregistered and is available at <https://osf.io/ar4hm>. Deviations from the accepted Stage 1 are explicitly documented at <https://osf.io/2hcyr/>. We coded a total of 74 variables for each article (see Table 2; full coding of all variables along with coding criteria/instructions are available at <https://osf.io/a9mfq/>). Here we report the variables likely to be of broad interest and most relevant for identifying gaps in the literature. Despite best efforts to avoid error, as with any project of this scale, coding errors may occur. We will maintain an updated version of all tables/figures and the associated database at <https://osf.io/dm9yx/>, and will document any errors, corrections, or comments we receive.

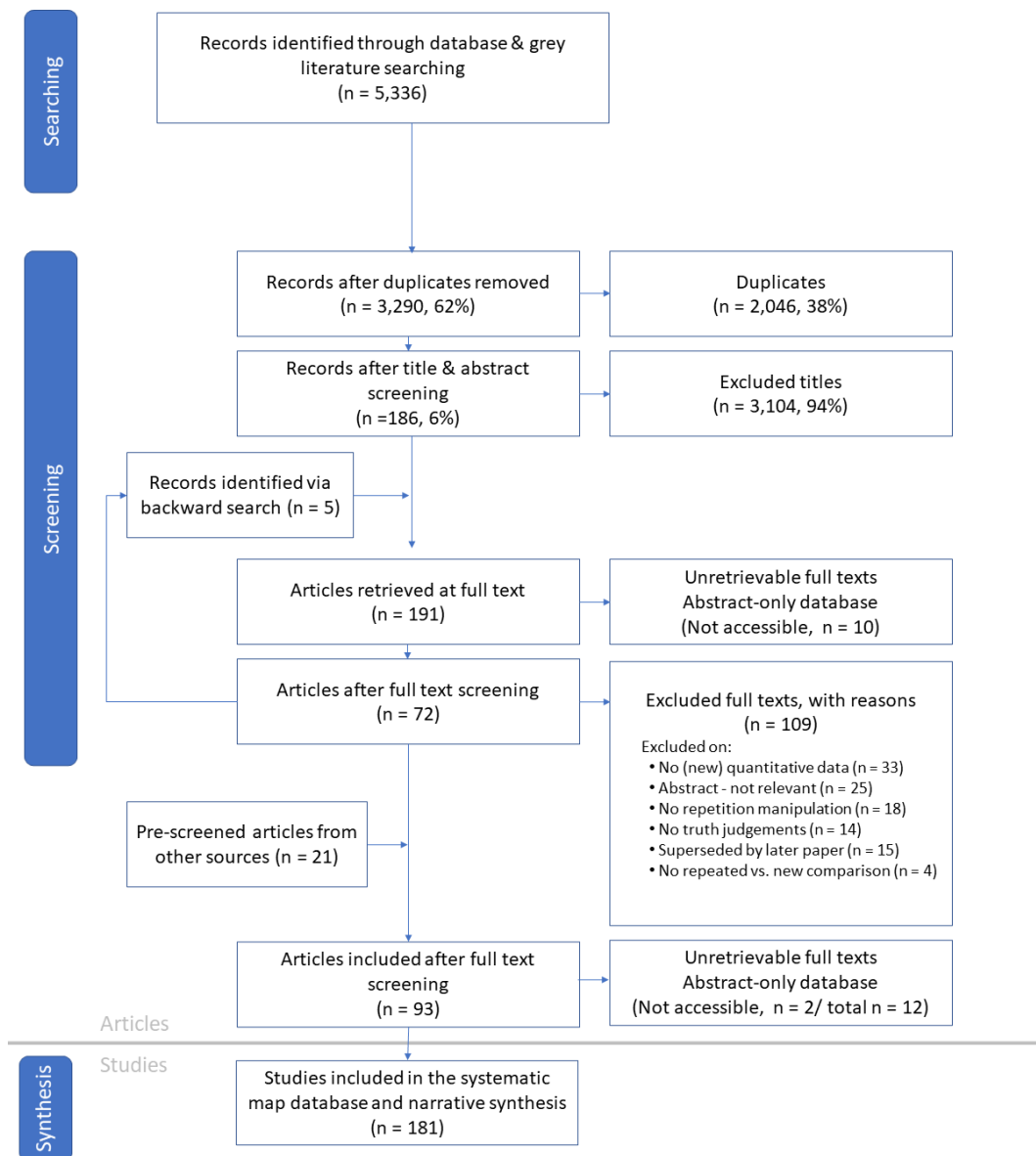


Figure 1. *ROSES flow diagram for systematic maps (version 1.0).*

3.3.2.1 Range of publication types, countries and experimental aims. In

order to understand the breadth of research conducted on the illusory truth effect, in this section we evaluate the range of article types, publication locations and dates, and the overarching aims of the included studies. Table 3 categorises the types of documents included in the map.

Table 3

Types of sources included in the systematic map by publication status

Article type	N
Published	
Peer reviewed journal article	57
Book chapter	1
Unpublished	
PhD thesis	8
Summary	8
Article	5
Preprint	5
MSc dissertation	4
Conference paper	3
UG dissertation	2

The majority of published articles appeared in psychology journals, followed by marketing, neuroscience, and education journals (see Figure 2). We used www.openaccessbutton.org to check whether the 58 published works were available open access. Seventeen were open access and the remaining 41 were behind a paywall.

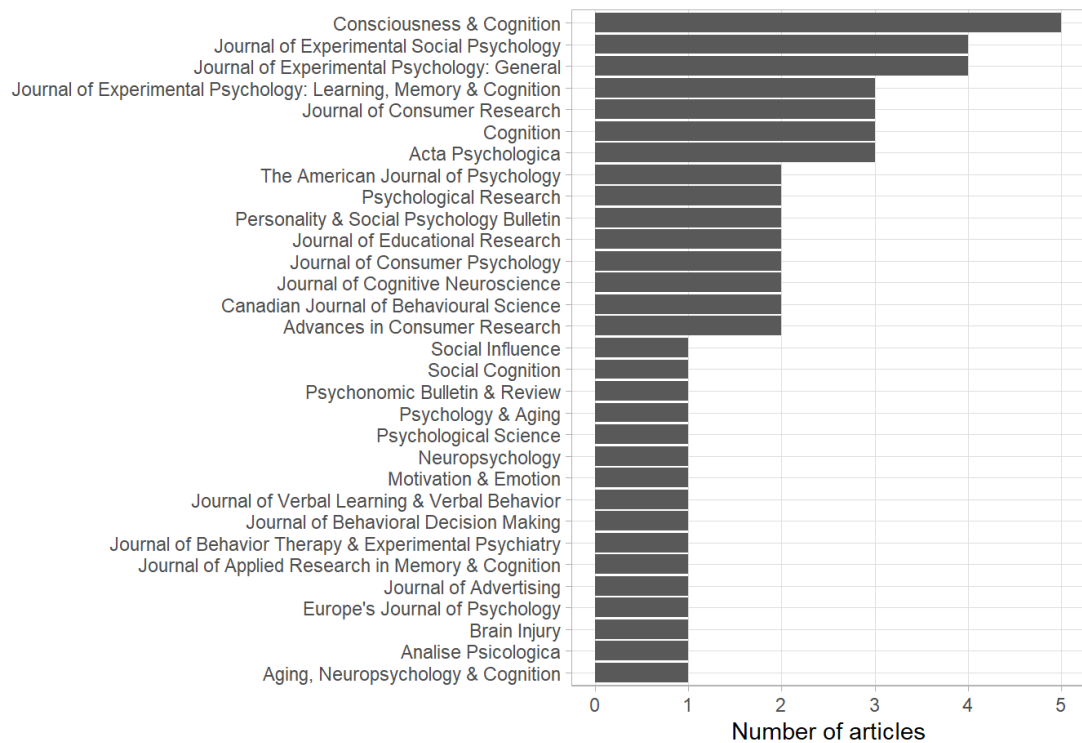


Figure 2. Journals that have published the illusory truth effect articles included in the map.

Since the first paper on the illusory truth effect was published in 1977 (Hasher, Goldstein, & Toppino, 1977), there has been a general upward trend in research on the topic (see Figure 3), with an increase since 2015 (2016: 7 papers; 2017: 6 papers; 2018: 6 papers; 2019: 8 papers; 2020: 15 papers³¹). Of the 93 papers in our map, 54 appeared in 2010 or later.

³¹ Note that electronic searches were conducted at the beginning of February 2020.

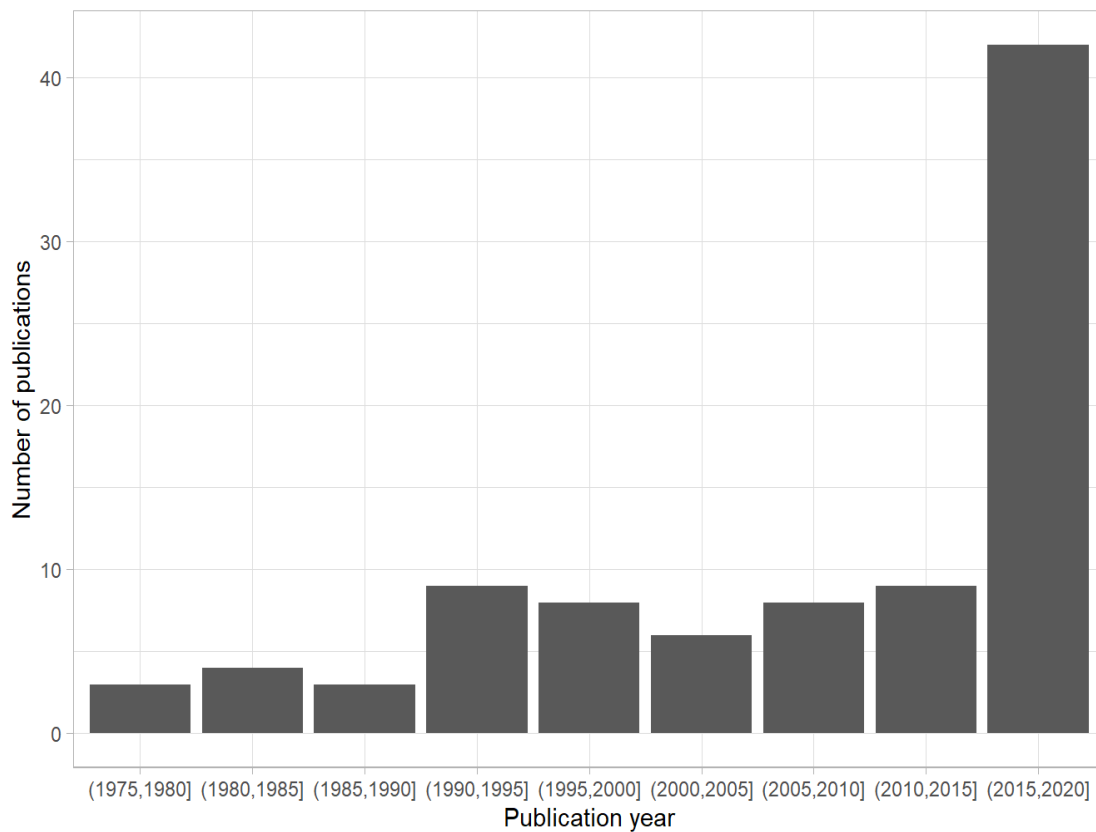


Figure 3. *Date of publication/completion articles included in the systematic map. The figure includes both published and unpublished studies. The square bracket means inclusive and the parentheses means exclusive (e.g., the range (1975 - 1980] excludes 1975 but includes 1980).*

Based on the first author's institutional location, all published studies were conducted in 12 Western countries, with nearly half conducted in the United States (see Figure 4). The lack of any studies from researchers in Asia, Africa, or Latin America appears to be a notable gap in the illusory truth effect literature. Although our exclusive use of English language search terms might have resulted in a sampling bias that missed work by authors from those regions, the vast majority of psychology literature is written in English. This gap warrants further investigation. If there are differences in the illusory truth effect based on culture or other global regional differences, the results in our systematic map cannot inform us about them.

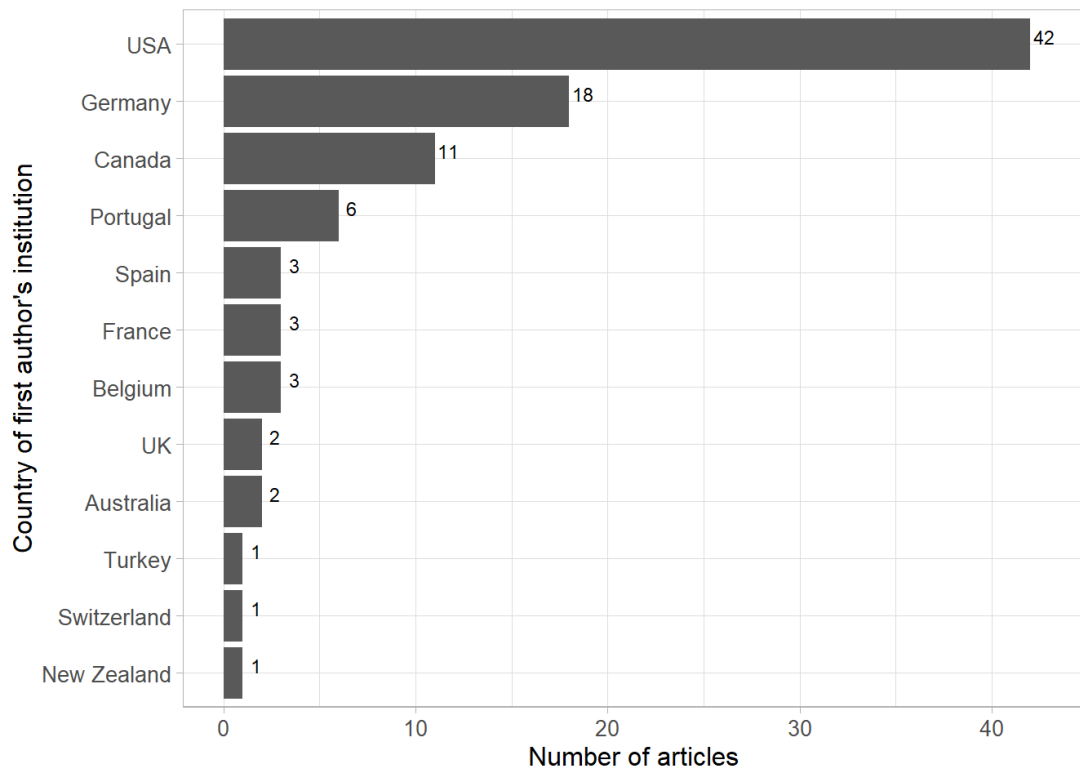


Figure 4. *Number of articles by country of first author's institution included in the systematic map.*

We aimed to code the primary purpose of each study to determine whether measuring the illusory truth effect was the main experimental goal, or whether the goal was to measure variation in the effect. We focused on the abstract to see whether the authors stated an explicit aim and corresponding results. Many studies (36 or 20%) did not specify a clear goal in the abstract. Just 46 (25%) studies described the results of an overall test of the illusory truth. This figure is not surprising given that the majority of studies address issues that assume an overall illusory truth effect exists, and instead focus on variation in other factors.

Many studies (69 or 38%) aimed to examine variations in the magnitude of the overall illusory truth (i.e. moderation or mediation), and 67 (37%) reported finding variation of some sort. However many studies focused on variations for outcomes other than the overall illusory truth effect (43 or 24%).

3.3.2.2 Experimental design, materials, measures and participants. This section evaluates the types of participant groups tested, and the range of conditions and materials used in order to assess the level of standardisation of experimental designs and the generalisability of the effect.

More than half of all studies used a student population (see Figure 5). There was minimal research on harder to reach groups such as clinical populations and younger and older participants, revealing a gap in the knowledge base about the nature of the illusory truth effect in children and older adults.

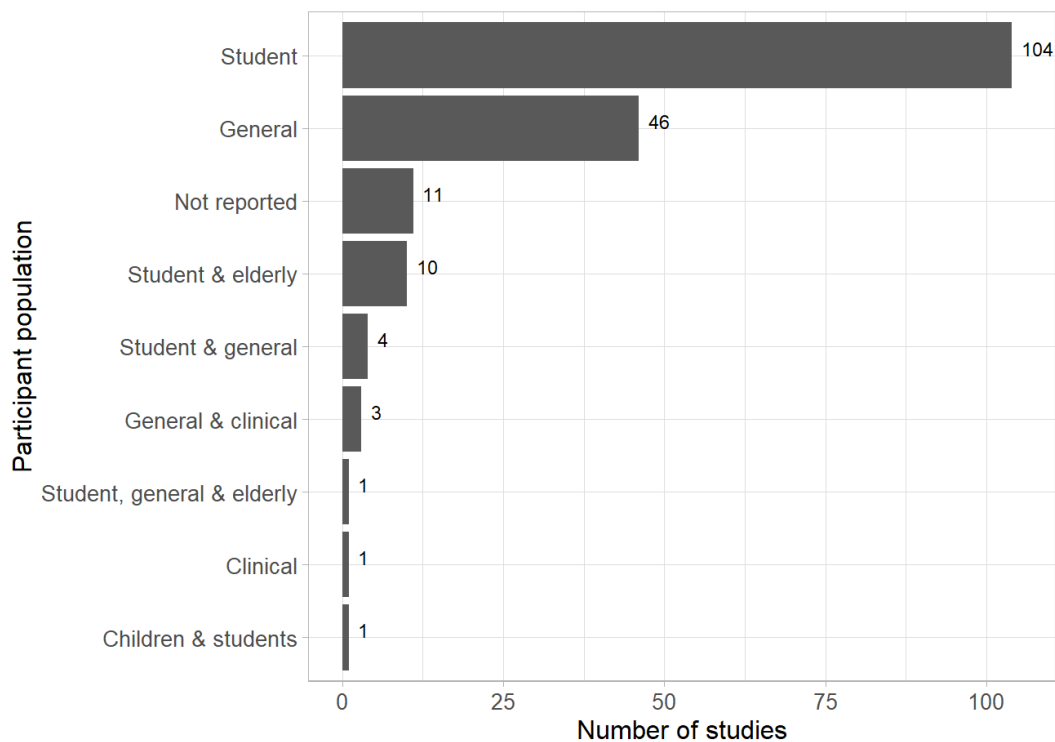


Figure 5. *Frequency and variety of participant populations within the included studies.*

Most studies were conducted in a lab or classroom (116 or 64%), followed by online (48 or 27%). Two studies (1%) were conducted in participants' homes which might represent a more naturalistic, generalisable context in which to measure the effect. Eleven studies (6%) did not report the setting, two studies (1%) used various settings, and we lacked information for two studies (1%).

Studies within the map overwhelmingly used trivia statements (135 or 75%) as the experimental stimuli (see Figure 6). This finding highlights a gap in the evidence that may affect the generalisability of the effect. What research there is beyond trivia statements suggests that the illusory truth effect occurs using a variety of other stimuli including statements about health, news headlines, and politics. Given the importance of such topics, future research should focus on these areas and other topics relating to deeply held beliefs (e.g., beliefs about climate change).

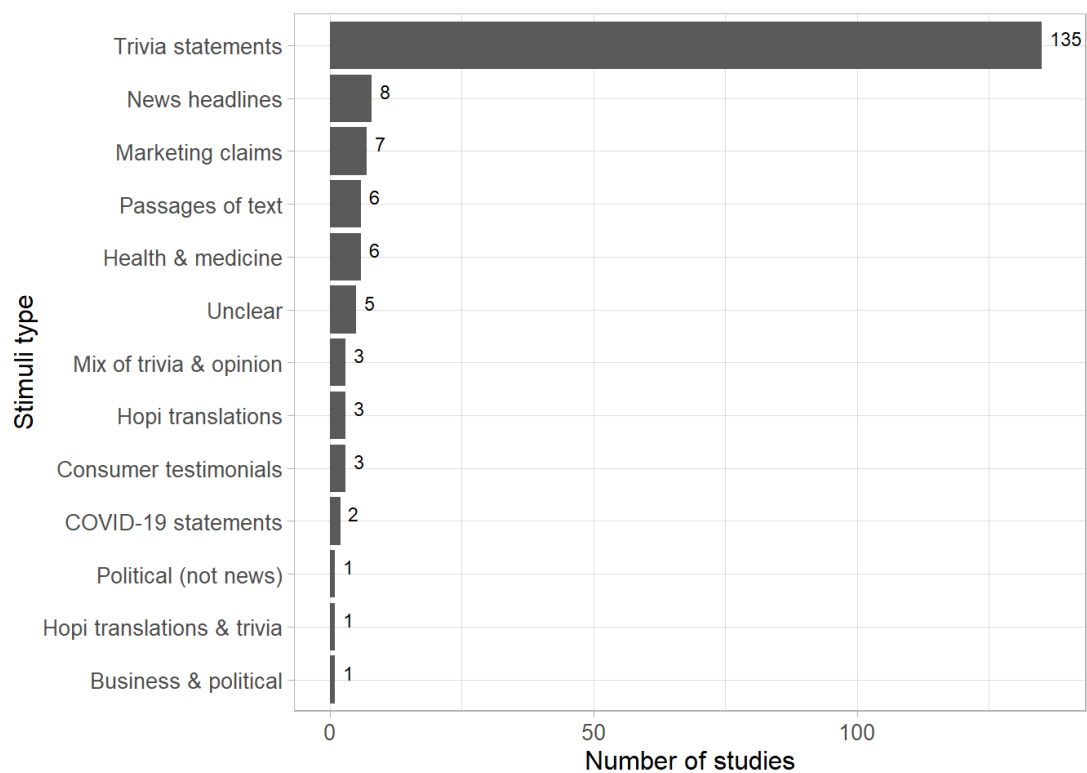


Figure 6. *Frequency and variety of experimental stimuli within the included studies.*

Few studies (15 or 8%) tested whether participants already knew the truth/falsity of the experimental stimuli. If participants already know the answers to some trivia questions, they may use their existing knowledge when judging truth, thereby diminishing the effect of repetition (although prior knowledge does not provide total protection from the effect, see Fazio, 2020). Using normed trivia statements does not completely avoid this issue: Participants correctly answered 36%

of the “unknown” statements from a normed set (Fazio, 2020). Disentangling prior knowledge from the effect of repetition looks to be an interesting direction for future studies.

We coded 55 different tasks or combinations of tasks carried out with the experimental stimuli during the exposure phase (we grouped tasks into meaningful clusters for the purposes of reporting; Figure 7). This level of task variability shows a lack of standardised method for testing the illusory truth effect. Furthermore, some tasks could affect participants’ ratings during the test phase. For example, evaluating stimuli might result in a different level of processing compared to just reading or hearing them (42 or 23%). Asking participants to rate their interest in the stimuli (29 or 16%) could imply that the statements are true and might inadvertently tap into processes that are similar to explicit truth judgements. Similarly, 37 (21%) studies required participants to give truth judgements during the exposure phase, which could encourage them to give consistent ratings during the test phase (Nadarevic & Erdfelder, 2014). Some studies that directly manipulate the exposure task have found that the choice of task moderates the effect. For example, participants rating interest (Brashier, Eliseev, & Marsh, 2020) or categorising statements (Nadarevic & Erdfelder, 2014) show the illusory truth effect, but those rating truth do not. Further synthesis of the literature could compare effect sizes as a function of exposure task, and this could be complemented by research in which the choice of task is systematically varied.

Similarly, there was no consistency in the filler tasks used during the retention interval between exposure and test. Fifty eight different tasks or combinations of tasks were reported, ranging from demographics questions to number puzzles to personality questionnaires. As with the exposure task, it is

possible that different filler tasks could influence the subsequent test phase. Sixty-nine (38%) studies did not specify the filler task, meaning that these studies cannot be evaluated for the influence of filler task on the effect.

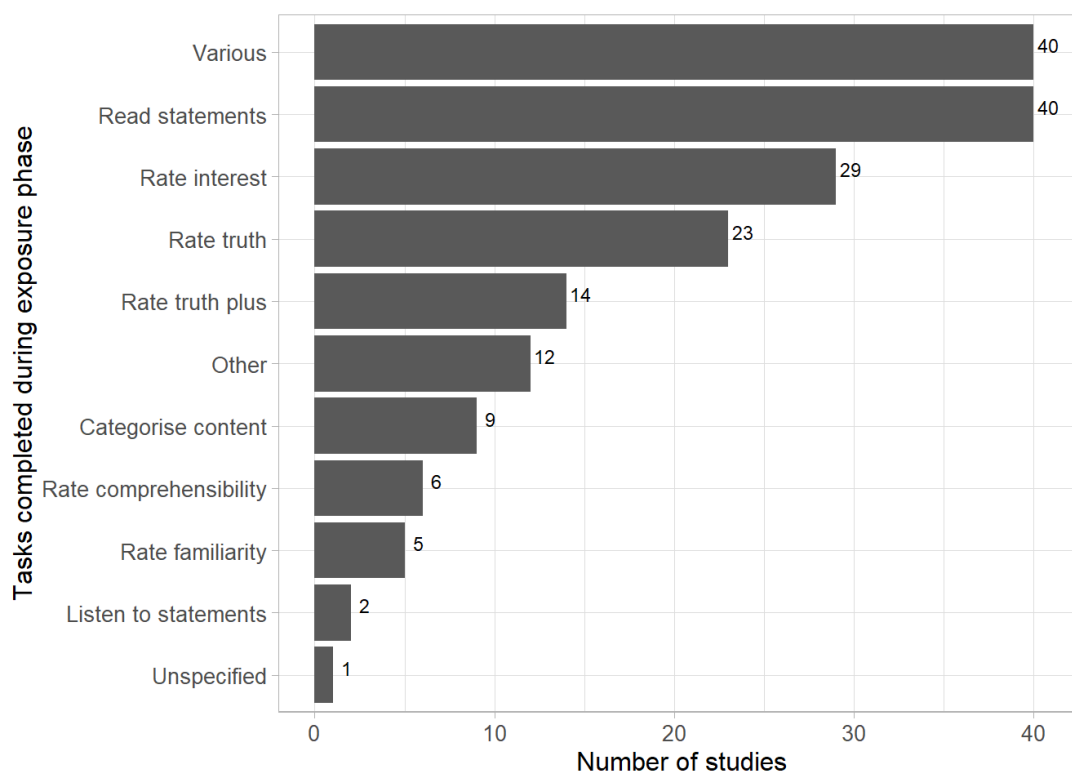


Figure 7. Frequency and range of tasks completed during the exposure phase. If the task involved rating truth and another task, they were coded as “rate truth plus”. Other combinations of two or more tasks were coded as “various”. All tasks involved reading or listening to the critical stimuli. If participants did not carry out any additional task with the critical stimuli, they were coded as “read statements” or “listen to statements”.

There was also great heterogeneity in the measures used to rate truth.

Nineteen “truth” measures were coded in the map, including continuous scales from 1 - 100, Likert-type scales with and without neutral points, and dichotomous judgements (Figure 8). In some cases, the truth measure varied within a paper without explanation. Measuring truth judgements in such diverse ways implies an underlying, latent truth continuum that can be measured in a binary or continuous way, yet there has been no validation or latent construct analysis in the literature. Given the quantity of evidence available, this area merits further synthesis to

investigate whether the illusory truth effect differs as a function of the way in which truth judgements are measured. Additionally, future experimental research should systematically vary the measure to investigate illusory truth as a function of truth measure. Based on the homogeneity of research questions being asked (i.e., does repetition affect truth?), the variability in approaches to measuring truth seems worth addressing. Ideally, the field could establish a few reliable, validated measures and use them consistently (or provide justification for using alternative measures).

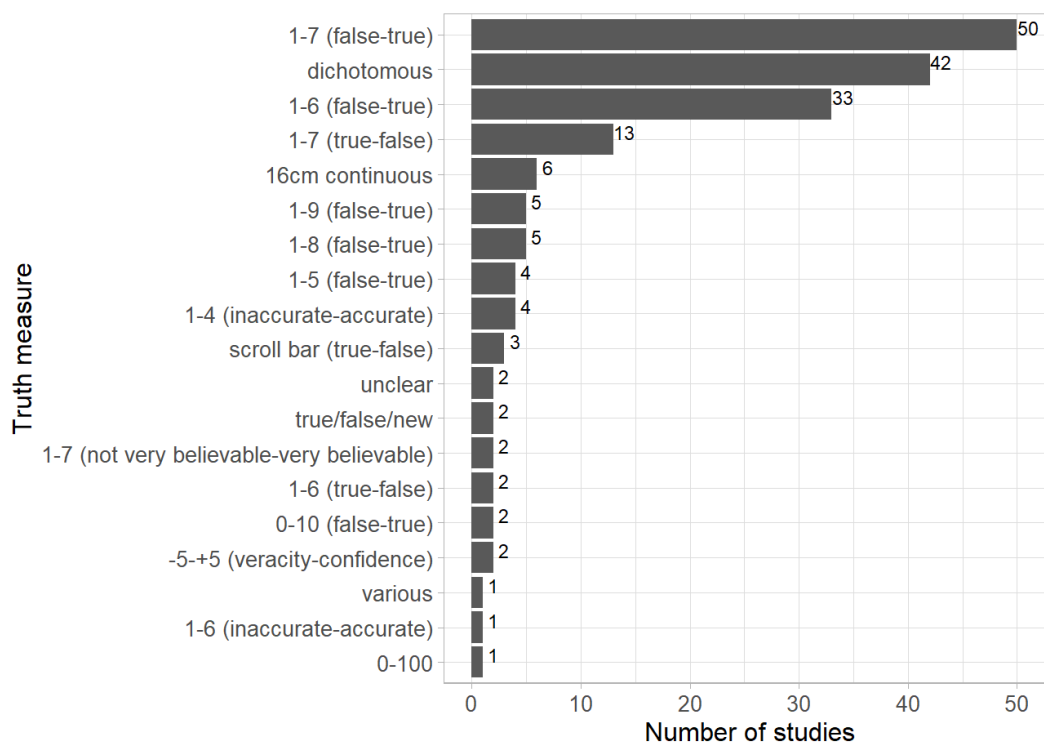


Figure 8. *Frequency and variety of truth measures within the included studies.*

In order to understand the illusory truth effect over time we need a range of retention intervals as well as studies that systematically track the effect over time using multiple retention intervals between exposure and test. We coded the length of the retention interval and the number of intervals used by each study. Overall, the vast majority of studies used a single retention interval, in most studies, the test

phase was conducted in the same session as the exposure phase³² (see Figure 9).

Relatively few studies used multiple testing intervals, and all but 12 (6%) of the test stages occurred within one month of exposure. The literature includes almost no studies testing long intersession intervals, examining the effect over time, or exploring the temporal boundaries of the effect.

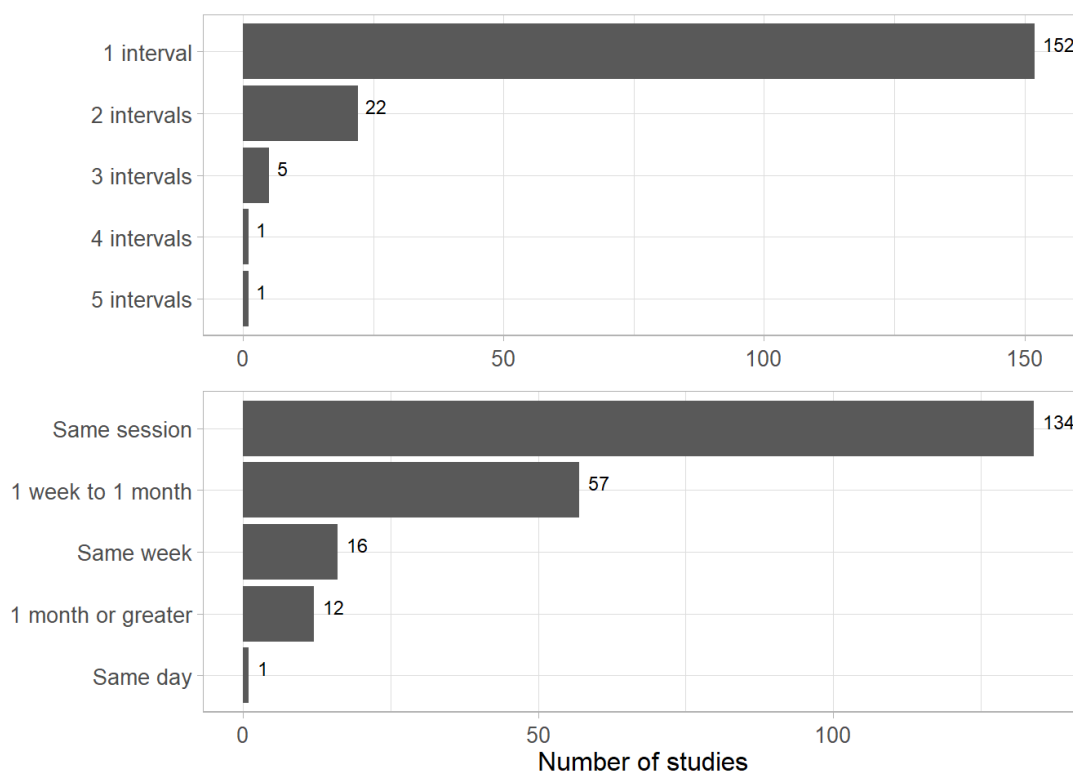


Figure 9. Top panel shows the number of retention intervals used in the 181 included studies. The bottom panel shows the length of the retention interval (i.e., time between exposure phase and test phase). Some studies used multiple retention intervals, $n = 220$.

Although many studies are motivated by the idea that repetition over time increases judged truth, relatively few studies varied the number of repetitions. At exposure phase the vast majority of studies (153 or 85%) presented the stimuli just once (see Figure 10). At test phase, almost all studies (167 or 92%) used a single

³² For the purposes of reporting we re-coded each idiosyncratic interval into the following categories: same session, same day, same week, 1 week to 1 month, 1 month or greater. Within the map database we coded each retention interval as reported in the paper (e.g., 2 minutes, 2 hours, etc). For the 29 studies that used two or more retention intervals, we coded each interval separately and thus have 220 intervals from the 181 studies.

session and presented participants with one exposure to the experimental stimuli (173 or 96%). Consequently the majority of studies are based on one presentation during the exposure and one during the test phase. Other combinations of repetitions are less studied, highlighting the need for studies that vary both the number of repetitions and the gaps between them to examine the illusory truth effect as it might occur in the real-world.

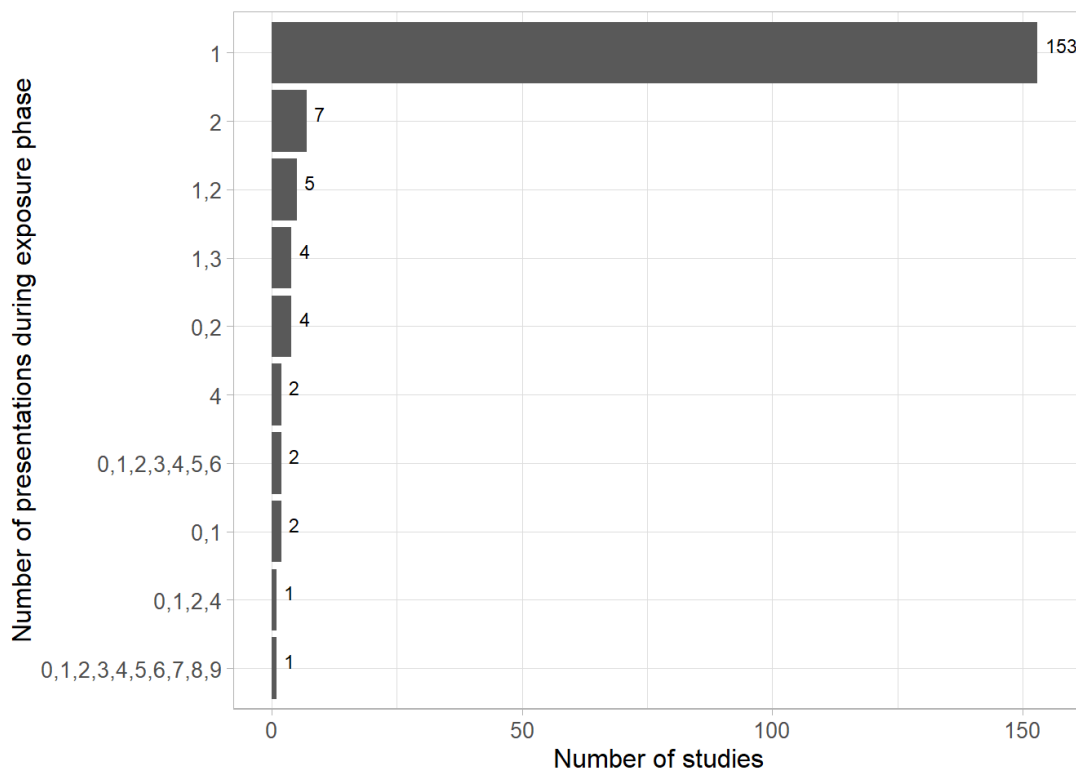


Figure 10. Number of presentations of experimental stimuli during the exposure phase within the included studies. For example, “1, 3” represents studies where individual stimuli were presented either 1 time or 3 times during the exposure phase.

Likewise, although most studies used verbatim repetition of stimuli (148 or 82%), exact repetition in the real world is relatively rare. Gist repetition (8 or 4%) is likely to be more representative of real life information acquisition where repetitions can occur multiple times from multiple sources with variations in prose. For real-world generality we need further research based on repetitions of content, rather than repetitions of exact wording.

3.3.2.3 Openness, transparency, reproducibility and completeness of reporting. In this section we evaluate the completeness of reporting within the evidence base, the frequency of “positive” results, and various transparency practices, in order to assess whether the studies provide enough information to verify that they are reproducible and robust.

3.3.2.3.1 Completeness of reporting. Transparent and complete reporting of sample size should include an explanation of the sample size selected and details of any data dropped from analyses. Study sample size ranged from 12 to 1478 ($M = 153$, $SD = 196$; see Figure 11), with online studies ($M = 331$) being larger than lab or classroom studies ($M = 89$). The majority of studies (139 or 77%) did not provide any rationale for the sample size selected. Only twenty-five (14%) provided a justification that included formal characteristics such as effect size or power level. Around half the studies (94 or 52%) analysed the data from all participants tested, and 68 studies reported exclusions³³. But 14 (8%) studies had unexplained discrepancies between the reported and analysed sample sizes, suggesting unreported exclusions or possible errors.

Conducting a meta-analysis requires reported effect sizes or the descriptive statistics necessary to calculate them. Around three quarters of studies (129 or 71%) reported the results of the overall illusory truth effect in the results section, and of those 74 (57%) reported the effect size. Just over half of studies (102 or 56%) reported the overall means for repeated versus new statements. In the remaining studies, the means were potentially calculable from information provided (51 or 28%), or the information was not reported (23 or 13%). Only 47 (26%) studies

³³ We lacked this information for five studies (3%).

reported the variance or SD for the critical means, 40 (22%) gave a range or provided some information that might make it possible to calculate the variability, but 89 (49%) studies did not provide measures of variance or enough information to calculate them³⁴. Based on this incomplete reporting, an accurate meta-analysis of the entire literature is not possible. However the database will allow researchers to identify meaningful groups of studies that might provide enough information to be meta-analysed.

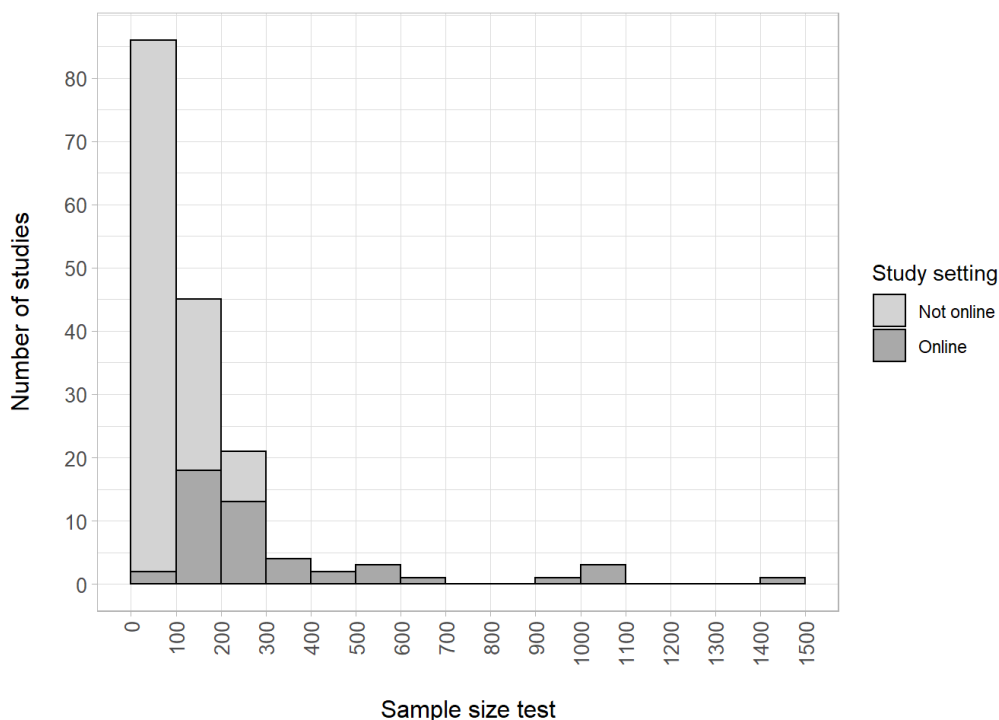


Figure 11. *Sample sizes at test of included studies, split by whether studies were conducted online or not. We lacked information for three studies, $k = 178$. Note that the analysed sample sizes may have been smaller if data were excluded.*

3.3.2.3.2 Transparency and reproducibility. We coded open-science practices to assess the transparency and potential reproducibility of the literature. Note that if the authors reported using an open practice (e.g., sharing materials) and evidence of that practice was available (i.e., *some* materials were shared), we coded

³⁴ We lacked this information for five studies (3%).

the study as using that open practice. We did not, however, verify that sufficient materials were shared to enable a replication attempt.

Most open practices were rare (see Figure 12). The most commonly used practice was sharing of materials, although this was largely driven by preprints and PhD theses. Only a small subset of papers reported sharing open data (24 or 26%), and we were able to access raw data for 17 (18%). Even fewer (7 or 8%) reported available analysis code, meaning that researchers interested in verifying the reproducibility of results could do so only for a minority of studies.

Fifteen (16%) papers reported a preregistered study, with seven of those appearing in 2020, indicating that preregistration is a new and possibly increasing practice in this literature. Although we did not carry out a comprehensive evaluation of those preregistration protocols, we note that several lacked comprehensive details about the procedures and analysis plans. As noted by others, a lack of detail is problematic because it does not sufficiently restrict researcher degrees of freedom (Bakker et al., 2020; Claesen, Gomes, Tuerlinckx, & Vanpaemel, 2019). That lack of precision might be particularly problematic for this literature given that the lack of methodological standardisation across the exposure task, filler task, and truth measures provides opportunities for researcher degrees of freedom. Preregistration represents an area for improvement: In addition to more preregistrations, the field needs more comprehensive preregistrations (or even better, Registered Reports) with sufficient detail to control type 1 error rates.

Thirteen articles (14%) described a built-in (“internal”) replication of a study reported in the same article, whereas only three (3%) included a replication of a study not reported in the same article (“external replication”). Replications are vital

for verification, and replicability is a necessary condition for the accumulation of knowledge. The absence of independent replications, combined with the lack of available code and data, creates uncertainty about the robustness of the evidence in the literature. Further, our estimates of open science practices might overstate their commonality because we coded the article as using an open science practice even if not all of the studies reported in that article did so.

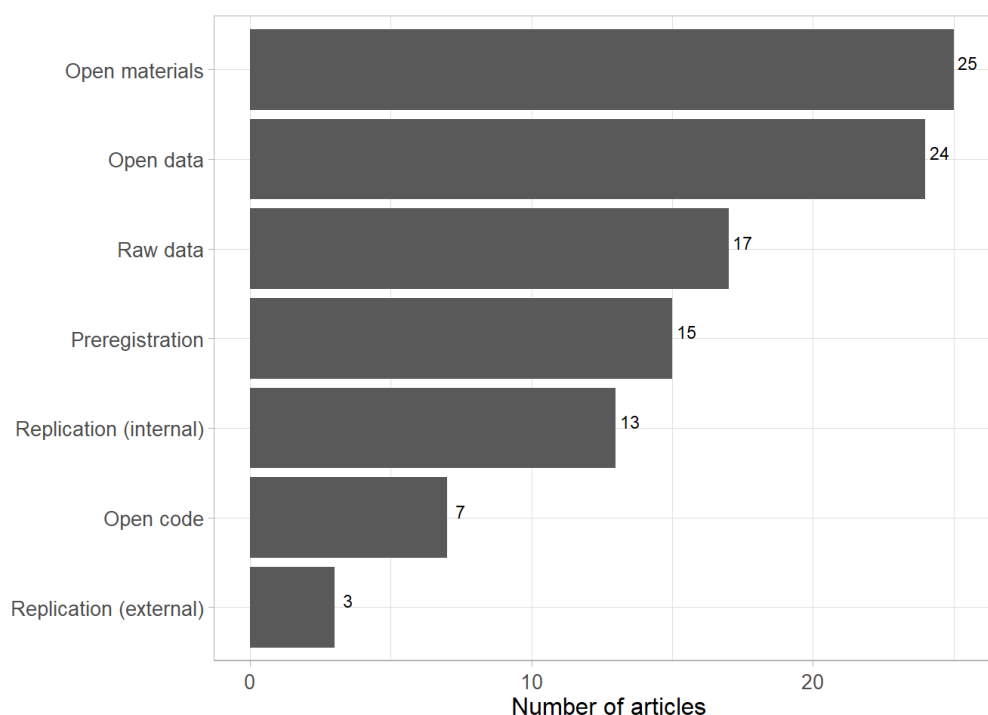


Figure 12. *Frequency of open science practices used within the 93 included articles in the systematic map. If one study within a paper used that open practice it was coded as using that practice.*

3.3.2.3.3 Publication bias. A literature without publication bias should include both positive and negative results. We coded the results of the overall illusory truth effect, as defined by the authors. Note that we have not evaluated the veracity of the claims about findings of the illusory truth effect. Nor have we formally assessed the magnitude of the reported effects (a future systematic review or meta-analysis could do so). Rather, we documented claims of having observed an

illusory truth effect. Therefore these tallies should not be used to assess the presence or absence of an effect.

Of the 129 studies (71%) that reported inferential statistics for the overall illusory truth effect, 124 (96%) reported that the effect was either statistically significant or observed³⁵. Surprisingly, the proportion of positive results was similar regardless of publication status or availability of open data. If all of these studies were testing real effects (not false positives), that means they averaged 96% power. However, within the psychological literature as a whole, power is estimated to be less than 50% (Cohen, 1990), and perhaps as low as 35% (Bakker, van Dijk, & Wicherts, 2012). We did not evaluate the power of the included studies, but given that sample sizes at test ranged from 12 to 1478, and only 25 studies reported some level of formal power analysis, it seems unlikely that all of these studies had $\geq 96\%$ power (if they did, the effect sizes under investigation would have to vary massively as well, and the sample sizes for individual studies would have needed nearly perfect calibration with the true effect size under study). The high proportion of positive results might instead provide evidence of publication bias. The percentage of statistically significant (i.e., positive) results in this literature is similar to that reported for other literatures or for the field as a whole: 95.56% (Sterling et al., 1995), 91.5% (Fanelli, 2010) and most recently 96% (Scheel, Schijen, & Lakens, 2020). In contrast, a recent assessment of Registered Reports which should be comparatively bias free showed just 44% (Scheel et al., 2020).

³⁵ Fifty two studies (29%) did not report the results of an overall illusory truth effect. It is possible that in some cases this was due to the result being non-significant. However without a preregistration, we do not know their a priori aims and whether they planned to analyse the overall effect.

3.4 Discussion

3.4.1 Key Findings

The aim of this map was to document the available evidence on the illusory truth effect. We identified 181 separate studies reported in 93 empirical articles, chapters, or theses. The research spans five decades, 12 countries, and is largely published in psychology journals. The literature includes many studies using verbatim repetition of trivia statements with student participants in a single session. It includes few studies that vary the number of repetitions or the persistence of the effect over time, tasks, and materials.

Overall, the majority of studies used fairly simple and quick data collection procedures that do not provide a strong test of the generality or practical importance of the illusory truth effect: Most studies did not look at the effects of delay, the effects of repeated exposures, or population differences. To increase the generalisability of the effect, future research should diversify beyond the frequently studied domains and focus on questions that help us understand how the effect might work in the real world, such as “how long lasting is the effect of single/multiple repetitions?”. The literature lacks the breadth of evidence to generalise beyond the commonly used participants groups and materials. Future research using carefully designed multi-lab studies, such as those conducted via the Psychological Science Accelerator, would be an appropriate way to ascertain the generalisability of findings in this literature.

In addition to using a restricted set of stimuli and populations, the experimental methodology was characterised by a lack of standardisation in the tasks and measures used to measure the illusory truth effect. There was large heterogeneity in the tasks used during the exposure phase and intersession interval, and there was

substantial variability in the way in which truth judgements were measured. Work is needed both to investigate the potential effect of this variability on the magnitude of the illusory truth effect and to standardise measures in order to increase the reliability and validity of subsequent research. Future research should focus both on synthesising the available evidence on these topics, and on systematically varying these factors within preregistered experiments.

While open science practices are increasing, the lack of available raw data and code means that attempts to reproduce the research would only be possible in a small minority of cases. This factor, along with a dearth of close replication studies, few preregistrations, and largely absent justifications for sample sizes raise concerns about the credibility and robustness of the literature. A lack of sample size justification alone does not mean that the study had low power. However, many literatures appear to be dominated by studies with relatively low power (Bakker et al., 2012; Cohen, 1990), suggesting that researchers are (or were) unaware of the problem of low power. And, studies that do include a power analysis likely are conducted by researchers who recognise the need for larger sample sizes. Consequently, significant results from studies that justified their sample size might be more likely to reflect true positive findings than those that did not. Consistent with the idea, the mean sample size for studies in the map that reported any form of sample size justification was more than double ($M = 277.4$, $SD = 305.1$) the mean for studies that did not ($M = 126.6$, $SD = 148.0$).

This map identified high levels of positive results within the literature, signifying potential publication bias. In addition the levels of incomplete reporting preclude a meta-analysis of the entire evidence base. There is no reason to believe that these issues are more or less severe in this literature than in other fields - most

fields have publication bias. Regardless of their prevalence these issues warrant attention and improvement. To assist future subgroup meta-analyses, we recommend that authors report full descriptive statistics for all measures as well as correlations among measures for repeated-measures designs. Ideally, all future research on the illusory truth effect will make raw data (and a codebook) available in a public repository such as the Open Science Framework.

3.4.2 How to Use this Systematic Map and Database

This map illustrates the quantity and diversity of research on the illusory truth effect. Although we coded articles for open science practices, we did not carry out a critical appraisal. Therefore a high prevalence of a particular type of evidence in this map indicates only that it has been studied frequently, and not that it has been studied well or that the evidence is strong. Further syntheses are required to make evaluations of effectiveness and effect size.

The map is accompanied by a database available at <https://osf.io/37xma/>. The database serves as a searchable resource on the illusory truth effect. This paper reports results that will be of general interest, but the database includes more information and makes it possible for researchers to filter based on specific variables of interest, to understand the areas that are well studied, which papers studied them, and where there is scope for further research.

Researchers may wish to conduct a meta-analysis on some subset of the literature. The systematic map database can be filtered based on specific areas of interest (e.g., studies that use health statements as stimuli) and codes #73 and #74 then can be used to identify whether means and measures of variance are reported for that subset of the literature. To progress from this map to a full systematic review is a

relatively small task since much of the time consuming aspects of the review, such as searching and screening, have already been completed. Before conducting a further review, we recommend that a full critical appraisal is completed as well as an update to include new evidence.

3.4.3 Limitations of the Systematic Map

Although we used the R package `litsearchr` (Grames et al., 2019b) to reduce bias in, and increase the diversity of, our search term selection, we recognise that as a team of psychologists we may have missed terms used in adjacent fields.

Additionally, due to resource constraints, all search terms were in English. Although the majority of psychological literature is written in English, there could be literature in other languages that our search terms did not identify. However we have clearly and transparently reported our search methodology, so the map could be updated with further searches in multiple languages.

Coding the primary goal and results of each study from each article's abstract was challenging due to unclear reporting. Whereas the majority of variables coded in this map were objective, these codes required more interpretation and may therefore be less reproducible. To help overcome this issue, the primary coder (ELH or SJW) sought a second opinion on these codes where necessary.

We coded the first author's institutional location as a proxy for the location in which the study was conducted. This measure is likely to be accurate in most cases, but it is possible that some studies were conducted outside of the lead author's home country.

When assessing open science practices, we coded a paper as having used a practice if there was any evidence of that open practice (e.g., a file containing data

was shared). We did not evaluate whether the shared materials were complete or usable (e.g., whether they included relevant data, a codebook, or runnable code), so we cannot be certain that they allow for reproducibility or that they would be sufficient for a replication. Equally, although we verified whether or not preregistration documents existed, we did not thoroughly review the details and the extent to which the procedures reported in the article matched those in the preregistration. Insufficiently detailed preregistrations might not adequately constrain researcher degrees of freedom and type 1 errors (Bakker et al., 2020; Claesen et al., 2019). In sum, our findings estimate the prevalence of open science practices but not whether those practices are working as intended.

3.4.4 Future Research Summary

Throughout the paper we highlight knowledge gaps in the current literature on the illusory truth effect. We see three general directions for future research: First, test the generalisability of the effect by using more diverse stimuli, participants, intervals, and numbers of repetitions. Multi-lab Registered Reports would be an ideal mechanism for such research. Second, examine the dependency of the effect on the choice of exposure task and truth measure by synthesising the current research. Last, increase the reliability of illusory truth research by standardising the exposure task and establishing validated truth measures.

Chapter 4: The Trajectory of Truth: A Longitudinal Study of the Illusory Truth Effect

Abstract

Repeated statements are rated as subjectively truer than comparable new statements, even though repetition alone provides no new, probative information (the *illusory truth effect*). Contrary to some theoretical predictions, the illusory truth effect seems to be similar in magnitude for repetitions occurring after minutes or weeks. This Registered Report describes a longitudinal investigation of the illusory truth effect ($n = 608$, $n = 567$ analysed) in which we systematically manipulated intersession interval (immediately, one day, one week, and one month) in order to test whether the illusory truth effect is immune to time. Both our hypotheses were supported: We observed an illusory truth effect at all four intervals (overall effect: $\chi^2(1) = 169.91$; $M_{\text{repeated}} = 4.52$, $M_{\text{new}} = 4.14$; H1), with the effect diminishing as delay increased (H2). False information repeated over short timescales might have a greater effect on truth judgements than repetitions over longer timescales. Researchers should consider the implications of the choice of intersession interval when designing future illusory truth effect research.

Keywords: illusory truth, repetition, truth judgement, longitudinal, Registered Report

4.1 Introduction

Human judgements are influenced not only by the informational value of the content we experience, but also by our subjective experience of information processing (Alter & Oppenheimer, 2009). When judging truth or accuracy, people rate repeated statements as subjectively truer than comparable new statements (the *illusory truth effect* or *repetition-induced truth effect*), even though repetition alone provides no new, probative information. That is, repetition generates the illusion of epistemic weight. Inferring truth from repetition is apposite in a world where most of the information people encounter is true, but repetition can create an illusion of truth for false information; the truth effect occurs for both true and false statements (Brown & Nix, 1996), and for both plausible and implausible ones (Fazio et al., 2019b). The effect seems robust to individual differences in cognitive ability (De keersmaecker et al., 2019). It persists even when participants are warned to avoid it (Nadarevic & Aßfalg, 2017), possess knowledge about the factual answer (Fazio et al., 2015), or are explicitly informed about which statements are true and which are false (Begg, Anas, & Farinacci, 1992; Gilbert et al., 1990; Skurnik, Yoon, Park, & Schwarz, 2005). Repeatedly reading misinformation might even reduce how unethical it feels to share that unambiguously false information on social media (Effron & Raj, 2019).

The illusory truth effect could bolster the tactics of propagandists, allowing them to amplify the believability of their message whether or not it is true (see Pennycook et al., 2018; Polage, 2012). Simply by repeating a statement, such as “no country currently has a functioning track and trace app” as Boris Johnson has said, or “President Barack Obama was born in Kenya” as Donald Trump persisted, a politician can increase belief in inaccurate or misleading information. Similarly, an

advertiser can repeat scientifically spurious claims as a means to increase belief in their product's effectiveness. Deliberate use of the illusory truth effect could amplify the believability of claims, from minor (Listerine prevents sore throats) to monumental (Iraq has weapons of mass destruction). Given the ramifications of repetition for belief, there is both a theoretical and ethical imperative to understand better the parameters of the effect.

4.1.1 Explanations, Predictions and Contradictions

In the standard illusory truth effect paradigm, a set of statements, half true and half false, are presented during an exposure phase. There then follows an intersession interval that varies in length from zero minutes to several weeks. During the subsequent test phase, participants rate the perceived truth of a mix of both repeated statements and previously unseen ones. The illusory truth effect is measured by comparing truth ratings for repeated versus new statements. All explanations of the illusory truth effect, including recognition, familiarity, and the most commonly accepted explanation -- processing fluency -- are closely related, rely on memory, and predict that the effect should vary over time. It is perhaps surprising, then, that research to date finds little evidence that the time between repetitions (i.e., the intersession interval) changes the effect (see Dechêne et al., 2010). The aim of our research was to systematically manipulate intersession interval in order to test whether the illusory truth effect is unaffected by the time between repetitions. We next briefly consider the potential mechanisms underlying the illusory truth effect because, even though the Dechêne et al. meta-analysis found little evidence for an effect of delay, all of the proposed mechanisms for the illusory truth effect predict such an interaction between intersession interval and the size of the illusory truth effect. Note, though, that our study did not attempt to distinguish among the possible

mechanisms. Instead, our study and the core of our literature review below focuses on the effect of interval duration.

4.1.1.1 Recognition and familiarity. Repetition is the key prerequisite for the illusory truth effect. The perception of repetition and explicit memory for the prior presentation might enhance the effect (Bacon, 1979; Hawkins & Hoch, 1992; Law, Hawkins, & Craik, 1998). In fact, the perception that a statement has been repeated might be more important than the repetition itself (Bacon, 1979; Hawkins & Hoch, 1992; Law et al., 1998). If a statement is familiar enough for readers to think that they read it before, then they are more likely to judge it to be true. With a longer delay, people should be less likely to detect the repetition (see Brown & Nix, 1996), and the feeling of familiarity should fade (Arkes et al., 1991), hence these mechanisms predict a reduced effect with longer delays.

4.1.1.2 Processing fluency. The predominant explanation for the illusory truth effect is fluency: Repetition leads to easier and more fluent processing. For example, semantic priming leads to faster and more accurate responses (Alter & Oppenheimer, 2009; Whittlesea, 1993). We experience fluency for stimuli we have seen recently, frequently, or for a prolonged time (Oppenheimer, 2008), and people might associate that fluency with truth (Oppenheimer, 2008; Schwarz & Jalbert, 2020). This mechanism rests on the idea that people will misattribute the fluency that comes from repetition to the informativeness or accuracy of the content, thereby increasing belief. Since fluency should be greater for recently repeated items, the illusory truth effect should decline with longer delays. If the source of fluency (in this case repetition) is conspicuous, however, people may correctly attribute their fluency to the repetition rather than to the content, eliminating the effect of repetition

on truth judgements (Alter & Oppenheimer, 2009; Nadarevic & Erdfelder, 2014; Oppenheimer, 2004).

4.1.1.3 Source dissociation. According to the source dissociation account, people may forget the original source of a statement (Arkes et al., 1989). In so doing, they might attribute the statement to a source other than the earlier presentation in the experiment, thereby enhancing the effect by increasing its credibility (Arkes et al., 1991, 1989). With source dissociation, participants experience familiarity without conscious recollection of the previous exposure. Like the *sleeper effect* of persuasion (see Kumkale & Albarracín, 2004 for a review), a source dissociation account predicts that the magnitude of the illusory truth effect should increase as memories of contextual details (i.e., source) are lost with time because the original repetition and the accompanying sense of familiarity will be misattributed to a source other than the earlier presentation; people remember the semantic content but not its source.

4.1.2 The Illusory Truth Effect over Time

In sum, each of the above explanations of the illusory truth effect predicts that the effect should be sensitive to the length of the delay between repetitions. Whereas the source dissociation account hypothesizes a larger illusory truth effect over time as contextual details are lost, the fluency, familiarity, and recognition accounts all predict a decrease in the effect over time because they are enhanced by recency. Note, though, that the fluency account predicts a reduced effect if participants realise that the statements were repeated. All four mechanisms are interrelated and might all contribute to the illusory truth effect synergistically (e.g., fluency and familiarity), antagonistically (e.g., fluency and source disassociation), or individually at different time points. Their precise relationship is an open question (Unkelbach et al., 2019).

Considered individually, these theoretical predictions are inconsistent with the 2010 meta-analysis that found no relationship between the size of the effect and the delay between repetitions (Dechêne et al., 2010; see Table 1); the magnitude of the illusory truth effect was comparable when repetitions occurred moments apart (Begg & Armour, 1991; Schwartz, 1982) or weeks apart (Arkes et al., 1991, 1989; Bacon, 1979; Gigerenzer, 1984; Hasher et al., 1977). These results suggest that people can be influenced by repetitions that are both fresh in memory or more stale, and that once a falsehood has been digested, it persists (Lewandowsky, Ecker, & Cook, 2017; Swire, Berinsky, Lewandowsky, & Ecker, 2017).

Table 1

Moderator Analysis of Delay Between Sessions (reproduced from Dechêne et al. (2010) meta-analysis)

	Session 2	<i>k</i>	<i>d</i>	95% CI		<i>Q_b</i>
				Lower Bound	Upper Bound	
Within-items						3.44 (2.74)
	Within day	9	.25 (.24)	0.07 (0.04)	0.43 (0.46)	
	Within week	11	.44 (.45)	0.31 (0.29)	0.57 (0.61)	
	Longer delay	10	.44 (.45)	0.32 (0.28)	0.56 (0.61)	
Between-items						< 1 (<1)
	Within day	25	.48 (.49)	0.39 (0.37)	0.57 (0.62)	
	Within week	14	.43 (.44)	0.32 (0.28)	0.54 (0.59)	
	Longer delay	12	.48 (.49)	0.36 (0.32)	0.59 (0.65)	

Note. Fixed-effects values are presented outside brackets, and random-effects values are within brackets. Within-items = the difference in ratings for repeated statements between exposure (session 1) and test phase (session 2). Between-items = the difference between truth ratings for new versus repeated statements during the test phase. For within-items, within day is descriptively smaller, however delay did not modify either within-items or between-items as shown by the non-significant goodness of fit statistic *Q_b*.

Few studies have directly manipulated the interval duration, and the comparisons in the meta-analysis were almost entirely across studies with different participants and tasks. The meta-analysis included only three studies, reported across two papers, that manipulated intersession interval. Both used long delays. One study

did not find a difference in the illusory truth effect with intersession intervals of one and two weeks (Gigerenzer, 1984), and another found no difference between one week and one month (Brown & Nix, 1996, Experiment 1). The third study found differences in ratings for false statements after one week, but that effect dissipated after one month and three months (Brown & Nix, 1996, Experiment 2).

Papers published since the 2010 meta-analysis that directly manipulated the interval duration report mixed results. In Nadarevic, Plier, Thielmann, and Darancó (2018, Experiment 2) retention interval (immediately versus 2 weeks) moderated the truth effect when statements were in a foreign language, but did not when the stimuli were in participants' native language. Nadarevic and Erdfelder (2014, Experiment 1) compared retention intervals of ten minutes and one week within-subjects. They reported no effect with a ten-minute interval, but an effect of $d_z = .54$ with a one-week delay. However, Silva, Garcia-Marques and Reber (2017, Experiment 1) observed the opposite pattern between-subjects: $d_z = 1.34$ with a few minutes delay and $d_z = .76$ ³⁶ after a one-week delay. These studies used different materials and tasks, so the effects of short delays on the illusory truth effect might depend heavily on seemingly minor variations in the study design. In the next section we describe the steps we have taken to minimise such variations and focus on the manipulation of interest.

4.1.3 Our Experiment

Despite nearly 40 years of research on the illusory truth effect, few studies have systematically varied the effect of intersession delay, and among that subset of studies, little consensus has emerged. Most claims about the effects of delay on the

³⁶ The effect sizes relate to the new vs. repeated comparison that was manipulated within-subjects, hence they are reported as d_z even though retention interval was manipulated between-subjects.

truth effect are based on cross-study comparisons, often of studies using different methods and designs. The lack of direct tests of the effect of delay likely results in part from the challenges of multi-session studies (e.g., ensuring that participants return to the lab multiple times).

Given the limited evidence, conflicting theoretical predictions, and practical importance, we designed a high-powered, preregistered, longitudinal investigation of the illusory truth effect. We systematically manipulated intersession interval in order to directly compare the magnitude of the illusory truth effect produced when statements are repeated (a) immediately, (b) after one day, (c) after one week, and (d) after one month. Systematic measurement of the effect of delay is a prerequisite of conducting more complex studies that build on the assumption that the size of the illusory truth effect is unrelated to intersession interval. For example, when would be the best time to use repetition as a corrective strategy to counter misinformation? We manipulated intersession interval within-subjects (two short and two long delays), using a simple, consistent design across sessions. Based on previous research, we expected to observe the illusory truth effect (i.e., repeated statements rated as subjectively truer) across our two short and two longer delays. Thus, we tested for a main effect of the illusory truth effect, and we examined whether the size of the effect differs across delays:

H1: We will observe the illusory truth effect. More precisely, we will observe a main effect of repetition averaging across all four delay durations.

H2: We will observe a repetition-by-interval interaction such that the size of the illusory truth effect will differ across the delay durations.

We took several steps to maximise the chances of observing the illusory truth effect. First, we used verbatim repetition of plausible but unknown trivia statements (the “classic” effect). Below, we describe the pre-testing measures we took to ensure that the truth of the statements were unknown to the participants. Second, during the exposure phase, we asked participants to assign each statement to a topic category rather than to judge its truth (Nadarevic & Erdfelder, 2014). Drawing attention to actual truth or falsity of statements at exposure could have reduced or eliminated the size of the effect by encouraging participants to be sceptical when giving their ratings (Brashier et al., 2020; Jalbert et al., 2020; Nadarevic & Erdfelder, 2014). Similarly, we did not inform participants that half of the items were false as that could have increased scepticism. In the real world, statements typically do not come with such warnings, so including them limits generality (Jalbert et al., 2020).

Our analysis procedures include a number of statistical enhancements as well. Previous work has treated items as a fixed factor, with impaired generalisability as a likely consequence (Judd, Westfall, & Kenny, 2012; Yarkoni, 2019). Studies that neglect stimuli as a potential source of variation can have higher false-positive rates, even if items are counterbalanced (Barr, Levy, Scheepers, & Tily, 2013). We analysed our data using linear mixed-effects modelling which permits the simultaneous modelling of subject and stimulus variability (Baayen, Davidson, & Bates, 2008). Second, given that truth ratings use discrete, Likert-style responses, we used ordinal logistic regression (a cumulative link mixed model) rather than an analysis that assumes those responses were based on an underlying continuous measure (e.g., t-tests, ANOVA, linear regression). Treating ordinal data as continuous can increase the rate of false positives or reduce power, especially in

factorial designs where interactions are of primary interest (Liddell & Kruschke, 2018).

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures (Simmons et al., 2012). Our preregistration, stage 1 manuscript, materials, and data are available on the Open Science Framework at <https://osf.io/nvugt/>.

4.2 Method

4.2.1 Participants

Participants were recruited online via Prolific and tested using Qualtrics. For recruiting purposes the experiment was described as a study about assessing a range of trivia statements. The full study description used on Prolific can be found in the “Materials & Procedures” component on the OSF.

Participants received the following compensation: Phases 1 and 2 - a total of £3.00 paid after phase 2, phase 3 - £1.00, and phase 4 - £1.00. In addition, participants who completed the entire study received a bonus payment of £1.00. We used the payment structure to motivate participants to continue with the study by paying them for phases 1 and 2 after completion of phase 2. As preregistered, we have included all participants in our analyses who passed the exclusion criteria detailed below, regardless of whether they completed all phases of the study. The final sample that completed all phases of the study without exclusion was $n = 507$, $M_{\text{age}} = 37.6$ (see Table 3).

At the outset, we used Prolific’s prescreening settings to ensure that participants had an approval rating of $\geq 99\%$, had completed at least 20 previous Prolific submissions, listed English as their first language and United Kingdom as their nationality, and were aged between 18 and 65 years. Participants who

completed the first phase of the study were listed on a custom allow-list. This Prolific feature enabled us to invite only those participants who took part in the first phase to complete the remaining phases.

4.2.2 Design

The design consists of two within-subjects factors: 2 (repetition: new vs. repeated) x 4 (retention interval: immediately vs. 1 day vs. 1 week vs. 1 month).

4.2.3 Sampling Plan

4.2.3.1 Smallest effect size of interest (SESOI). We planned to balance false-negative and false-positive rates by setting power to 95% and establishing a Type I error rate (alpha) of 5%. Our dependent variable was a Likert-type scale ranging from 1 (*definitely false*) to 7 (*definitely true*). Given the limited previous evidence about the effect size of the interaction between time and repetition, we powered to detect a small time-by-repetition interaction: namely, a difference in illusory truth effect no smaller than a tenth of a scale point across two arbitrarily chosen intervals. This difference equates to about 0.14 on the log odds scale. We chose this threshold as it represents a conservative scenario for detecting an interaction. For instance, if the effect only emerged at the last time point, we could detect that effect with high probability as long as the difference in truth judgements between repeated and new statements was at least a tenth of a scale point larger at that time point than at any other time point. This conservative approach also allowed us to detect a wide variety of other patterns, such as a gradually increasing or decreasing repetition effect, or an effect that reaches asymptote at the second time point, so long as the variance these patterns introduce was at least as large as our SESOI.

4.2.3.2 Comparing our SESOI to previous research. A tenth of a scale point equates to a Cohen's d of approximately 0.20, or a d_z of approximately 0.34 (with the caveat that this estimate was made possible by treating the ratings data as continuous rather than discrete, and treating stimuli as fixed rather than random). For context, we considered the 17 studies published since the 2010 meta-analysis that used a 6 or 7 point Likert-type scale. The range of scaled raw effects was [0.13, 1.30], with a mean of 0.45. Thus our SESOI of a tenth of a scale point (about 0.14 log odds) was on the bottom of the distribution of reported effects, smaller than the smallest raw effect reported in this literature.

4.2.3.3 Power analyses. Our power analyses, as well as our main analyses, were performed using R 3.6.2 (R Core Team, 2019). We used Monte Carlo simulation to estimate power, simulating data based on estimates for the variance components from an ordinal logit re-analysis of Nadarevic and Erdfelder (2014, Experiment 1; retrieved from the OSF <https://osf.io/eut35/>). Our re-analysis differed from Nadarevic and Erdfelder's analysis in that we fit a cumulative link (ordinal) mixed model using the ordinal package in R, version 2019.4-25 (Christensen, 2019), which included participants and stimuli (statements) as random factors. The random effects included by-subject and by-stimulus random intercepts and by-subject and by-stimulus random slopes for interval (immediately, 1 day, 1 week, 1 month) and repetition (repeated, new) and their interaction. A report of our re-analysis that includes R code is available in the "Reanalysis of Nadarevic & Erdfelder (2014)" component in the OSF project.

Our analysis confirmed all of their findings except that we did not detect a main effect of the statements' actual truth or falsity on truth ratings, which suggests that the effect described in the original study may be an artefact of unmodelled

stimulus variance. Both in Nadarevic and Erdfelder's study, and in the present study, the stimuli are pre-tested to be of uncertain truth or falsity to participants. Thus, an effect of actual truth was neither expected, nor the experimental focus. We therefore excluded statement truth or falsity in our analyses. We retained the parameter estimates and wrote code to simulate new data on a seven-point scale based on these values (see the OSF repository for details).

Although we were able to estimate a model with maximal random effects (Barr et al., 2013) on the Nadarevic and Erdfelder (2014) data, doing so for our more complex 2 (repetition: new vs. repeated) x 4 (retention interval: immediately vs. 1 day vs. 1 week vs. 1 month) design, in which all factors are both within subject and within stimulus, would have been computationally infeasible given the large number of parameters to estimate. A 2x4 design implies eight fixed effect estimates: intercept, effect of repetition, three predictors to code the effect of interval, and three more to code the repetition-by-interval interaction. A maximal random effects model would imply estimation of 64 random effects parameters corresponding to those eight fixed effects, 16 variances (8 for subjects and 8 for stimuli) and 48 covariances. We simplified the model based on the finding that when performing a significance test for some effect of interest, a model that includes only those random slopes related to that effect performs as well as a maximal model (Barr, 2013). We were able to confirm through simulation that using such a 'minimally sufficient' specification within the cumulative logistic mixed-effects modelling framework maintained the false positive rate at a nominal level (see Figure 1 below for power simulation results with an effect size of zero). Following this approach, for the test of the interval-by-repetition interaction, the only random slopes included in the model were the by-subject and by-statement random slopes for the three predictors coding

the interaction. For the test of the main effect of repetition, the only random slopes included in the model were by-subject and by-statement random slopes for the single predictor coding that factor.

Preliminary benchmarking of our simulation code determined that estimating cumulative link mixed models was too slow to be used to search for plausible subject Ns for power; estimation of a single model on a large simulated data set took as long as 36 hours when the model included only those random effects required to maintain nominal error rates for the interval-by-repetition interaction. Given this limitation, we adopted the strategy of seeking initial estimates of the sample N needed to power the interaction to 95% using a linear-mixed effects model instead of the ordinal model. We then confirmed these initial estimates using the ordinal model. The sensitivity curves in Figure 1 show that we have 95% power to detect an effect of at least .07 (on a log odds scale) for the main effect (H1) and at least 0.14 for the interaction (H2). Further details about the power simulations can be found in the OSF repository under the “Power Calculation” component.

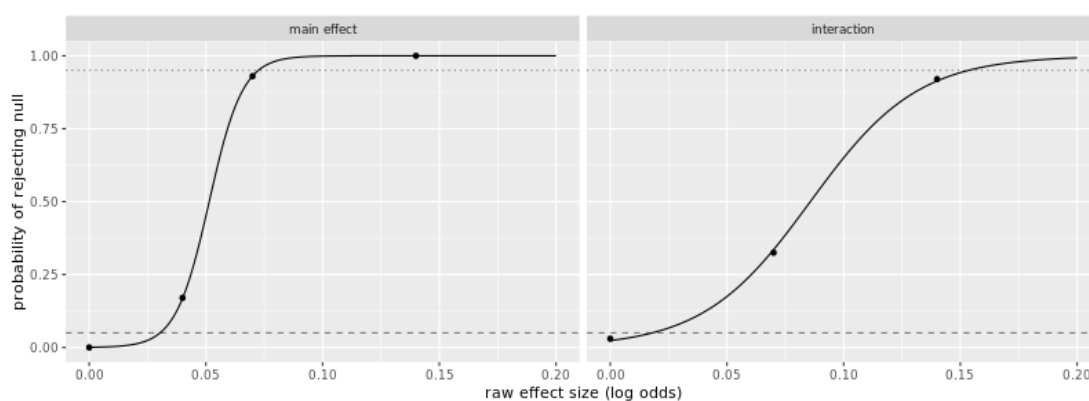


Figure 1. *Estimated sensitivity curves for a sample with a final N of 440 participants (based on a starting N of 608 with dropouts). Each point in the plot was based on 100 simulations, and the curves were obtained by fitting a logistic regression model to the data. For reference, the dashed line is at the 5% null rejection rate and the dotted line is at the 95% rejection rate.*

4.2.3.4 Sampling plan. We used Prolific to recruit participants to the study.

Following discussions with Prolific, our simulated data assumed a participant attrition rate of 5% between phases 1 and 2, 10% (of the remaining N) between phases 2 and 3, and a further 10% (of the remaining N) between phases 3 and 4. Based on the parameters from the Nadarevic and Erdfelder (2014) data, the attrition rates described above, predicted data loss due to exclusions, and 128 stimuli, the number of participants required to detect either effect with the linear mixed-effects model was 608 participants³⁷ at phase 1, and at least 440 participants who provided useable data for the final phase after exclusions. Note that for counterbalancing purposes, the number of subjects starting the study must be a factor of eight. See Table 3 for details of the n recruited, excluded and analysed at each phase.³⁸

4.2.4 Materials

The trivia statements used in our experiment were drawn from those used in two previous illusory truth effects papers, De keersmaecker et al. (2019) and Nadarevic and Erdfelder (2014). The 120 statements used by De keersmaecker et al. (Experiment 3) were originally compiled by Unkelbach and Rom (2017) and were provided via email by Jonas De keersmaecker (05 August, 2019). We amended spellings from US to UK English, corrected misspellings, and excluded statements

³⁷ Prolific recruited 631 participants to fulfil our target of 608 participants. The platform continues to recruit participants until the target number of participants have *completed* the experiment.

³⁸ We removed the following text from the Stage 1 report that described the process we would use to recruit additional participants if we had fewer than 440 useable participants after phase 4: “If at the end of phase 4, and after applying exclusions to all phases, we have usable data from fewer than 440 participants, we will recruit additional new participants. We will calculate the observed attrition and exclusion rate over the course of the study and use this rate to calculate the additional number of participants we should recruit to achieve a sample of about 440 usable participants. The process of recruiting additional participants will continue either until we achieve 440 usable participants, or we have recruited 700 participants to phase 1 of the study, whichever comes first. This maximum participant N is based on funding constraints. In the case that we have usable data from fewer than 440 participants, we will report the smallest effect size that we had 95% power to detect (a sensitivity analysis).”

that were unclear or were not clearly true or false, leaving 95 statements. The 176 statements used by Nadarevic and Erdfelder were retrieved from the OSF (<https://osf.io/eut35/>). We translated those items from German to English using Google Translate, and then asked a native German speaker to check the translations. We excluded statements that we could not translate, that were not clearly true or false, or that were specific to German participants, leaving 145 statements. The combined set of 240 items included 124 false and 116 true statements, and covered a wide range of domains (e.g., history, geography, science).

4.2.4.1 Pre-testing materials. To ensure that the truth of the statements would generally be unknown to UK participants, we pre-tested them using 78 UK participants recruited via Prolific. Statement pre-testing was conducted prior to submitting the stage 1 Registered Report. All materials, data and code from pre-testing can be found in the “Materials & Procedures” component of the OSF project. All analyses were written using R (R Core Team, 2019) and the tidyverse suite of packages (Wickham et al., 2019). We used Prolific’s prescreener settings to select participants with an approval rating of $\geq 95\%$, aged between 18 and 65 years, who listed English as their first language and United Kingdom as their nationality.

We randomly split the 240 statements into four sets of 60 (each including 29 true and 31 false statements). Participants were randomly assigned to one of the four sets and statements were presented in random order. Each set of statements was evaluated by 19-21 participants using the same scale as the experimental dependent variable -- a Likert-type scale ranging from 1 (*definitely false*) to 7 (*definitely true*). The experiment was self-paced and took approximately 8-10 minutes to complete. Upon completion of the experiment, participants received a payment of £1.30. We excluded data from one participant who self-reported using technical aids to find

answers to the question (final sample: $N = 77$, $M_{\text{age}} = 33.0$). We then selected the 64 true statements and the 64 false statements with mean truth ratings closest to the centre of the scale (i.e., the ones that participants were least certain were true/false).

The resulting final set of 128 trivia statements contained 57 statements from De keersmaecker et al. (2019) and 71 from Nadarevic and Erdfelder (2014), with truth ratings ranging from $M = 3.50$ to $M = 4.53$. The statements were randomised into eight stimulus sets of 16 items that were counterbalanced across participants and across judgement phases (see the Materials & Procedures component for the stimulus lists, code, and csv file). Each set included eight true (e.g., “The area between the eyebrows is called the Glabella”) and eight false (e.g., “A galactic year takes 2500 terrestrial years”) statements. During all phases of the experiment, statements appeared on the screen one at a time, with the rating scale positioned directly below the statement.

4.2.5 Procedure

The experiment comprised four phases. Phase 1 followed a typical illusory truth effect procedure: Participants read statements in an initial exposure phase, and then rated the truth of a set of statements during a test phase. A screen recording of phase 1 is available in the “Materials & Procedures” component on the OSF (Heycke & Spitzer, 2019).

4.2.5.1 Phase 1. During the exposure phase participants read the 64 statements (half true, half false) that were repeated over the course of the longitudinal experiment. The statements were presented in a different randomised order for each participant. Previous research suggests that if participants rate truth during the exposure phase, they may try to give consistent ratings during the test phase (Nadarevic & Erdfelder, 2014). Rather than asking participants to rate the truth

of each item, we ensured that they read the items by asking them to assign each statement to a topic category. Response options were: (1) Art & Entertainment, (2) Geography, (3) History & Politics, (4) Language, (5) Science, Nature & Technology, and (6) Sports.

Immediately afterward, participants completed the first test phase.

Participants saw 16 new statements and 16 old statements repeated from the initial exposure phase. For both the new and old statements, half were true and half were false. These 32 statements were presented in a different randomised order for each participant. Participants were asked to judge the truth of each statement on a Likert-type scale ranging from 1 (*definitely false*) to 7 (*definitely true*). Finally, participants completed demographic information about their age, gender, first language, and nationality. They also reported any technical difficulties and whether they had looked up answers. Phase 1 of the experiment took about 15 minutes to complete and, like all phases, was participant-paced throughout.

4.2.5.2 Phases 2 to 4. In the three further phases, participants completed only the test phase from the initial session: One day later, in phase 2, participants read 16 new statements and 16 repeated statements (from the initial exposure phase), and rated each on the same 7-point scale. This procedure was followed in phase 3 after one week, and phase 4 after one month. In every phase the statements were presented in a different randomised order for each participant. The 32 items used in each test phase were sampled without replacement, so each repeated statement appeared in only one of the four test phases, and the new items were not repeated. Participants reported demographic information only after the first session, but reported any technical difficulties and whether they had looked up answers after each phase. Phases 2 - 4 each took approximately 6 - 8 minutes to complete.

To increase the chances of participants completing all four phases of the experiment, we allowed some flexibility in the retention intervals: For the one-day condition, participants were able to complete the session between 08.00 and 22.00 the following day. For the one-week condition, participants were able to complete the session seven to eight days after phase 1. For the one-month condition, participants were able to complete the session 28 to 32 days after phase 1.

Debriefing took place at the end of phase 4. The timing of the debrief was explained at the end of each test phase, and contact details for the first author were provided. For participants who dropped out over the course of the experiment, we used their Prolific IDs to send them the debrief after all data collection had been completed. If participants self-reported looking up the answers to questions during any phase of the study, they automatically were directed to a modified debrief that explained that they would not be invited to take part in further phases of the study.

Prior to the full phase 4 debrief, we used funnel debriefing to check whether participants had guessed the broad purpose of the study. These questions were not preregistered but came after all phases of the study were complete, so did not interfere with the experimental design. We asked participants the following sequence of questions: “Do you have an idea about what we were testing in this study? (If yes, please describe what you think we were testing.)”; “Did you notice that some statements were repeated during the study?”; and “Why do you think we repeated some statements?”. Participants then saw the full debrief.

Following the phase 4 debriefing we asked participants to share their views on the future of this research area using two optional questions: Participants were asked an open-ended question about the topics researchers should focus on and they were asked to rank the importance of five prespecified topics (see “Materials &

Procedures/FutureResearch”). These questions were purely exploratory; the results are shared in the supplement at <https://osf.io/9zwca/>.

4.3 Analysis Plan

4.3.1 Outcome-Neutral Criteria

Below we discuss the outcome-neutral criteria that relate to the fidelity of our manipulation and the associated hypothesis tests.

4.3.1.1 Manipulation checks. Since the only prerequisites to the classic illusory truth effect are that plausible but unknown trivia statements are repeated verbatim, a manipulation check was unnecessary for the present experiment. While recognition and processing fluency are the primary explanations for the illusory truth effect, neither are theoretically specified preconditions. Furthermore, the process of attributing fluent processing of a stimulus to truth is considered to be automatic (Bornstein & D’Agostino, 1992), and the experience of fluency is “at the periphery of conscious awareness, resulting in a vague or “fringe” experience of ease” (Reber, Fazendeiro, & Winkielman, 2002, p.3). Accordingly, fluency is typically studied via its influence on judgements, such as truth, rather than asking participants directly (Reber et al., 2002). In respect to recognition, the effect occurs even when people cannot recall encountering the statements previously (Begg et al., 1992), and conscious recognition might reduce the effect (Alter & Oppenheimer, 2009; Oppenheimer, 2004). Thus self-report measures of perceived processing fluency or recognition are not informative in this case. Also, in a longitudinal design, such checks might have revealed the purpose of the study or have caused participants to respond differently in future phases.

Instead, central to the fidelity of this study was that participants should not know whether each statement was true or false. As reported in the Materials section,

we pre-tested the statements and selected those that participants rated closest to the centre of the scale (range $M = 3.50$ to $M = 4.53$). Thus, the statements should have been sensitive to the repetition manipulation, and there was no indication that floor or ceiling effects would occur. Additionally, we excluded any participants who responded uniformly to all statements (see Exclusion Criteria below). This exclusion criteria served as a general attention check that reduced the possibility that participants had not seen the statements before when they were repeated in later phases.

4.3.1.2 Missing data. Given the longitudinal nature of our experiment, we took several steps to help reduce missing data and avoid loss of power. First, as detailed in the Sampling Plan, we planned to collect data until we had usable data from our target N of 440 participants at phase 4. We exceeded our target of 440 because attrition was lower than expected. Second, within the Qualtrics settings we selected “forced response” for all statements to ensure that participants gave ratings for every statement. Third, when recruiting participants we: (a) wrote a clear study description for Prolific explaining that the study involved four phases; (b) used Prolific’s prescreening settings to select participants who were likely to remain active on the site (i.e., those with more than 20 submissions completed and an approval rate of $\geq 99\%$); (c) paid a fair hourly rate plus bonus for completing all four phases; and (d) minimised the attrition rate between phases 1 and 2 by paying participants upon completion of phase 2.

We included data from participants who did not complete all four phases. As a form of mixed-effects regression, cumulative link mixed models gracefully handle missing data and thus avoid any need for listwise deletion or data imputation.

4.3.2 Analytic Reproducibility

The analysis script was written in R Markdown (Allaire et al., 2020). All analyses used R version 3.6.2 (R Core Team, 2019) with the following add-on packages: tidyverse 1.3.0 (Wickham et al., 2019) for data wrangling and visualisation, ordinal 2019.12.10 (Christensen, 2019) for fitting cumulative link mixed models, emmeans 1.4.5 (Lenth, 2020) for follow-up analyses and equivalence tests, and rmarkdown 2.1 (Allaire et al., 2020) for compiling the analysis script. To ensure reproducibility, we created an R package truthiness 1.2.4 (available in the repository or via <https://cran.r-project.org/web/packages/truthiness/index.html>) with functions preprocess() to pre-process and anonymise the raw data, and reproduce_analysis() to re-compile the master R Markdown analysis script (we also included an R Markdown template in the package). Finally, we prepared a Singularity 3.5 software container (<https://sylabs.io>) to ensure that all analyses were performed with the appropriate software versions.

Raw data and code are available in the project repository. We matched data from the same participant across all phases of the study using their Prolific ID. To preserve anonymity, Prolific IDs were removed before we shared the raw data.

4.4 Results

4.4.1 Exclusion Criteria

Our exclusion criteria were split into participant-level and phase-level exclusions. Table 2 and the list below illustrate the sequence in which exclusion criteria were applied in our data preprocessing code (see the illusory-truth-analysis R Markdown template available through the truthiness package or in the “Analytic Code” component of the OSF). Therefore, if a participant could be excluded for

multiple reasons, the reason for exclusion was coded as the first applicable exclusion criterion.

Table 2

Sequence of Application of Preregistered and Non-Preregistered Exclusion Criteria

Application order	Exclusion criteria
Participant-level	
1	Duplicate sessions recorded*
2	Consent to data collection across all four phases was absent*
3	English not first language
4	Used technical aids to answer question(s)
5	Responded uniformly across an entire phase of the study
6	Failed to complete all phases in a reasonable amount of time
7	No ratings data*
8	Other: participant asked for their data to be withdrawn*
Phase-level	
9	Consent for phase was absent*
10	Failed to complete all of the ratings in the phase

Note. Non-preregistered criteria are marked with an asterisk. “No ratings data” means that there was no more data left for that subject following application of the phase-level exclusion criteria. This occurred if, for example, a participant partially completed phase 1 before dropping out. Data for that phase would be excluded based on the phase-level criterion “Failed to complete all of the ratings in the phase”, leaving no ratings data for that participant at all, and so we also deleted their participant-level information.

We excluded all data from any participant who did not meet the following criteria during *any phase* of the study: self-reported having any language other than English as their first language ($n = 11$), self-reported using technological aids to answer question(s) ($n = 3$), or who responded uniformly (e.g., always answer 1) to all topic categorisations (phase 1 only) or to all truth ratings (in any phase; $n = 7$). To account for overly fast or slow completion of the study in the absence of an experimenter observing data collection in person, we excluded participants who completed the study in more than 40 minutes (for phase 1, and 30 minutes all other phases), or less than 3 minutes (for phase 1) and 1 minute (all other phases; $n = 22$). In addition to the preregistered exclusion criteria, we excluded participants who, during phase 1, did not consent to complete all four phases ($n = 10$), or who started duplicate sessions ($n = 8$), and we manually excluded a participant who requested

that their data be withdrawn ($n = 1$). We also removed participant-level information for participants who had no ratings data left after applying all phase-level exclusions ($n = 2$). In total, the participant-level exclusion criteria resulted in the removal of 64 participants from the sample.

Phase-level exclusions were applied after any participant-level exclusions (i.e., on phases that remained after removing data from those 64 participants). At the phase level, we excluded data from any participant who elected to end their participation prior to completing an entire phase of the study ($n = 13$; see Table 3). For example, if a participant elected to end phase 3 of the study before completing it, we retained their data for phases 1 and 2 of the study but not for phases 3 or 4. In addition to the preregistered exclusion criteria, we excluded data from phases where a participant did not provide consent ($n = 1$). In total, the phase-level exclusion criteria resulted in the removal of 14 phases.

Table 3

Participants Recruited, Excluded, Retained, and Analysed, Separated by Experimental Phase and Gender

Phase	Gender	N recruited	N excluded	N retained	N analysed
1	Female	386	6	380	364
	Male	212	8	204	198
	Gender variant	2	0	2	2
	Prefer not to say	3	0	3	3
	(Missing)	28	28	0	0
	TOTAL		631	42	589
2	Female	365	10	355	346
	Male	201	2	199	194
	Gender variant	1	0	1	1
	Prefer not to say	3	0	3	3
	(Missing)	4	0	4	0
	TOTAL		574	12	562
3	Female	347	7	340	337
	Male	197	1	196	191
	Gender variant	1	0	1	1
	Prefer not to say	3	0	3	3
	(Missing)	3	0	3	0
	TOTAL		551	8	543
4	Female	329	7	322	322
	Male	192	9	183	183

Gender variant	0	0	0	0
Prefer not to say (Missing)	2	0	2	2
TOTAL	526	16	510	507

Note. “Missing” refers to participants who did not finish phase 1 and therefore did not report their gender. Four of these participants were erroneously invited back to future phases because they started multiple sessions at phase 1. Participants who started multiple sessions during any phase were excluded from analyses. “N retained” is the number of participants after exclusions were applied at the end of each phase. “N analysed” is the number of participants after exclusions were retroactively applied. For example, if a participant responded uniformly to all statements during phase 4, their data were excluded from all previous phases.

Retention was better than anticipated across all phases of the study. The combination of exclusions and dropouts resulted in 567 total participants providing data for at least one phase; of the 526 participants who attempted phase 4, data from 507 were analysed (Table 4).

Table 4

Summary of Exclusions, Dropouts, and Attrition by Phase

Phase	Recruited	Attempted	Excluded	Retained	Analysed	Dropout	Excluded	Attrition
1	NA	631	42	589	567	NA%	10.1%	NA%
2	589	574	12	562	544	2.5%	5.2%	7.7%
3	566	551	8	543	532	2.7%	3.4%	6.1%
4	545	526	16	510	507	3.5%	3.6%	7.1%

Note: “Retained” is the number of participants after exclusions were applied at the end of each phase. “Analysed” is the number of participants after exclusions were retroactively applied. For example, if a participant responded uniformly to all statements during phase 4, their data were excluded from all previous phases.

4.4.2 Confirmatory Analyses

After importing the data and applying the exclusion criteria, we fitted a cumulative link mixed model to the data using `clmm()` from the `ordinal` package (Christensen, 2019) which allowed us to model the effects of repetition, interval, and the repetition-by-interval interaction in log-odds space, with a set of thresholds (cut points) representing the thresholds between the ordinal response categories. We fit

models using ‘flexible’ thresholds (the default), which allows the distance between the six cut-points making up the seven point scale to vary freely. This proved the best fitting approach on the Nadarevic and Erdfelder (2014) data. The “Simulated Data & Analyses” component of the project repository contains HTML reports with results from applying the analysis script to data simulated under three different hypothetical scenarios: null main effect and null interaction ([analysis_all_null.html](#)), significant main effect and null interaction ([analysis_main_effect.html](#)), and significant main effect and significant interaction ([analysis_interaction.html](#)). Please consult the repository for further details, including the full annotated code and an appendix with information on how to reproduce the results.

The fixed effects of repetition (new, repeated) and interval (immediately, one day, one week, one month) were coded using deviation-coded numerical predictors. As described above, the models included participants and stimuli as random factors, and the minimally sufficient random slopes for the fixed effect being tested. Each of the two hypotheses were tested using likelihood ratio tests, comparing models with and without the fixed effect or effects of interest, with all random effects held constant. As laid out in the supplementary materials, we started with descriptive statistics and a visualisation of the results, followed by inferential statistics. We followed the preregistered plan summarised in Table 7; using test 1 to test the main effect (H1), and tests 3 and 4 to test the repetition-by-interval interaction (H2)³⁹.

4.4.2.1 Test of main effect (H1). We tested the main effect of repetition using a χ^2 test with one degree of freedom, and with $\alpha = .05$ (Table 7, test 1). Supporting H1, there was a significant main effect of repetition when collapsing over

³⁹ See the stage 1 Registered Report for a detailed description of the processes we would have followed had the results of tests 1 and/or 3 been non-significant (<https://osf.io/9mncq> p. 22).

interval, $\widehat{\beta}_R = 0.57$ ($SE = 0.04$), $\chi^2(1) = 171.88$, $p < .001$. Ratings averaged across items and participants were higher for repeated statements ($M = 4.51$, $SD = 1.45$) than those for new statements ($M = 4.13$, $SD = 1.34$).⁴⁰

4.4.2.2 Test of repetition-by-interval interaction (H2). Next, we tested the repetition-by-interval interaction using a χ^2 test with three degrees of freedom and $\alpha = .05$ (Table 7, test 3). There was a significant repetition-by-interval interaction, $\widehat{\beta}_{R:I1} = -0.47$ ($SE = 0.05$; immediately vs. one day), $\widehat{\beta}_{R:I2} = -0.67$ ($SE = 0.07$; immediately vs. one week), $\widehat{\beta}_{R:I3} = -0.84$ ($SE = 0.07$; immediately vs. one month), $\chi^2(3) = 121.15$, $p < .001$. This significant interaction supports H2, indicating that the illusory truth effect varies over time. The size of the illusory truth effect decreased over time; the difference between the two ratings (repeated minus new) decreased as the interval increased (see Table 5). The data showed greater variability across participants than across stimuli (Figure 2).

Table 5

Mean Ratings and SDs for Repeated Versus New Statements, and Their Difference, by Interval.

Interval	Repeated M (SD)	New M (SD)	Difference
immediately	4.80 (1.54)	4.12 (1.36)	0.68
1 day	4.53 (1.45)	4.14 (1.36)	0.39
1 week	4.41 (1.38)	4.14 (1.33)	0.27
1 month	4.28 (1.37)	4.14 (1.32)	0.14

⁴⁰ Text from the Stage 1 report that considered the possibility of a non-significant outcome has been removed: “By itself, the result of this test is not diagnostic for either hypothesis, because if the illusory truth effect varies over time (repetition-by-interval interaction), this could yield either a significant or non-significant main effect, depending on how the effect is distributed over time. For example, the effect could be countervailed if it appeared at two time points, and was reversed at two time points. Or, it could be underpowered if the effect was present for one interval but not at the others.”

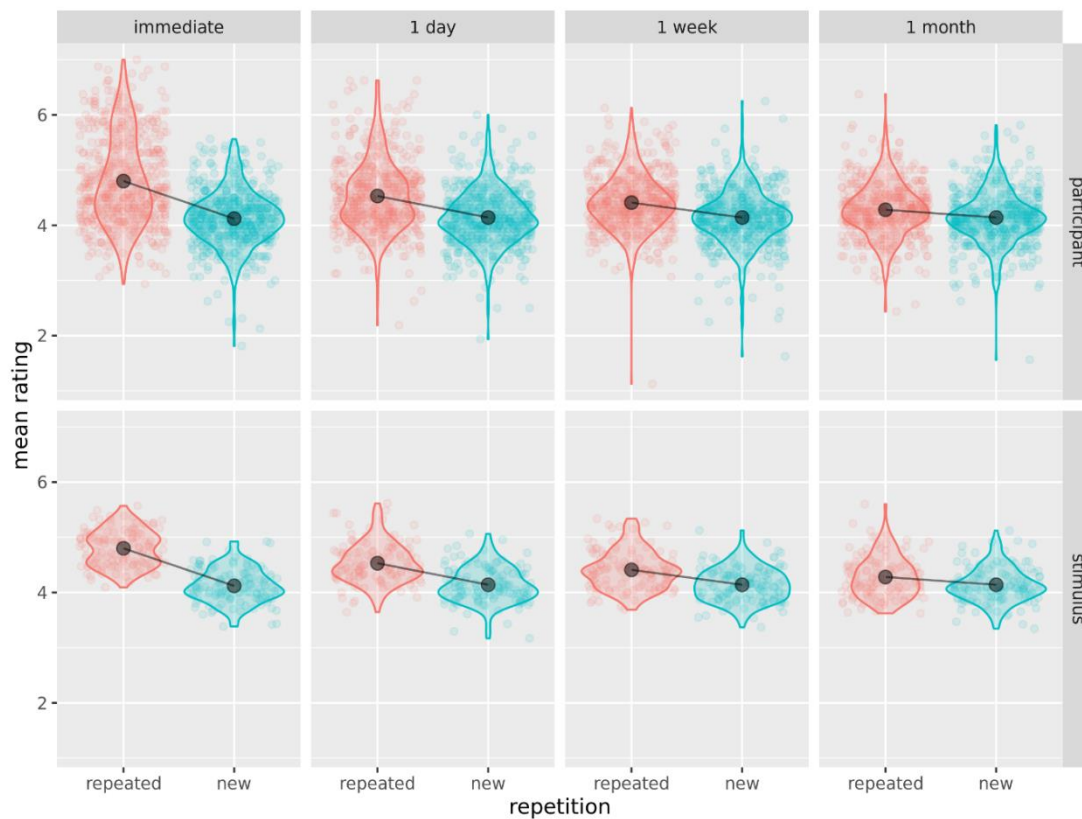


Figure 2. *Effect of repetition across interval, cell means (black points, line) plotted against participant means (top row) and stimulus means (bottom row).*

We followed this significant result using the function `emmeans()` to attempt to localise the effect, testing the effect at each of the four intervals, and using a Holm-Bonferroni stepwise procedure (Holm, 1979) to keep the familywise error rate at .05 (Table 7, test 4). Pairwise comparisons revealed that at every interval, estimated marginal means for repeated statements were significantly higher than those for new statements, indicating that the illusory truth effect was present at all four phases (Table 6).

Table 6

Planned Comparisons of the Simple Effect of Repetition at Each Interval, with Holm-Bonferroni Correction

Contrast	Interval	Estimate	SE	Z ratio	Reject null
repeated - new	Immediately	1.04	0.05	22.68	True
repeated - new	1 day	0.56	0.03	18.64	True
repeated - new	1 week	0.37	0.03	11.00	True
repeated - new	1 month	0.20	0.04	5.53	True

Table 7

Summary of Experimental Design from Research Questions to Results

Question	Hypothesis	Test no	Analysis plan	Power analysis	Results
Is there a time-invariant illusory truth effect?	H1: We will observe a main effect of repetition averaging across all four delay durations.	1	Fit a cumulative link mixed model (as detailed in the “Simulated Data & Analyses” component on the OSF) and conduct χ^2 test with one degree of freedom, with $\alpha = .05$.	95% power to detect an effect of .07 or larger on the log odds scale (about a twentieth of a scale point on a seven-point scale). Based on 440 participants completing phase 4.	Supporting H1, there was a significant main effect of repetition when collapsing over interval, $\widehat{\beta}_R = 0.57$ ($SE = 0.04$), $\chi^2(1) = 171.88$, $p < .001$.
		2	IF tests 1 and 3 are non-significant: Test for the absence of the main effect using an equivalence test with bounds of $\Delta_L = -0.14$ and Δ_U of 0.14 on a log odds scale.	95% power to reject the null of a raw effect greater than .085.	
Does the illusory truth effect vary over time?	H2: We will observe a repetition-by-interval interaction such that the size of the illusory	3	Fit a cumulative link mixed model (as detailed in the “Simulated Data & Analyses” component on the OSF) and test the repetition-by-interval interaction using a χ^2 test with	95% power to detect an effect of a tenth of a scale point, (about .14 on the log odds scale) between two arbitrarily chosen time points: If an illusory truth effect	Supporting H2, there was a significant repetition-by-interval interaction, $\widehat{\beta}_{R:I1} = -0.47$ ($SE = 0.05$; immediately vs. one day), $\widehat{\beta}_{R:I2} = -0.67$ ($SE = 0.07$; immediately vs. one

truth effect will differ across the delay durations.		three degrees of freedom and $\alpha = .05$.	only emerges at very the last time point, we can detect it with 95% power as long as it is at least a tenth of a scale point. Based on 440 participants completing phase 4.	week), $\hat{\beta}_{R:13} = -0.84$ ($SE = 0.07$; immediately vs. one month), $\chi^2(3) = 121.15, p < .001$.
	4	IF test 3 is significant: Use emmeans() to attempt to localise the effect, testing the effect at each of the four intervals, and using a Holm-Bonferroni stepwise procedure to keep the familywise error rate at .05.	N/A	Pairwise comparisons revealed that at every interval, estimated marginal means for repeated statements were significantly higher than those for new statements, indicating that the illusory truth effect was present at all four phases (Table 6).
	5	IF test 3 is non-significant: Test for the absence of an interaction effect using an equivalence test considering all six possible pairwise comparisons of the illusory truth effect across intervals to see whether they fall within the bounds of $\Delta_L = -0.14$ and Δ_U of 0.14 on a log odds scale	With $ \Delta = .14$, 37% power to reject H_0 if the true value is 0, about 18% power if true value is .07 or smaller. With $ \Delta = .20$, 93% power if the true value is 0, 75% power if the true value is .07 or smaller, 18% power if the true value is .14 or smaller. For results with $.14 < \Delta < .20$, see equivtest.html in the repository.	

4.4.2.3 Model validation. As noted above, our cumulative link mixed-modelling approach makes fewer and more reasonable assumptions about the data as compared to a traditional approach using ANOVA. Unlike ANOVA, the cumulative link mixed-modelling approach does not assume that the data come from a continuous, unbounded scale; nor does it assume equal psychological distances

between response categories. This leaves fewer assumptions to be tested than in a conventional ANOVA-based analysis. The main two assumptions behind cumulative link mixed-models are the proportional odds assumption and multivariate normality of the random effects in logit space. Although there has been some discussion of testing the proportional odds assumption for models without random effects (Harrell, 2015) we know of no accepted way to test this assumption for models with random effects. Also, there is no clear consensus in the mixed-modelling literature on how to check the assumption for multivariate normality, with a primary difficulty being distinguishing effects of an ill-fitting model from effects of the data structure or model fitting procedure (Loy, Hofmann, & Cook, 2017). Therefore, we opted to check our model fit using graphical methods. In particular, we generated a reference distribution by simulating data from the model parameter estimates and plotted these distributions against the observed data distributions for participants and stimuli (see Figure 3). Should more systematic model checking procedures be developed in the future, our data are freely available for re-evaluation.

Figure 3 suggests a good fit to the data, except that the model underestimates the by-subject variability for repeated statements at the two earliest phases (immediately, 1 day). However, the effects are so large at these two phases that it seems extremely unlikely that a more complex model that accounted for this overdispersion would yield different results.

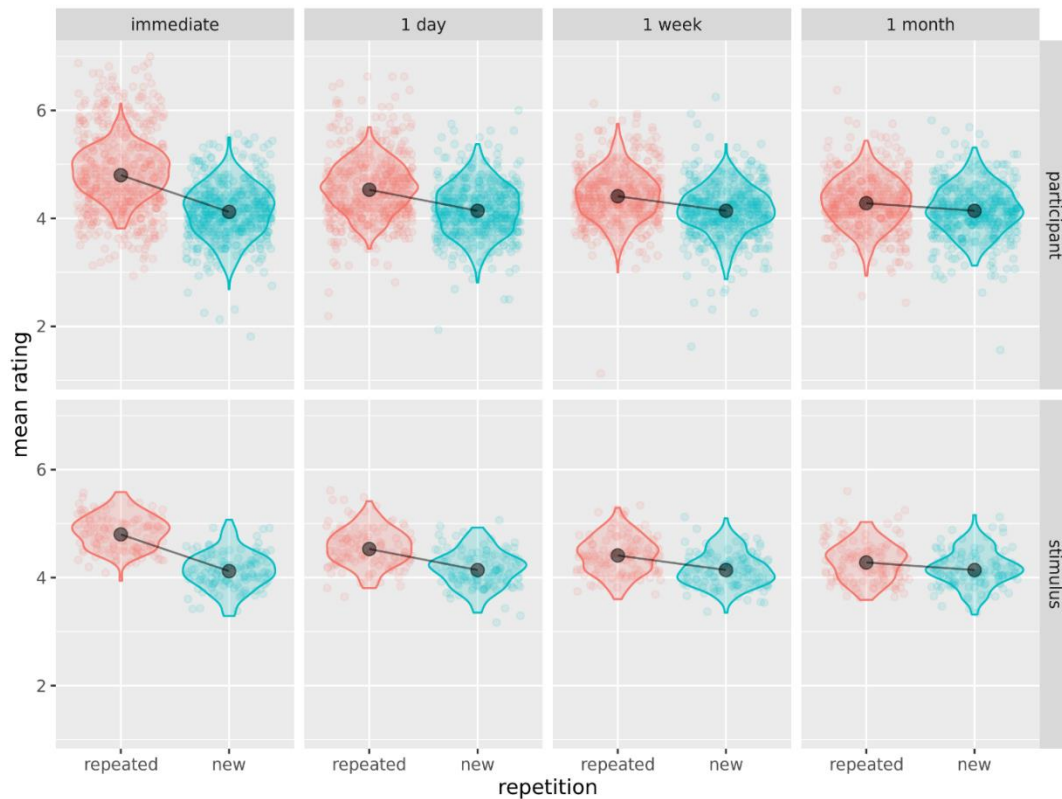


Figure 3. *Model validation: Plot of observed participant/stimulus means (points) against simulated data distributions (violins) and cell means (black points, line).*

4.4.3 Exploratory Analyses

We found that fitting ordinal models on such a large dataset required unreasonable amounts of computation time (up to 24 hours per model on a standard desktop computer). For expedience, our exploratory analyses used standard regression models and correlation instead of ordinal models. The analyses primarily focus on factors that might modulate the illusory truth effect: attention during exposure phase, test phase completion time, and participant age⁴¹. However, first we explored what proportion of participants showed the illusory truth effect.

4.4.3.1 How many participants displayed the predicted effect? We were interested in how many participants behaved consistently with the prediction that

⁴¹ We conducted the age and attention during exposure phase analyses based on feedback from reviewers of our Stage 1 manuscript.

repeated statements would be rated as truer than new statements (J. W. Grice et al., 2020). In order to calculate an overall illusory truth effect for each participant, we combined across all 128 statements and subtracted scores for new items from those for repeated items. Four hundred and eighty-three participants out of 567 (85.2%) showed higher mean truth ratings for repeated statements (see Figure 4). Calculated phase-by-phase, the proportion of participants showing this pattern declined with time: immediately (75%), 1 day (72%), 1 week (71%), and 1 month (61%).

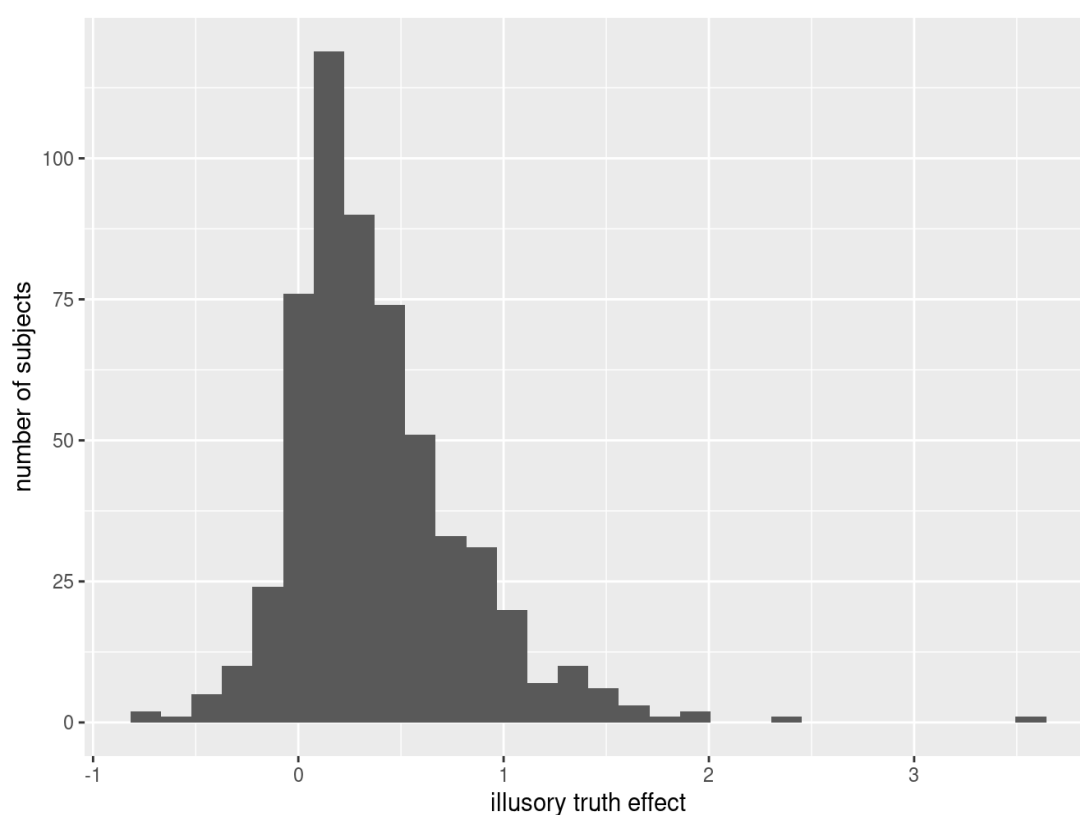


Figure 4. *Distribution of participants showing an overall effect of the illusory truth effect.*

4.4.3.2 Seriousness check. Next, we explored how seriously participants took their participation in the study, and whether seriousness interacted with the illusory truth effect. We used the exposure task—assigning the statements to topic categories—as a proxy for participant seriousness. Each statement had a correct

category⁴² (e.g., “A polo game is divided into periods of 7.5 minutes called 'chukkas'.” would be categorised as “sports”). If participants concentrated during the exposure phase then generally they should have accurately assigned statements to the correct category. One caveat is that some statements might have addressed unfamiliar topics for some participants, meaning that they would have guessed the answer.

Generally participants were accurate in their categorisations ($M_{proportion\ correct} = .912$, $SD = .06$; see Figure 5). In a linear model, just 0.53% of the variation in the illusory truth effect was explained by exposure task accuracy, $F(1, 565) = 3.99$, $p = .046$, $R^2 = .01$. A Spearman's rho correlation revealed a small correlation between the overall illusory truth effect and exposure task accuracy $r_s = .09$, $p = .040$. Thus, the seriousness with which participants approached the task was largely unrelated to the size of the illusory truth effect.

⁴² Some statements fit into two categories. Where this was the case, we considered either categorisation to be correct.

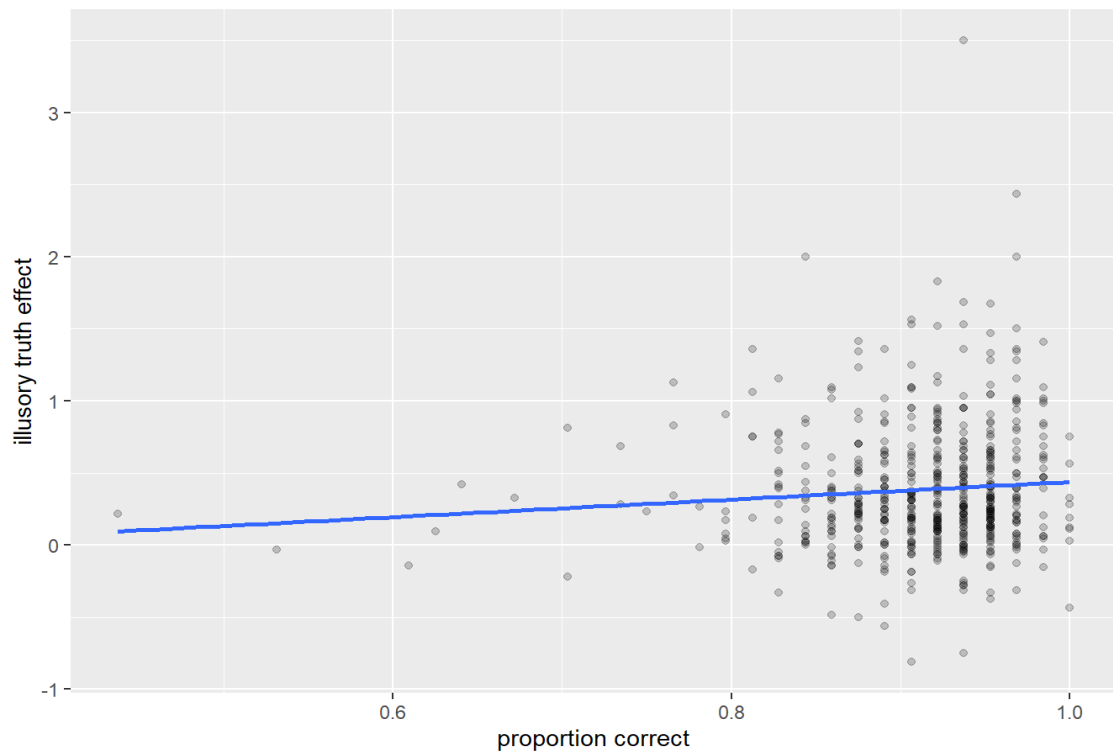


Figure 5. *Illusory truth effect by category judgement accuracy.*

4.4.3.3 Phase duration. Next, we investigated whether there was a relationship between the illusory truth effect and how long people took to complete each test phase. Participants who spent longer completing the task might have expended more effort and given higher ratings for repeated statements. Since phase duration varied with interval (i.e., phases 1 and 4 were longer than 2 and 3), we modelled log duration as a function of phase and used the residuals from this model to predict the size of the illusory truth effect. A Spearman's rho correlation revealed a small correlation between the overall illusory truth effect and phase duration $r_s = .04$, $p = .045$. However a linear model relating the two parameters was non-significant, $F(1, 2148) = 1.96$, $p = .162$, $R^2 < .001$. Consistent with the results from the seriousness analysis, there is little evidence that participants' attentiveness to the task was associated with the size of the illusory truth effect. However, these results

should be interpreted with the caveat that our stringent criteria for study participation might have selected for conscientious participants.

4.4.3.4 Age. Since all explanations of the illusory truth effect rely on memory, and memory performance varies with age, we investigated whether age modulated either the main effect, or the trajectory of the effect over time. Participant ages ranged from 18 to 65 years, with a mean of 37.24 years ($SD = 12.08$; see Figure 6)⁴³.

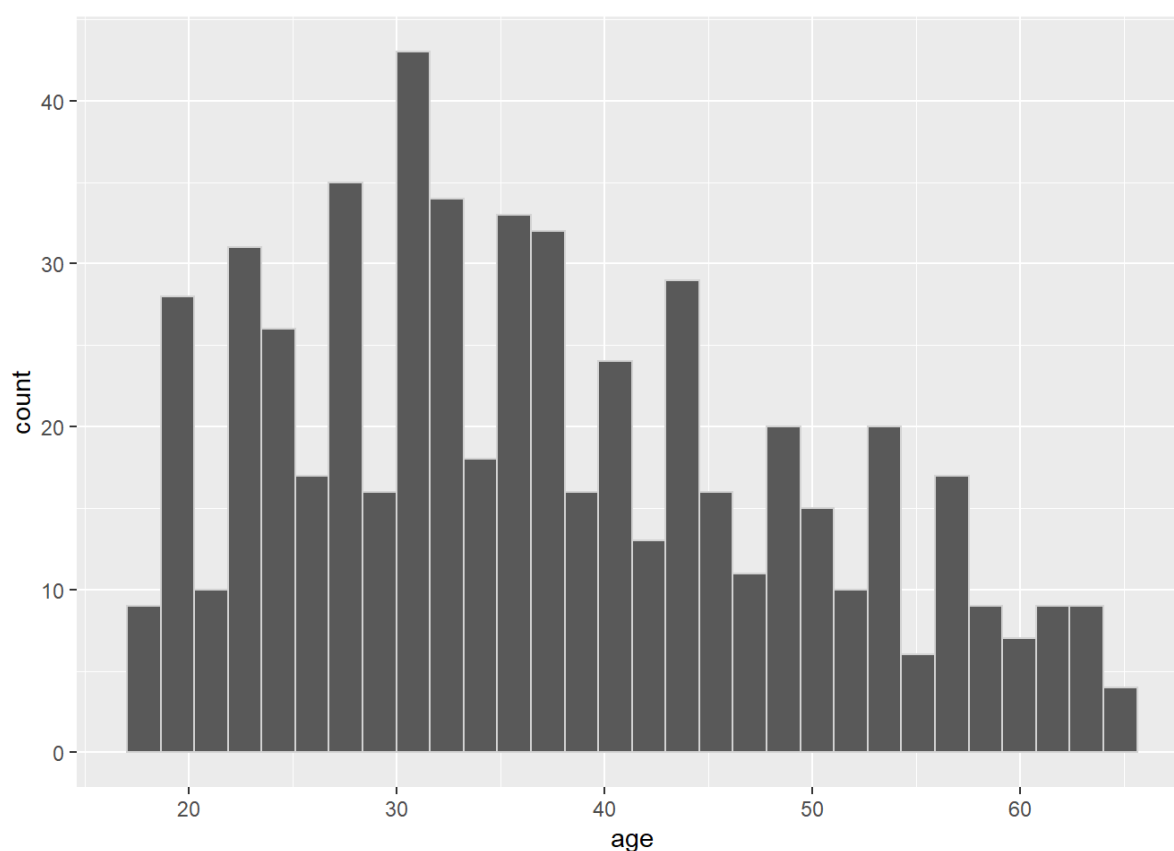


Figure 6. *Distribution of participants' age.*

The Spearman's rho correlation between the overall main effect of illusory truth and age was non-significant, $r_s = .06$, $p = 0.154$. Similarly, the results of the linear model were non-significant, $F(1, 565) = 0.54$, $p = .464$, $R^2 < .001$, suggesting

⁴³ An adjusted version of Figure 6, using one bin per year, can be found in Appendix F. This footnote and appendix are thesis-specific and do not appear in the published version.

little relationship between a participant's age and the overall size of their illusory truth effect.

Next, we examined whether age modulated the decline in the illusory truth effect over time. The finding of a consistently decreasing trend in the confirmatory analysis above, with no discontinuities or asymptotic behaviour, suggests that the data are amenable to polynomial regression, which provides a more concise mathematical summary of the trajectory than the 2x3 factorial analysis we used above. As a first step, we used model comparison to determine whether the trajectory was best described as a linear, quadratic, or cubic polynomial function of interval. A cubic model provided the best fit for our data. We found that variance in trajectory by age was statistically significant, $\chi^2(3) = 11.08, p = .011$. To interpret the trajectory-by-age interaction, we modelled the predicted trajectories for two age groups: participants aged 25 years (about 1 SD below the mean age) and participants aged 50 years (about +1 SD above the mean age). As can be seen in Figure 7 the trajectories vary by age, with the primary difference concentrated at the immediate interval, where older participants showed a bigger truth effect (0.76 versus 0.59 for 50 versus 25 year olds, respectively), and at 1 month with older participants showing a smaller truth effect (0.11 versus 0.18, respectively; see Figure 7).

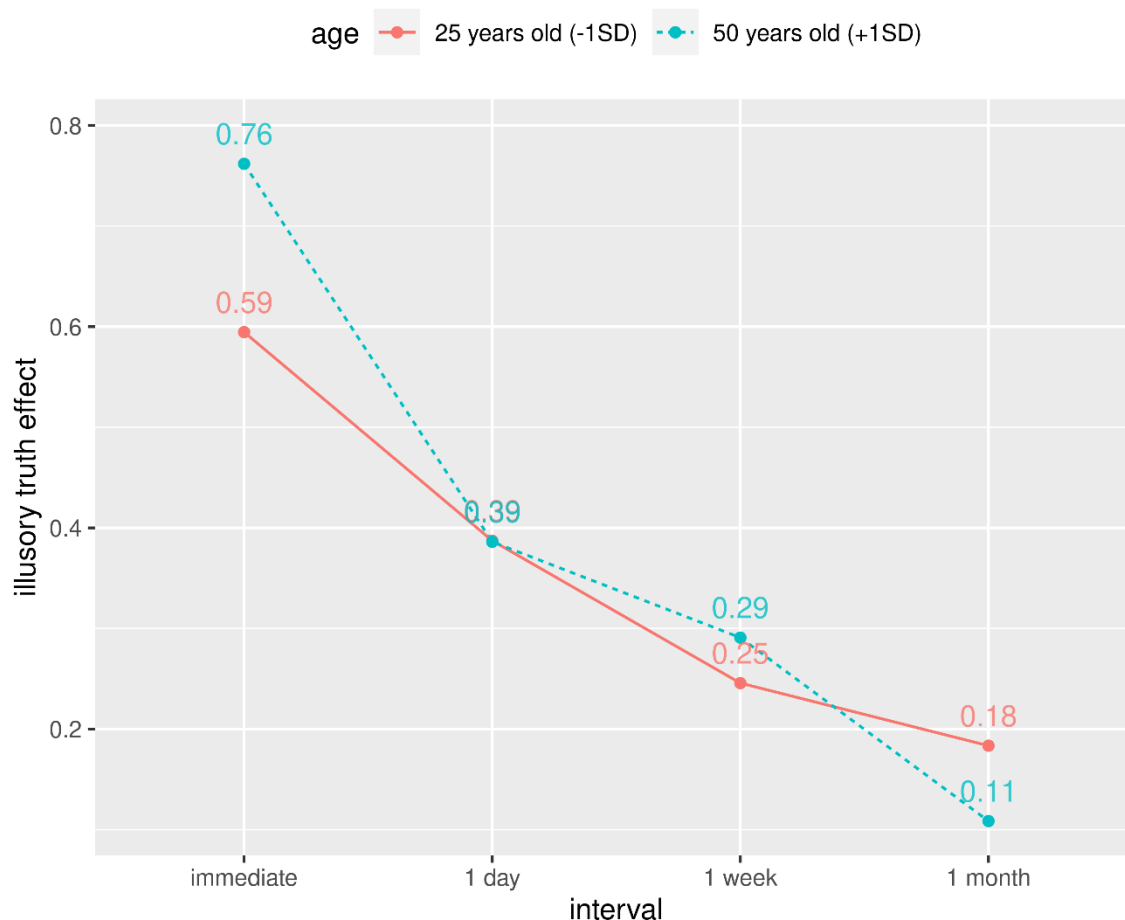


Figure 7. Model predictions for the trajectory of the illusory truth effect for two ages.

4.4.4 Supplement

The illusory truth effect has typically been analysed using ANOVAs. We have outlined the reasons that cumulative link mixed modelling is more appropriate for these data. For comparison, we report the results (in the online supplementary materials at <https://osf.io/ngrw7/>) of a 2 (repetition: new vs. repeated) x 4 (retention interval: immediately vs. 1 day vs. 1 week vs. 1 month) repeated measures ANOVA using participants' mean truth ratings as the dependent variable. Although we did not use this ANOVA for our confirmatory hypothesis test (H2), the results were in line with the findings from our cumulative link mixed model.

Additionally, the online supplement documents further exploratory analyses of statement length and topic, the results of three funnel debrief questions used to ascertain whether participants had guessed the purpose of the overall study, as well as questions about the topics researchers should study in the future (see Appendix E).

4.5 Discussion

We used a repeated measures, longitudinal design to investigate the trajectory of the illusory truth effect over time: immediately, one day, one week, and one month. Both of our hypotheses were supported: We observed a main effect of the illusory truth effect when averaging across all four delay conditions (H1). The illusory truth effect was present at all four intervals, but the size of the effect diminished as the interval duration increased (H2). The repeated-minus-new difference was largest when tested immediately (0.67) and shrank after one day (0.39), one week (0.27), and one month (0.14). This reduction in the illusory truth effect over time is inconsistent with an earlier meta-analysis that found no relationship between the size of the effect and intersession interval across studies (Dechêne et al., 2010), but it is consistent with one between-subjects study showing a smaller effect after one week than after a few minutes (Silva, Garcia-Marques, et al., 2017, Experiment 1).

The reduced effect after a delay is consistent with the recognition, familiarity, and processing fluency explanations of the illusory truth effect. All three explanations predict larger effects for recently repeated items and smaller effects as feelings of recognition, familiarity or fluency fade with time.

A caveat to the processing fluency account occurs when the source of fluency is obvious (e.g., when participants recognise that statements have been recently repeated). In such cases, participants might not use processing fluency to make their

judgements of truth, thereby eliminating the effect (Alter & Oppenheimer, 2009; Nadarevic & Erdfelder, 2014; Oppenheimer, 2004). Our results challenge this fluency discounting explanation because the size of the illusory truth effect was greatest when tested immediately, when participants should be most aware that some statements had been repeated. Similarly, the source disassociation hypothesis predicts that the illusory truth effect should increase with time as people forget that they saw the statements during the experiment, remembering only the semantic content and attributing it to a source outside the experiment. Here we find the opposite.

Our exploratory analyses revealed that most participants (85.2%) showed the illusory truth effect, suggesting that the effect is reliable across participants. However, fewer people might show the effect if participants had been warned that some statements would be false. We chose not to do so because in the real world, false statements are not accompanied by warnings (Jalbert et al., 2020).

The overall illusory truth effect was not associated with individual differences in participants' diligence in performing the task. However, we cannot be sure that participants were attentive the whole time because our online study only provided an overall measure of phase duration. It is possible that participants who spent longer during a phase actually might have been distracted and less attentive.

Our exploratory analyses revealed little association between age and the overall illusory truth effect, but the trajectory of the effect over time did vary with age. Compared to younger participants (25 years), older participants (50 years) showed a bigger truth effect at the immediate interval and a smaller effect at the 1-month interval. If future research replicates this relationship between age and the repetition-by-interval interaction, then this pattern might reflect a reduction in

memory in older adults: During the immediate phase, reduced working memory performance might lead to a feeling of familiarity (without explicit recognition) that could increase truth ratings. At one month, a decline in longer term memory could mean that statements are neither recognised nor familiar, resulting in a smaller truth effect.

4.5.1 Constraints on Generality (COG) Statement

We used Prolific to recruit UK adults (18 - 65 years) with English as a first language. Given that the illusory truth effect has been observed in a range of countries using a range of languages, we expect the repetition-by-interval interaction to generalise to other WEIRD adults (Western, Educated, Industrialized, Rich and Democratic; Henrich, Heine, & Norenzayan, 2010), regardless of language spoken. However we lack evidence showing that the results will generalise beyond this population. Our pre-screen questions on Prolific might have selected for more conscientious participants, but there is no evidence that personality differences contribute to the illusory truth effect (De keersmaecker et al., 2019).

Given that we modelled stimuli as random effects and show that the illusory truth effect is not reliant on particular trivia items, we expect our results to generalise to other sets of statements that comply with these two critical features: 1) ambiguous veracity—if participants already know the answer to questions, they may use existing knowledge. Future replications or extensions should follow our statement pre-testing procedure to ensure that statements are generally unknown; 2) topic—statements should not relate to dearly held beliefs. Considering that the vast majority of illusory truth effect research uses trivia statements as stimuli, it is not clear that this result will generalise to topics on which participants hold strong prior beliefs. Our exposure task ensured that participants read and processed the statements, but we did not ask

participants to rate them for truth. We do not know whether the effect of delay would be robust if participants rated truth at exposure because that might lead them to give consistent responses (Nadarevic & Erdfelder, 2014).

We do not think the effect is likely to rely on the presentation medium (computer vs. paper and pencil), and we have no reason to expect temporal or historical context to be relevant in observing the repetition-by-interval interaction. That said, data collection occurred during the COVID-19 pandemic (December 2020 to January 2021) when the UK was under lockdown or semi-lockdown. The timing of the study might partially account for the high retention across the experiment. We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

4.5.2 Future Research

This Registered Report is the first study to manipulate intersession interval systematically over two short and two longer time periods, so the results should be replicated to ensure their generalisability, particularly because theories on the underlying mechanisms of the effect make contradictory predictions about the effect's trajectory. Future research should also consider the real-world implications of the observed differences and attempt to calibrate them. For example, what is the implication of a 0.67 increase on a 7-point Likert-type scale in a person's truth judgement compared to 0.14? Given the reduction to a 0.14 difference after one month, future research may also investigate the temporal boundaries of the effect.

4.5.3 Conclusion

The aim of this study was to investigate whether the size of the illusory truth effect depends on the intersession interval. We found the size of the effect declined

over time. Whereas previous research suggested that the effect size was unrelated to the interval between exposure and test phases (Dechêne et al., 2010), our results suggest that researchers should consider the implications of the choice of an intersession interval when drawing inferences about the illusory truth effect. An effect that diminishes with time is consistent with the recognition, familiarity, and processing fluency explanations of the effect: As feelings of recognition, familiarity or fluency decrease, so does the effect of repetition on judged truth. The repetition-by-interval interaction implies that when false information is repeated over short timescales it may have a greater effect on truth judgements than repetitions that are far apart.

Chapter 5: General Discussion

In the introductory chapter, I reviewed the literature on two truth effects: the linguistic concreteness effect and the illusory truth effect. I then set out to answer three research questions about the robustness of those truth effects, using transparent, preregistered methods: 1) “Does a linguistic concreteness effect replicate?”; 2) “What is the evidence for the generality and pervasiveness of the illusory truth effect?”; 3) “Is the illusory truth effect immune to time?”. In Chapter 2, two studies attempted to replicate the linguistic concreteness effect whereby concretely worded statements are rated as truer than their abstract equivalents. Chapter 3 used systematic mapping to catalogue and review more than 4 decades of illusory truth effect research, documenting the prevalence, methods, findings, and transparency. Chapter 4 describes a four-phase longitudinal study investigation of the trajectory of the illusory truth effect over time.

I begin this chapter by highlighting the key results from each of the three empirical chapters. I then synthesise these findings and discuss their broader implications. Before turning to the limitations of the present research, I discuss the limitations of illusory truth effect research as a whole and offer potential solutions. I conclude with recommendations for future lines of research.

5.1 Overview of Results

Chapter 2: Does a linguistic concreteness effect replicate? Across two preregistered replication studies, I did not observe evidence of the linguistic concreteness effect. Based on the linguistic category model (Semin & Fiedler, 1988, 1991), Hansen and Wänke (Experiment 1, 2010) found that participants judged concretely worded trivia items as more likely to be true than abstractly worded versions with the same content. The explanation, according to the model (Semin &

Fiedler, 1988), was that the concretely worded statements contained verbs such as “wrote” that conjure a clear, easily verifiable image, while the abstract versions (“is by”) did not. I was interested in the original study because the manipulation appeared so subtle that it could be used, either alone or combined with the illusory truth effect, for nefarious purposes without the addressee noticing. While this is not the only potential mode of using this effect, the fact that it could be deployed by non-scientists with questionable ethical intentions added to the justification for pursuing this line of research. My intention was to investigate how the effect could be combined with the illusory truth effect to potentially produce even greater effects on beliefs. But because the effect had only been shown in one well-cited paper (Hansen & Wänke, 2010), I first chose to closely replicate the original study to ensure that I was building on reliable results.

Confirmatory findings from two separate cohorts (a classroom as per the original experiment, and an online extension) revealed that truth ratings were similar for statements regardless of the concreteness or abstractness of the contents. I supplemented the frequentist analyses with Bayes Factors (e.g., Kass & Raftery, 1995). A Bayes Factor compares the predictive adequacy of one hypothesis with the other, and quantifies the evidence for both models (Wagenmakers et al., 2018; Wagenmakers, 2007). Additionally, I used equivalence testing to test for the absence of an effect larger than the specified smallest effect size of interest (Lakens et al., 2018) which in this case was half the size reported in the original study. The results of all tests were consistent with each other, and collectively did not provide evidence for a linguistic concreteness effect on truth judgements.

The purpose of replication studies is to provide additional evidence that should either increase or decrease our confidence in the reliability of the tentative

evidence provided by an original study or set of studies. Although no single study can be considered definitive about the existence or absence of an effect, both replications produced evidence more consistent with the absence of an effect than with the original effect, and therefore raise doubt about the reliability of a linguistic concreteness effect. With greater than 95% power to detect an effect half the size of the original, both replications minimised the chances of obtaining a false negative. The balance of evidence from this large ($N = 466$ combined across two experiments), well-controlled, preregistered study, conducted in consultation with the original author, compared to the original study ($N = 46$), suggests it is much more likely that the original was a false positive than any other explanation. I therefore decided not to pursue this line of research and focus solely on the illusory truth effect. I began by systematically synthesising the literature.

Chapter 3: What is the evidence for the generality and pervasiveness of the illusory truth effect? Though studies are abundant in some areas of illusory truth research, there is a lack of research on some fundamental aspects of the effect. I used systematic mapping to locate and catalogue the breadth of evidence on the illusory truth effect using preregistered, transparent, and reproducible methods. The process identified 181 studies reported across 58 published articles and chapters and 35 unpublished sources⁴⁴. The associated map database is freely available to other researchers and can help motivate future empirical work. The map highlights areas that have been frequently studied and knowledge gaps where there is limited or weak evidence. Most of the included studies used verbatim repetition of trivia statements in a single testing session with a short delay between exposure and test. Like much of

⁴⁴ Note that the sources identified by the systematic map are not necessarily high-quality. The mapping process simply synthesises the available evidence.

the psychology literature, the studies were conducted at Western universities and often used convenience samples.

There was a dearth of research using stimuli on topics of public interest such as political or health statements. Other limitations of the literature include the need for studies examining effects of multiple repetitions and longer intervals between the exposure and test phases. Notably there was no standardisation in either the exposure task (55 different tasks) or the truth measure used as the dependent variable (19 measures of truth). Many studies failed to report effect sizes, or the descriptive statistics necessary to calculate them. Although there is a promising trend towards open research practices in the most recent publications, the literature as a whole is characterised by an absence of independent replication attempts, a lack of preregistered experiments, and unavailable data and code, meaning that verifying replicability and robustness is only possible for a small subset of the literature.

Throughout Chapter 3 I highlight gaps in the illusory truth effect literature. These can be summarised in terms of three directions for future research. First, it is important to examine the background assumptions (auxiliary hypotheses) being tested along with the main hypothesis. For instance, investigating the dependency of illusory truth on the multiplicity of exposure tasks and truth measures that are assumed to be equivalent or have no influence on the effect (e.g., examining the extent to which the effect is reliant on people not judging truth during the exposure task). Second, there is a need to increase the reliability of illusory truth research by standardising the exposure task and establishing validated truth measures. Last, future research should increase the external validity and generalisability of the effect by varying the number of repetitions, intervals, stimuli, and participants - in particular, recruiting participants from non-Western countries and using a variety of

stimuli such as those relating to politics and news headlines. These lines of research will extend our knowledge to situations closer to real-world repetitions embedded in everyday life.

To my knowledge, this work represents the first example of systematic mapping in the psychology literature. Therefore, in addition to synthesising the literature and stimulating future research, it provides a guide for others who wish to take this approach.

Chapter 4: Is the illusory truth effect immune to time? The size of the illusory truth effect decreased as the time between the exposure and test phases increased. The few previous studies that varied intersession interval also manipulated other factors and reported mixed results. Hence the aim of this longitudinal experiment was to use the basic illusory truth design to focus only on the time manipulation using a well-powered design. With 567 participants analysed, the experiment had 95% power to detect the smallest effect reported in previous literature. The stimuli were evaluated in a norming study to ensure that the actual truth of the trivia statements would be generally unknown, and I used cumulative link mixed models to model stimuli as well as participant variability.

Using a within-subjects design, participants initially read the to-be-repeated statements and assigned each statement to a topic category. Then four times, after different intervals (i.e., immediately, one day, one week, one month), participants rated the truth of these statements, along with equal numbers of previously unseen statements. In contrast to the meta-analysis that found the effect was unaffected by time (Dechêne et al., 2010), here the effect diminished at every interval. The method section, along with supplementary materials (including a video of the experimental

procedure) provide a recipe for replicating the effect, and the “Constraints on Generality” section specifies the conditions under which I expect the effect to replicate.

The repetition-by-interval interaction implies that when information is repeated across short timescales, it may have a greater effect on truth judgements than repetitions that are weeks apart. If the effect generalises beyond the lab, such an interaction is particularly concerning in contexts such as targeted misinformation campaigns where falsities are communicated frequently over short timescales. Beyond the potential real-world consequences, the result has important implications for theory. Until now, largely due to the results of the meta-analysis by Dechêne and colleagues (2010), illusory truth effect research has usually proceeded on the assumption that the delay between exposure and test phases did not influence the effect of repetition (but see Nadarevic & Erdfelder, 2014; Silva, Garcia-Marques, et al., 2017) and was therefore an irrelevant feature of the design. For example, “... the effect size of the illusory truth effect is unrelated to the interval between the presentation and test phase.” (De keersmaecker, Roets, Pennycook, & Rand, 2018, p. 14). This is an illustration of an auxiliary hypothesis that we assume to be true whenever we test our main hypothesis (Uygun-Tunç & Tunç, 2020). When the main theoretical hypothesis and auxiliary hypotheses are bundled together, as is typical, it is hard to interpret what apparently consistent or inconsistent results mean for the main hypothesis (Uygun-Tunç & Tunç, 2020). Based on the results from Chapter 4, researchers should now consider the intersession interval a critical element of their research design⁴⁵.

⁴⁵ If the design falls within the conditions we expect the effect to generalise to, as described in the “Constraints on Generality” section.

Exploratory analyses focused on factors that might have modulated the illusory truth effect, such as attention during exposure phase, measured using accuracy during the topic categorisation task, test phase completion time, and participant age. There was little variation in the categorisation task: Most participants were very accurate and task performance was largely unrelated to the effect. Similarly test phase completion time seemed unrelated to the size of the illusory truth effect. However, the trajectory of the effect over time did vary with age. Compared to younger participants, older participants displayed a larger truth effect at the immediate interval and a smaller effect after 1 month. This analysis was exploratory, and the experiment was not powered for this test, although I did have a reasonable range of ages as a result of collecting data online. The reliability of this finding should therefore be tested in an experiment targeted at investigating the interaction between age and the effect over time.

5.2 Synthesis

This thesis provides evidence both for and against the existence of truth effects. On the one hand, the linguistic concreteness effect likely does not exist. On the other, the illusory truth effect appears to be robust, especially with trivia statements and over short timescales. Both findings contradict prior literature. Taken together, the results of all three empirical chapters point to a range of issues within previous truth effects research. Specifically, despite being well powered for the effect of interest and designed in consultation with the first author of the original experiment, the two replication experiments in Chapter 2 did not observe the intended effect. The most likely explanation is that the original was a false positive. The cataloguing study in Chapter 3 identified important gaps in the literature and issues around methodological standardisation, transparency, and publication bias.

Finally, in Chapter 4 the results of the extension study (designed, unlike most previous research, to manipulate only the effect of interest) contradicted the meta-analysis on the topic by Dechêne et al. (2010), who concluded that delay between sessions did not moderate the illusory truth effect. The latter observation, along with the other issues identified in Chapter 3, raise doubts about the reliability of the findings reported in that meta-analysis. These doubts are amplified by the fact that the meta-analysis combines the results of multiple studies conducted before many in the field started adopting more rigorous research practices.

Although the aim of close replications is to recreate the conditions of the original experiment as directly as possible, no replication can be exact (Simons, 2014; Stroebe, 2019) and this raises questions regarding the interpretation of the findings. There are many reasons why a replication study might produce a different outcome from the original. The original finding may have been a false positive. That possibility would be made more likely if there were analytic and reporting flexibility (Gelman & Loken, 2014; Simmons et al., 2011) or through selective reporting of “positive” results (Greenwald, 1975; Sterling et al., 1995). Alternatively, the replication attempt might be a false negative due to lack of power or play of chance. Or as some commentators have argued, an unsuccessful replication attempt might reflect a failure to understand the complexity of the original effect or methods (e.g., Gilbert, King, Pettigrew, & Wilson, 2016). In such cases, failing to elicit the phenomenon of interest would indicate a failure in the method rather than a challenge to the original finding (Ebersole et al., 2020).

I attempted to minimise the possibility of a false negative by using sufficient power and the original measures, materials, and procedures. Additionally, I adapted some stimuli to reflect cultural differences between our participants and the original

sample, and analysed both the original and adapted stimuli sets. The first author endorsed the replication attempt, agreeing that my design included the (known) features required to observe the effect and that any changes were theoretically irrelevant (based on our current understanding). After the study's completion, they offered no *post hoc* "hidden moderator" context-based explanation. It is therefore unlikely that the failure to replicate was due to poor replication design, and more likely that the original was a false positive.

With this in mind, the results should increase uncertainty about the reliability of other truth effects research that uses subtle manipulations, particularly those that were conducted prior to the credibility revolution, and that have been infrequently studied since. Some examples include the effect of perceptual fluency on truth, as studied by comparing trivia statements printed in "highly visible" colours such as dark blue and red, compared to "moderately visible" colours including light blue, green, and yellow (Reber & Schwarz, 1999)⁴⁶. Or other linguistic fluency effects, such as the "rhyme as reason effect" whereby unfamiliar rhyming aphorisms (e.g., "Woes unite foes") were judged to be more accurate than their non-rhyming equivalents (e.g., "Woes unite enemies"; McGlone & Tofiqbakhsh, 2000). That is not to say that these effects will not reliably replicate but that we should be sceptical of such effects until we have stronger evidence.

This view is supported by the results of a multi-lab, registered replication report that investigated whether subtle differences in wording resulted in different judgements of guilt (Eerland et al., 2016). In the original study participants judged behaviour described using the imperfective aspect (i.e., what a person *was doing*) as

⁴⁶ This result is reported as $p < .05$, one-tailed. In the absence of a preregistration we do not know whether the decision to use a one-tailed test was a priori or post hoc.

being more intentional, and they also imagined the behaviours in more detail, than behaviour described using perfective aspect (i.e., what a person *did*). Twelve close replication attempts failed to detect an effect of grammatical aspect, and several labs found results in the opposite direction to the original. Together, this result and those in Chapter 2 suggest that people do not pay attention to subtle differences in wording when inferring truth or guilt. Or that a small manipulation cannot drive a large effect.

Repetition provides a stronger, more obvious manipulation, the effects of which are investigated in Chapter 4. Beyond the key finding that the effect diminishes with time, we can learn several factors from this study. First, in what is one of the largest and arguably most controlled single studies of the illusory truth effect to date, the effect appears robust, certainly over short timescales with trivia statements as stimuli. Second, the effect occurs in the majority of participants. Overall illusory truth effect scores calculated for each participant (combined across all statements), revealed that 85% of participants displayed the predicted effect. When analysed at phase-level, the proportion of participants displaying the effect reduced with time. Both observations lend indirect support to the theory that the effect is not moderated by individual differences (see De keersmaecker et al., 2019), and that memory processes underlie the effect.

Third, the effect of repetition still appears after one month based on a single repetition. The fact that the effect persists over time means that it could apply in the real world. The negative implications of the effect are clear but there is a potential positive application of the effect: Repetition increases people's beliefs in information irrespective of the actual truth or falsity of the content. Consequently, the effect can be used to reinforce important socio-political messages communicating true information (see Unkelbach & Speckmann, 2021). But with a note of caution that

when communicating evidence the aim should be to inform, not persuade (see Blastland, Freeman, van der Linden, Marteau, & Spiegelhalter, 2020). Plainly, no single strategy is enough to fight misinformation, but with an awareness of the negative and positive implications of the effect, and a better understanding of how the effect changes with time, the illusory truth effect can be included as an intervention in our misinformation toolbox.

Even though the experiment was not designed to test a particular theory of the underlying mechanisms of the illusory truth effect, the results are also illuminating in this area, and lend support to certain explanations. The recognition, familiarity, and processing fluency explanations are all predicated on the idea that feelings of familiarity or fluency are misattributed as being informative about the truth of a statement. All three explanations predict the pattern observed in Chapter 4: larger illusory truth effects for recently repeated items, and smaller effects as feelings of recognition, familiarity, or fluency fade over time. However, a diminishing effect over time contradicts the source dissociation hypothesis. This theory holds that when people recall the semantic content of a statement but not its source, they may attribute the statement to multiple sources outside the experiment (Arkes et al., 1991, 1989). This in turn provides convergent validity, because numerous sources are perceived to have repeated the statement. Were this to be the case, the size of the illusory truth effect should increase as memories of the source fade with time. Yet I find the opposite.

Although the results generally support the most likely explanation of illusory truth - the processing fluency account, they have important implications for discounting, a mechanism that is thought to underpin fluency. When people notice the source of fluency (e.g., repetitions within the experiment), they realise it is

irrelevant to the judgement at hand, and discount it as informative (Alter & Oppenheimer, 2009). Indeed it seems intuitive that “people might get sceptical of statements repeated in close succession” (Bacon, 1979, p. 247). However, this does not reflect the pattern of results in the longitudinal study: The illusory truth effect is largest at the “immediate” interval, when participants have seen the repeated statements moments before. Since the source of the repetition is the experiment, participants should discount processing fluency founded on temporally close repetitions. The fact that the pattern of results contradicts this prediction, indicates that if processing fluency underlies the illusory truth effect, people do not discount repetition as an informative source, even when they are aware of it. Consequently, the most promising explanation of illusory truth proves to be an unsatisfactory account in this case.

The presence and direction of the repetition-by-interval interaction also contradicts some previous work investigating the illusory truth effect over time (e.g., Nadarevic & Erdfelder, 2014). Issues highlighted in Chapter 3 may provide some insight into potential explanations. As noted, there is no standard illusory truth effect paradigm, and results assumed to be based on manipulated factors (i.e., interval) cannot be disentangled from the influence of other changes to the experimental procedure, such as the exposure and retention tasks. For example, Nadarevic and Erdfelder (2014, Experiment 1) did not observe the illusory truth effect with an exposure task rating interestingness, a ten-minute delay, and a number puzzle retention interval task. Here it is possible that the exposure task, that might tap into similar processes as explicit truth judgements, could explain the result. Similarly, when comparing the present study to previous research with no power analyses, relatively small samples sizes, and inferential statistics that do not model stimuli, the

associated obfuscating consequences cannot be ruled out as explanations for inconsistent findings.

On the basis of these results, there are several other interesting avenues for future research. As I have highlighted throughout this thesis, most illusory truth research uses trivia statements as stimuli. I used trivia statements in the longitudinal study specifically because the effect appears reliable with these stimuli, and therefore the experimental design could isolate the effect of time. Consequently, the effect should be replicated using topics relating to dearly-held beliefs, such as political statements. Similarly, the design used one just repetition, so we do not know how frequency of repetitions interacts with delay. Last, the experiment provided preliminary evidence that age interacts with the effect over time. A focused investigation into how age moderates the effect would be interesting in and of itself, and could provide insights into the effect's reliance on memory.

5.3 Limitations of Illusory Truth Effects Research

Based primarily on the findings from Chapter 3, this section takes a step back and presents a general overview of the limitations that affect illusory truth effects research as a whole. I include suggestions for potential solutions, several of which were used within this thesis. Although replicability has been the primary focus of the credibility revolution (Finkel, Eastwick, & Reis, 2017), replication does not guarantee that the manipulation worked as intended, or that the measures and interpretations are valid (Nosek et al., 2021). Therefore, I consider the limitations of illusory truth effects research through the lens of four types of validity: statistical conclusion validity, internal validity, external validity, and construct validity (Fabrigar et al., 2020; Shadish, Cook, & Campbell, 2002).

5.3.1 Statistical conclusion validity

Statistical conclusion validity holds when reasonable conclusions have been drawn regarding the relationship between the variables of interest.

Problem: Statistical validity is violated when two types of error occur: A Type I error/false positive (i.e., incorrectly concluding that an effect exists), or a Type II error/false negative (i.e., incorrectly concluding that there is no effect). Threats to statistical conclusion validity include QRPs (e.g., John et al., 2012; Simmons et al., 2011) such as p-hacking and HARKing, low statistical power, and unreliable measures. As highlighted in Chapter 1 (section 1.7.1) the number of psychology researchers that have engaged in QRPs is disconcertingly high (John et al., 2012). The proximal antidotes for QRPs include well-specified preregistrations or Registered Reports, and at a distal range, improving research incentives and culture. However, as noted in Chapter 3, only 16% of papers included one or more preregistered studies⁴⁷, and many of those lacked detail. Sharing of raw data and code was relatively rare, as was the reporting of any type of sampling plan (e.g., power analysis). Although these findings are not surprising considering that the focus on these practices is relatively new, it means that we do not have the tools necessary to evaluate the statistical validity of most illusory truth research⁴⁸.

Solution: I took several precautions to increase the validity of my statistical conclusions. Those measures included using the Registered Reports publishing format. By creating a two-stage process, “Registered Reports represent a major advance for statistical conclusion validity” (Vazire, Schiavone, & Bottesini, 2020 p.4). As discussed in Chapter 1, this publication format reduces bias by creating a

⁴⁷ I coded preregistration at paper level. Therefore if, for example, a paper included four studies but only one was preregistered, the paper was coded as having been preregistered.

⁴⁸ For example, the first preregistration in the literature was in 2017.

process where all decisions (hypotheses, methodology, sample size, analysis plan, publication decision) occur *a priori*, meaning that QRPs such as p-hacking and HARKing are neither possible nor incentivised. HARKing is particularly egregious in terms of statistical validity because when hypotheses are altered *post hoc* to follow the data, there is no way they can be disconfirmed. As detailed in Chapters 2 and 4, all experiments within this thesis were adequately powered to detect the effect of interest, and false-negative and false-positive rates were balanced by using 95% power and establishing a Type I error rate of 5%. To improve the statistical validity of truth effects research, these practices need to become the norm. This can be accomplished in part by offering training in such practices, and by promoting collaborative research that increases both sample sizes and team expertise.

5.3.2 Internal validity

Internal validity relates to the extent to which a study establishes a cause-and-effect relationship.

Problem: All illusory truth research assumes a causal relationship between repeated statements and increased truth judgements. In a tightly controlled experimental setting with randomisation this is a reasonable inference. An alternative explanation is that the effect is merely an artefact of demand characteristics. Demand effects represent a bias that occurs when participants infer the experimental hypotheses and attempt to validate them by behaving in line with perceived expectations (Orne, 1962). Though recent studies question the existence of such effects, especially for online research (e.g., Mummolo & Peterson, 2018). Perhaps more pertinent is the possibility that by giving participants just one parameter with which to rate statements, they have no choice but to use the truth scale to communicate any experiences or feelings, whether or not they relate to truth. We

may just be capturing a feeling of fluent processing for familiar statements, rather than truth.

Solution: One way of interrogating this causal relationship would be to vary the wording of the truth judgement questions or give participants a variety of measures on which to rate the statements. Furthermore, measuring truth judgements is the equivalent of measuring a behavioural intention rather than measuring behaviour itself. Theoretically if an individual believes that something is true versus false, it will affect their behaviour. Therefore, future research could focus on measuring the behaviours that naturally extend from belief. For example, willingness to place money on statements that feel true.

5.3.3 External and ecological validity

External validity examines whether the findings of study generalise to other contexts beyond those in the study (e.g., other people, stimuli, or measures), and ecological validity refers specifically to whether the study findings can be generalised to real-life settings.

Problem: Presumably most authors would like to draw inferences from their studies that apply beyond their specific participants and stimuli. Yet almost no studies specify which populations the finding should generalise to, and we might therefore assume that the study's generality is unconstrained (see Simons, Shoda, & Lindsay, 2017). However, as previously discussed, the generality of the illusory truth effect is currently constrained because most studies have been designed to present one repetition of trivia statements to university students. There is a scarcity of evidence regarding how the effect works in real-world contexts where it might be

most important to understand, such as health, politics, and the environment, and on larger numbers of repetitions over various timescales.

Solution: In Chapter 4 the external validity of the longitudinal experiment is explicitly reported in the “Constraints on Generality” section. Specifically, I discuss the populations I expect the effect to generalise to, and the conditions under which I expect it to replicate. Including Constraints on Generality sections in future truth effects research will allow others to evaluate where additional research on generality is required (see Simons et al., 2017). Furthermore, future research should focus investigating the illusory truth effect in conditions where misinformation may be rife, such as political contexts.

5.3.4 Construct validity

Construct validity captures the extent to which inferences made about the measured variables relate to the intended construct.

Problem: Belief, as measured by truth judgements, is a latent variable that cannot be observed or measured directly. It is unlikely that any operationalisation of an effect will be a pure reflection of the intended construct (Fabrigar et al., 2020). Moreover, within the illusory truth effect literature there is no agreement about how belief should be measured. As highlighted in the systematic map, there is great heterogeneity in the measures used, with 19 different truth measures coded including dichotomous judgements and Likert-type scales with and without mid-points. There is likely to be some variability associated with these different scales. Using just one measure of belief within an experiment means that we cannot easily separate what reflects a true change in belief and what is due to a combination of a change in belief and the task or measure.

Solution: To increase latent construct validity, future research could use multiple tasks and measures within an experiment to distinguish effects specific to the methodology from those relating to the underlying belief structure.

5.4 Limitations of the Present Research

The individual limitations of each study are described in the corresponding chapter. I will not repeat the limitations here but instead focus on the real-world generalisability of the repetition-by-interval interaction, and the online testing setting that was used for two of the three experiments reported in this thesis.

5.4.1 Real-world generalisability

Chapter 4 provides compelling evidence for a repetition-by-interval interaction when investigated in a controlled experimental setting using trivia statements. Therefore, we cannot make broad claims about the repetition-by-interval interaction and other types of topics, such as those relating to healthcare or politics. And we do not know to what extent the result is reliant on the selected exposure task where participants categorised statements. The observation that the effect of a single repetition persists, albeit diminished, after one month indicates that the effect is strong enough to endure beyond a single experimental session. But this effect is based on a somewhat heavy-handed manipulation – multiple repeated statements interspersed with new statements. And when illusory truth is tested with all repeated stimuli, rather than mixed lists of repeated and new, the effect disappears (Dechêne, Stahl, Hansen, & Wänke, 2009). Thus, we do not know whether more subtle, real-world repetitions, embedded in real experience, would show the same effect. Consequently, we cannot say with certainty that the repetition-by-interval interaction will occur outside the structured experimental experience.

5.4.2 Online experimental testing

Conducting studies online is one way to collect increased sample sizes more easily. However remote testing has been criticised because the participants are “invisible”, and we cannot directly observe their behaviour. Face-to-face research permits researchers to meet participants, verify basic demographic information, ensure that participants have not completed the research before, control the setting of the experiment, and observe participants during the experiment (Rodd, 2019). In contrast, when research is conducted online, there is a higher degree of uncertainty about who the participants are, and the experimental setting (Rodd, 2019). These factors are particularly important for research where the experimental outcomes are reliant on aspects of the setting such as a lack of noise, specific lighting, or on complex computer software performing.

To overcome this apparent lack of control, I used a range of methods and preregistered exclusion criteria designed to maximise data quality and minimise the effect of careless responders. These methods are detailed in their respective chapters but to summarise briefly, they included: 1) Paying participants a fair hourly rate; 2) Using Prolific’s prescreener questions to recruit participants with appropriate demographic features and high approval ratings; 3) Excluding data from participants who failed to complete the experiment within the reasonable minimum and maximum completion times; 4) Excluding data from participants who responded uniformly across an entire phase of the study; 5) Using participant performance in the lab study (where participants were observed) as a guide for excluding potential cheaters in the online study (Chapter 2); 6) Using performance on the exposure task to check participant attention (Chapter 4).

A second criticism levelled at online research is that in facilitating recruitment of large, “easy to access” and “inexpensive” samples, online platforms support research that can compromise construct and external validity (Finkel et al., 2017). The research process involves a sequence of interconnected decisions in which we try to balance conflicting scientific goals simultaneously (McGrath, 1981). Research carried out in the controlled conditions of an online experiment is necessarily a trade-off between some goals (e.g., establishing causality, having sufficient power to detect the effect of interest) and others that are inherently incompatible, namely external validity (McGrath, 1981).

There is no way to prioritise *all* the features of high-quality science in a single study with finite resources. Instead, decisions are made to optimise an experiment’s contribution and informativeness based on the features that are important for that research context and question (Finkel et al., 2017). In the replication and extension studies in Chapters 2 and 4, the primary aims of the studies were to establish the presence or absence of the effects and provide an estimate of their size. Therefore, the appropriate research practices included a controlled experimental setting, close replication, and adequate power. Furthermore, real behaviour occurs online: reading the news, discussions on social media, and work (among other things) all take place on phones and computers. Therefore to understand how people behave in the real world, we need to understand how they behave online (Danvers, 2019). This includes making decisions about what articles to read and share, and judgements about their truth. Thus, while online research may not be suitable for more elaborate scenarios, fundamental research on truth judgements is highly compatible with online experimental methods. That said, when considering the field as a whole, more research that considers other cues, such as

who is communicating and in what setting, will help to establish the validity and generalisability of truth judgements research across contexts (Anderson et al., 2019).

Overall, online data collection, while not a panacea, enables researchers to collect larger samples from beyond the typical university student pool (Ledgerwood et al., 2017). Using the Prolific platform allowed me to collect data that I would not otherwise have been able to collect (i.e., data from 608 participants, returning at four time points over the course of a month, while the UK was in COVID-19 lockdown). Furthermore, in Chapter 2 (see sections 2.2.2.1 and 2.3.2.1) where I conducted the same study in the lab and online, the results were extremely similar in terms of variance and effect sizes ($d_z = 0.08$; 95% CI: [-0.03, 0.18] lab study, $d_z = 0.11$; 95% CI [-0.01, 0.22] online).

5.5 Future Directions

Chapter 3 details the gaps in the literature and directions for future research identified via systematic mapping of the entire (available) illusory truth literature. Here I do not recap them all but instead I propose future research that investigates the real-world generalisability of the effect.

5.5.1 Gist repetition

In Chapter 3 I noted that most illusory truth effect research (82%) uses identical statements at exposure and test. Although verbatim repetition is a strategy sometimes used by politicians and advertisers, such repetition is not representative of most information acquisition in the real world. The majority of claims in our environment are based on semantically similar content repeated by different sources with variations in language. Previous research on the illusory truth effect suggests that the effect occurs for verbally similar but semantically contradictory stimuli

(Garcia-Marques, Silva, Reber, & Unkelbach, 2015) but for real-world generality we need more research on verbally different but semantically similar statements.

5.5.2 Avoiding the effect

When asked about future research at the end of the longitudinal study reported in Chapter 4, many participants showed interest in whether the effect can be avoided if one is aware of it. Research on false information shows that attempts to tag statements as false may ironically increase their truthiness as the statements become more familiar with repetition while the context is forgotten (Skurnik et al., 2005). Thus, efforts to warn people about the illusory truth effect (i.e., repeated statements feel more true), rather than correcting falsehoods at the point of consumption, could be a more effective strategy. Previous research found that such warnings reduced the illusory truth effect when the exposure and test phases were in the same session but had no effect with an intersession interval of 1 week, though the authors note they may have been underpowered for the 1-week retention (Nadarevic & Aßfalg, 2017).

Research similar to the longitudinal study in Chapter 4, but which includes instructions to *avoid* the effect, would be informative about whether the illusory truth effect can potentially be prevented over short and long timescales. In addition, it would be enlightening to manipulate when the instructions are given. Nadarevic and Aßfalg provided instructions after the exposure phase but before the test phase, however instructions prior to the exposure phase might produce a stronger effect. Such studies would complement work on the effectiveness of fact-checking (e.g., Brashier, Pennycook, Berinsky, & Rand, 2021) performed before (prebunking), during (labelling), and after exposure (debunking).

5.5.3 Manipulating truth base rates

In the Introduction I discussed people's natural propensity towards truth, a reflection of the base rates in daily life where most incoming information is true. However, base rates will vary depending on the context. For example, in settings such as the doctor's surgery or within the classroom one might expect especially high levels of honesty. There are also situations where we expect high frequencies of lies, for example, police interviews or political campaigns. It is possible that in contexts with high base rates of lying, repetitions reinforce falsity rather than truth. Future investigations could examine the impact of source of information, both in respect to who is communicating and in what setting.

5.5.4 Metric calibration

In Chapter 4 I observed an illusory truth effect at 4 time points and concluded that repetition makes statements feel truer. As with other constructs measured in psychology, the subjective feeling of truth is not directly observable but inferred via ratings on a Likert-type scale. At the immediate intersession interval, the difference in ratings between new and repeated statements was 0.68, and 0.14 after a month. Yet these observed changes in perceived truth do not reflect the magnitude of change on the unobserved dimension of truth (Blanton & Jaccard, 2006). The concrete implications of such changes deserve further research, especially in how they calibrate to meaningful real-world outcomes. For example, what change in perceived truth would be enough for someone to bet money on the statement being true, or tell a family member the information as a trivia fact, or persuade the reader to have a vaccination, or to vote in a certain way? Future research should focus on developing non-arbitrary truth metrics that link changes in ratings to meaningful behaviours (Blanton & Jaccard, 2006), or as a first step, to behavioural intentions. This research

programme is non-trivial and challenging but with a better understanding of how changes on a Likert-type scale equate to changes in opinions, choices, and behaviour, we will know what effect sizes we should care about, and we can develop meaningful smallest effect sizes of interest (SESOI) that can be used to determine power in new studies.

5.6 Conclusion

Arguably there has never been a more important time to understand how people decide what is true and what is not, and to learn about the factors that lead to inflated truth judgements, particularly when those factors are independent of the message content. As the misinformation age gathered impetus in the 2010s, scientists realised that they were not equipped with the most appropriate methods or incentives to investigate this behaviour in an unbiased manner. Throughout this thesis I have expounded the values of transparent research practices in producing robust, replicable findings. It is clear that these are not tangential issues: The consequence of our old methods is a literature populated by false positives in which we cannot distinguish what is true and what is not. Just as misinformation misleads the public, so too the false positives in our literature misdirect scientists, causing them to chase false effects. We should now be sceptical about studies that did not adopt best practices, especially those with small sample sizes and without preregistrations, because when their findings are tested using open, unbiased practices, effects do not replicate, and claims compiled in meta-analyses do not hold.

This thesis has shown the importance of employing those best practices: Designing studies with predetermined, analysis-specific sample sizes for predefined hypotheses, using preregistered methods and analysis plans, and making those plans and associated outputs open. Through their transparency and openness, the studies

presented here have progressed our understanding of truth effects and we now have a better grasp of those effects that are robust and those that are not. Throughout I focused on replicability, attempting to replicate previous work, designing studies with an increased likelihood of replicability, and providing the materials and recipes necessary to attempt those replications. Furthermore, the transparent methods applied allow for computational reproducibility. As a result of the methods employed, the research herein might now be considered the starting point for future truth effects research. Overall, this thesis illustrates that when truth effects research uses rigorous, transparent, and unbiased methods, it paints a different picture from that of the existing literature.

References

- Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., ... Iannone, R. (2018). rmarkdown: Dynamic Documents for R [Computer software]. Retrieved from <https://CRAN.R-project.org/package=rmarkdown>
- Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., ... Iannone, R. (2020). rmarkdown: Dynamic Documents for R [Computer software]. Retrieved from <https://github.com/rstudio/rmarkdown>
- Allen, C., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLoS Biology*, *17*(5), e3000246. <https://doi.org/10.1371/journal.pbio.3000246>
- Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235. <https://doi.org/10.1177/1088868309341564>
- Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rökkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin*, *45*(6), 842–850. <https://doi.org/10.1177/0146167218798821>
- Arkes, H. R., Boehm, L. E., & Xu, G. (1991). Determinants of judged validity. *Journal of Experimental Social Psychology*, *27*(6), 576–605. [https://doi.org/10.1016/0022-1031\(91\)90026-3](https://doi.org/10.1016/0022-1031(91)90026-3)
- Arkes, H. R., Hackett, C., & Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making*, *2*(2), 81–94. <https://doi.org/10.1002/bdm.3960020203>

- Asp, E., Khan, L., Jonason, A., Adkins-Hempel, M., Warner, K., Pardilla-Delgado, E., ... Tranel, D. (2020). Second-guessing of Spinoza: Psychophysiological and behavioral evidence that believing is default during proposition comprehension. *PsyArXiv*. <https://doi.org/10.31234/osf.io/s3tfk>
- Aust, F., & Barth, M. (2018). *papaja*: Create APA manuscripts with R Markdown [Computer software]. Retrieved from <https://github.com/crsh/papaja>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Bacon, F. T. (1979). Credibility of repeated statements: memory for trivia. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(3), 241–252. <https://doi.org/10.1037/0278-7393.5.3.241>
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. <https://doi.org/10.1177/1745691612459060>
- Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Cromptoets, E. A. V., Ong, H. H., Nosek, B. A., ... Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLoS Biology*, *18*(12), e3000937. <https://doi.org/10.1371/journal.pbio.3000937>
- Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 328. <https://doi.org/10.3389/fpsyg.2013.00328>

- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bastos, M. T., & Mercea, D. (2017). The Brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, *37*(1), 38–54. <https://doi.org/10.1177/0894439317734157>
- Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, *121*(4), 446–458. <https://doi.org/10.1037/0096-3445.121.4.446>
- Begg, I. M., & Armour, V. (1991). Repetition and the ring of truth: biasing comments. *Canadian Journal of Behavioural Science*, *23*(2), 195–213. <https://doi.org/10.1037/h0079004>
- Begg, I. M., Armour, V., & Kerr, T. (1985). On believing what we remember. *Canadian Journal of Behavioural Science*, *17*(3), 199–214. <https://doi.org/10.1037/h0080140>
- Belmore, S. M., Yates, J. M., Bellack, D. R., Jones, S. N., & Rosenquist, S. E. (1982). Drawing inferences from concrete and abstract sentences. *Journal of Verbal Learning and Verbal Behavior*, *21*(3), 338–351. [https://doi.org/10.1016/S0022-5371\(82\)90659-4](https://doi.org/10.1016/S0022-5371(82)90659-4)
- Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. <https://doi.org/10.1037/a0021524>

- Beukeboom, C. J., Tanis, M., & Vermeulen, I. E. (2013). The language of extraversion. *Journal of Language and Social Psychology, 32*(2), 191–201. <https://doi.org/10.1177/0261927X12460844>
- Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *The American Psychologist, 61*(1), 27–41. <https://doi.org/10.1037/0003-066X.61.1.27>
- Blastland, M., Freeman, A. L. J., van der Linden, S., Marteau, T. M., & Spiegelhalter, D. (2020). Five rules for evidence communication. *Nature, 587*(7834), 362–364. <https://doi.org/10.1038/d41586-020-03189-1>
- Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: accuracy and bias. *Psychological Bulletin, 134*(4), 477–492. <https://doi.org/10.1037/0033-2909.134.4.477>
- Bornstein, R. F., & D’Agostino, P. R. (1992). Stimulus recognition and the mere exposure effect. *Journal of Personality and Social Psychology, 63*(4), 545–552. <https://doi.org/10.1037/0022-3514.63.4.545>
- Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications, 10*(1), 1–14. <https://doi.org/10.1038/s41467-018-07761-2>
- Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... van ’t Veer, A. (2014). The replication recipe: what makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224. <https://doi.org/10.1016/j.jesp.2013.10.005>

- Brashier, N. M., Eliseev, E. D., & Marsh, E. J. (2020). An initial accuracy focus prevents illusory truth. *Cognition*, *194*, 104054.
<https://doi.org/10.1016/j.cognition.2019.104054>
- Brashier, N. M., & Marsh, E. J. (2019). Judging truth. *Annual Review of Psychology*, *71*(1), 499-515. <https://doi.org/10.1146/annurev-psych-010419-050807>
- Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(5), 1-3.
<https://doi.org/10.1073/pnas.2020043118>
- Brown, A. S., & Nix, L. A. (1996). Turning lies into truths: referential validation of falsehoods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1088–1100. <https://doi.org/10.1037/0278-7393.22.5.1088>
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. <https://doi.org/10.3758/s13428-013-0403-5>
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience*, *14*(5), 365–376. <https://doi.org/10.1038/nrn3475>
- Calvillo, D. P., & Smelter, T. J. (2020). An initial accuracy focus reduces the effect of prior exposure on perceived accuracy of news headlines. *Cognitive Research: Principles and Implications*, *5*(1), 55.
<https://doi.org/10.1186/s41235-020-00257-y>

- Chambers, C. D. (2013). Registered Reports: a new publishing initiative at Cortex. *Cortex*, 49(3), 609–610. <https://doi.org/10.1016/j.cortex.2012.12.016>
- Chambers, C. D., & Tzavella, L. (2020). Registered Reports: past, present and future. MetaArXiv. <https://doi.org/10.31222/osf.io/43298>
- Chou, H.-Y., & Yeh, M.-H. (2018). Minor language variations in campaign advertisements: the effects of pronoun use and message orientation on voter responses. *Electoral Studies*, 51, 58–71. <https://doi.org/10.1016/j.electstud.2017.10.006>
- Christensen, R. H. B. (2019). ordinal --- Regression models for ordinal data [Computer software]. Retrieved from <https://cran.r-project.org/web/packages/ordinal/index.html>
- Claesen, A., Gomes, S. L. B. T., Tuerlinckx, F., & Vanpaemel, W. (2019). Preregistration: comparing dream to reality. <https://doi.org/10.31234/osf.io/d8wex>
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12), 1304–1312. <https://doi.org/10.1037/0003-066X.45.12.1304>
- Coles, N. A., Tiokhin, L., Scheel, A. M., Isager, P. M., & Lakens, D. (2018). The costs and benefits of replication studies. PsyArXiv. <https://psyarxiv.com/c8akj/>
- Comtois, D. (2018). Summarytools: tools to quickly and neatly summarize data [Computer software]. Retrieved from <https://CRAN.R-project.org/package=summarytools>

- Corker, K. S. (2018). Strengths and weaknesses of meta-analyses. PsyArXiv.
<https://doi.org/10.31234/osf.io/6gcnm>
- Dahl, D. B. (2016). Xtable: export tables to latex or html [Computer software].
Retrieved from <https://CRAN.R-project.org/package=xtable>
- Danvers, A. (2019, June 23). Credibility and its discontents. *Psychology Today*.
Retrieved May 14, 2021, from
<https://www.psychologytoday.com/us/blog/how-do-you-know/201906/credibility-and-its-discontents>
- De keersmaecker J., Dunning, D., Pennycook, G., Rand, D. G., Sanchez, C., Unkelbach, C., & Roets, A. (2019). Investigating the robustness of the illusory truth effect across individual differences in cognitive ability, need for cognitive closure, and cognitive style. *Personality and Social Psychology Bulletin*, 46(2), 204–215. <https://doi.org/10.1177/0146167219853844>
- De keersmaecker J., Roets, A., Pennycook, G., & Rand, D. G. (2018). Is the illusory truth effect robust to individual differences in cognitive ability, need for cognitive closure, and cognitive style? *Unpublished manuscript*, 1–38.
- de Groot, A. M. (1989). Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(5), 824–845.
<https://doi.org/10.1037/0278-7393.15.5.824>
- de Raadt, A., Warrens, M. J., Bosker, R. J., & Kiers, H. A. L. (2021). A comparison of reliability coefficients for ordinal rating scales. *Journal of Classification*.
<https://doi.org/10.1007/s00357-021-09386-5>

- Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2009). Mix me a list: context moderates the truth effect and the mere-exposure effect. *Journal of Experimental Social Psychology, 45*(5), 1117–1122.
<https://doi.org/10.1016/j.jesp.2009.06.019>
- Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: a meta-analytic review of the truth effect. *Personality and Social Psychology Review, 14*(2), 238–257. <https://doi.org/10.1177/1088868309352251>
- Dowle, M., & Srinivasan, A. (2018). Data.table: extension of ‘data.frame’ [Computer software]. Retrieved from <https://CRAN.R-project.org/package=data.table>
- Dreyfuss, E. (2017, November 2). Want to make a lie seem true? Say it again. And again. And again. Retrieved December 18, 2018, from <https://www.wired.com/2017/02/dont-believe-lies-just-people-repeat/>
- Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., ... IJzerman, H. (2020). Many Labs 5: testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science, 3*(3) 309–331.
<https://doi.org/10.1177/2515245920958687>
- Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., ... Prenoveau, J. M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science, 11*(1), 158–171.
<https://doi.org/10.1177/1745691615605826>
- Effron, D. A., & Raj, M. (2019). Misinformation and morality: encountering fake-news headlines makes them seem less unethical to publish and share.

Psychological Science, 31(1), 75-87.

<https://doi.org/10.1177/0956797619887896>

Elliott, W. B., Rennekamp, K. M., & White, B. J. (2015). Does concrete language in disclosures increase willingness to invest? *Review of Accounting Studies*, 20(2), 839–865. <https://doi.org/10.1007/s11142-014-9315-6>

Everett, J. A. C., & Earp, B. D. (2015). A tragedy of the (academic) commons: interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology*, 6, 1152. <https://doi.org/10.3389/fpsyg.2015.01152>

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A validity-based framework for understanding replication in psychology. *Personality and Social Psychology Review*, 24(4) 316-344 <https://doi.org/10.1177/1088868320931366>

Fanelli, D. (2010). “Positive” results increase down the hierarchy of the sciences. *Plos One*, 5(4), e10068. <https://doi.org/10.1371/journal.pone.0010068>

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3), 891–904. <https://doi.org/10.1007/s11192-011-0494-7>

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>

Fazio, L. K. (2020). Repetition increases perceived truth even for known falsehoods.

PsyArXiv. <https://doi.org/10.31234/osf.io/2u53a>

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does

not protect against illusory truth. *Journal of Experimental Psychology:*

General, *144*(5), 993–1002. <https://doi.org/10.1037/xge0000098>

Fazio, L. K., Rand, D. G., & Pennycook, G. (2019a). Repetition increases perceived

truth equally for plausible and implausible statements. PsyArXiv.

<https://doi.org/10.31234/osf.io/qys7d>

Fazio, L. K., Rand, D. G., & Pennycook, G. (2019b). Repetition increases perceived

truth equally for plausible and implausible statements. *Psychonomic Bulletin*

& Review, *26*(5), 1705–1710. <https://doi.org/10.3758/s13423-019-01651-4>

Fazio, L. K., & Sherry, C. L. (2020). The effect of repetition on truth judgments

across development. *Psychological Science*, *31*(9), 1150–1160.

<https://doi.org/10.1177/0956797620939534>

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2017). Replicability and other features

of a high-quality science: toward a balanced and empirical approach. *Journal of Personality and Social Psychology*, *113*(2), 244–253.

<https://doi.org/10.1037/pspi0000075>

Firke, S. (2018). Janitor: simple tools for examining and cleaning dirty data.

[Computer software]. Retrieved from [https://CRAN.R-](https://CRAN.R-project.org/package=janitor)

[project.org/package=janitor](https://CRAN.R-project.org/package=janitor)

- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: unlocking the file drawer. *Science*, *345*(6203), 1502–1505.
<https://doi.org/10.1126/science.1255484>
- Fujita, K., Trope, Y., Liberman, N., & Levin-Sagi, M. (2006). Construal levels and self-control. *Journal of Personality and Social Psychology*, *90*(3), 351–367.
<https://doi.org/10.1037/0022-3514.90.3.351>
- Funder, D. (2014). Notice: PSPB articles by authors with retracted articles at PSPB or other journals: Stapel, Smeesters, and Sanna. *Personality and Social Psychology Bulletin*, *40*(1), 132–135.
<https://doi.org/10.1177/0146167213508152>
- Garcia-Marques, T., Silva, R. R., & Mello, J. (2017). Asking simultaneously about truth and familiarity may disrupt truth effects. *Análise Psicológica*, *35*(1) 61-71. <https://doi.org/10.14417/ap.1121>
- Garcia-Marques, T., Silva, R. R., Reber, R., & Unkelbach, C. (2015). Hearing a statement now and believing the opposite later. *Journal of Experimental Social Psychology*, *56*, 126–129. <https://doi.org/10.1016/j.jesp.2014.09.015>
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460–465.
- Gerlanc, D., & Kirby, K. (2015). BootES: bootstrap effect sizes [Computer software]. Retrieved from <https://CRAN.R-project.org/package=bootES>
- Gibbons, M. (1999). Science's new social contract with society. *Nature*, *402*, C81–4.
<https://doi.org/10.1038/35011576>

- Gigerenzer, G. (1984). External validity of laboratory experiments: the frequency-validity relationship. *The American Journal of Psychology*, *97*(2), 185–195.
<https://doi.org/10.2307/1422594>
- Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, *46*(2), 107–119. <https://doi.org/10.1037/0003-066X.46.2.107>
- Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on “Estimating the reproducibility of psychological science”. *Science*, *351*(6277), 1037. <https://doi.org/10.1126/science.aad7243>
- Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59*(4), 601–613. <https://doi.org/10.1037/0022-3514.59.4.601>
- Giner-Sorolla, R. (2019). From crisis of evidence to a “crisis” of relevance? Incentive-based answers for social psychology’s perennial relevance worries. *European Review of Social Psychology*, *30*(1), 1–38.
<https://doi.org/10.1080/10463283.2018.1542902>
- Giner-Sorolla, R., Carpenter, T., Lewis, Jr, N. A., Montoya, A. K., Aberson, C. L., Bostyn, D. H., ... Soderberg, C. (2019). Power to detect what? Considerations for planning and evaluating sample size. Retrieved from <https://osf.io/jnmya/>
- Gong, H., & Medin, D. L. (2012). Construal levels and moral judgment: some complications. *Judgement and Decision Making*, *7*(5) 628–638.

- Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, *66*, 877–902. <https://doi.org/10.1146/annurev-psych-010814-015321>
- Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *The American Psychologist*, *59*(2), 93–104. <https://doi.org/10.1037/0003-066X.59.2.93>
- Grames, E. M., Stillman, A. N., Tingley, M. W., & Elphick, C. S. (2019a). An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods in Ecology and Evolution*, *10*(10), 1645–1654. <https://doi.org/10.1111/2041-210X.13268>
- Grames, E. M., Stillman, A., Tingley, M., & Elphick, C. (2019b). litsearchr: Automated search term selection and search strategy for systematic reviews [Computer software]. Downloaded from <https://elizagrames.github.io/litsearchr/>
- Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20. <https://doi.org/10.1037/h0076157>
- Greifeneder, R., Bless, H., & Pham, M. T. (2011). When do people rely on affective and cognitive feelings in judgment? A review. *Personality and Social Psychology Review*, *15*(2), 107–141. <https://doi.org/10.1177/1088868310367640>
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (eds.), *Syntax and Semantics 3: Speech Arts* (Vols. 1–3, pp. 41–58). New York: Academic Press.

- Grice, J. W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M. (2020). Persons as effect sizes. *Advances in Methods and Practices in Psychological Science*, 3(4), 443–455.
<https://doi.org/10.1177/2515245920922982>
- Gunnell, K., Poitras, V. J., & Tod, D. (2020). Questions and answers about conducting systematic reviews in sport and exercise psychology. *International Review of Sport and Exercise Psychology*, 1–22.
<https://doi.org/10.1080/1750984X.2019.1695141>
- Haddaway, N. R. (2018). Open synthesis: on the need for evidence synthesis to embrace open science. *Environmental Evidence*, 7(1), 26.
<https://doi.org/10.1186/s13750-018-0140-4>
- Haddaway, N. R., Bethel, A., Dicks, L. V., Koricheva, J., Macura, B., Petrokofsky, G., ... Stewart, G. B. (2020). Eight problems with literature reviews and how to fix them. *Nature Ecology & Evolution*, 4(12), 1582–1589.
<https://doi.org/10.1038/s41559-020-01295-x>
- Haddaway, N. R., Collins, A. M., Coughlin, D., & Kirk, S. (2015). The role of google scholar in evidence reviews and its applicability to grey literature searching. *Plos One*, 10(9), e0138237.
<https://doi.org/10.1371/journal.pone.0138237>
- Haddaway, N. R., Feierman, A., Grainger, M. J., Gray, C. T., Tanriver-Ayder, E., Dhaubanjari, S., & Westgate, M. J. (2019). EviAtlas: a tool for visualising evidence synthesis databases. *Environmental Evidence*, 8(1), 22.
<https://doi.org/10.1186/s13750-019-0167-1>

- Haddaway, N. R., Macura, B., Whaley, P., & Pullin, A. (2018a). ROSES flow diagram for systematic maps. Version 1.0. *Figshare*.
<https://doi.org/10.6084/m9.figshare.6085940>
- Haddaway, N. R., Macura, B., Whaley, P., & Pullin, A. S. (2018b). ROSES RepOrting standards for Systematic Evidence Syntheses: pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. *Environmental Evidence*, 7(1), 7.
<https://doi.org/10.1186/s13750-018-0121-7>
- Hansen, J., & Wänke, M. (2010). Truth from language and truth from fit: the impact of linguistic concreteness and level of construal on subjective truth. *Personality and Social Psychology Bulletin*, 36(11), 1576–1588.
<https://doi.org/10.1177/0146167210386238>
- Hardwicke, T. E., Wallach, J. D., Kidwell, M. C., Bendixen, T., Crüwell, S., & Ioannidis, J. P. A. (2020). An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014-2017). *Royal Society Open Science*, 7(2), 190806.
<https://doi.org/10.1098/rsos.190806>
- Harrell, F. E. (2015). Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis. New York: Springer. <https://doi.org/10.1007/978-3-319-19425-7>
- Harzing, A. W. (2007). Publish or Perish. Retrieved April 24, 2019, from <https://harzing.com/resources/publish-or-perish>

- Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107–112. [https://doi.org/10.1016/S0022-5371\(77\)80012-1](https://doi.org/10.1016/S0022-5371(77)80012-1)
- Hassan, A., & Barber, S. J. (2021). The effects of repetition frequency on the illusory truth effect. *Cognitive Research: Principles and Implications*, *6*(1), 38. <https://doi.org/10.1186/s41235-021-00301-5>
- Hawkins, S. A., & Hoch, S. J. (1992). Low-involvement learning: memory without evaluation. *The Journal of Consumer Research*, *19*(2), 212–225. <https://doi.org/10.1086/209297>
- Hawkins, S. A., Hoch, S. J., & Meyers-Levy, J. (2001). Low-involvement learning: repetition and coherence in familiarity and belief. *Journal of Consumer Psychology*, *11*(1), 1–11. https://doi.org/10.1207/S15327663JCP1101_1
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29. <https://doi.org/10.1038/466029a>
- Heycke, T., & Spitzer, L. (2019). Screen recordings as a tool to document computer assisted data collection procedures. *Psychologica Belgica*, *59*(1), 269–280. <https://doi.org/10.5334/pb.490>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.
- Huang, H.-W., & Federmeier, K. D. (2015). Imaginative language: what event-related potentials have revealed about the nature and source of concreteness effects. *Language and Linguistics*, *16*(4), 503–515. <https://doi.org/10.1177/1606822X15583233>

- Huxley, A. (1932). *Brave new world*. London: Macmillan Education UK.
- Iacobucci, G. (2019). Vaccination: “fake news” on social media may be harming UK uptake, report warns. *BMJ (Clinical Research Ed.)*, *364*, 1365.
<https://doi.org/10.1136/bmj.1365>
- Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, *58*(6), 543–549. <https://doi.org/10.1016/j.jclinepi.2004.10.019>
- Isager, P. M., van Aert, R. C. M., Bahník, Š., Brandt, M. J., DeSoto, K. A., Giner-Sorolla, R., ... Lakens, D. (2020). Deciding what to replicate: a formal definition of “replication value” and a decision model for replication study selection. *MetaArXiv*. <https://doi.org/10.31222/osf.io/2gurz>
- Jalbert, M., Newman, E., & Schwarz, N. (2020). Only half of what I’ll tell you is true: expecting to encounter falsehoods reduces illusory truth. *Journal of Applied Research in Memory and Cognition*, *9*(4), 602–613.
<https://doi.org/10.1016/j.jarmac.2020.08.010>
- James, K. L., Randall, N. P., & Haddaway, N. R. (2016). A methodology for systematic mapping in environmental sciences. *Environmental Evidence*, *5*(1), 7. <https://doi.org/10.1186/s13750-016-0059-6>
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.
<https://doi.org/10.1177/0956797611430953>

- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. <https://doi.org/10.1037/a0028347>
- Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *The American Psychologist*, *58*(9), 697–720. <https://doi.org/10.1037/0003-066X.58.9.697>
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. <https://doi.org/10.1080/01621459.1995.10476572>
- Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, *32*(1), 1–24. <https://doi.org/10.1006/jmla.1993.1001>
- Keren, G., & Teigen, K. H. (2004). Yet another look at the heuristics and biases approach. In D. J. Koehler & N. Harvey (eds.), *Blackwell handbook of judgment and decision making* (pp. 89–109). Malden, USA: Blackwell Publishing Ltd. <https://doi.org/10.1002/9780470752937>
- Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., ... Bahník, Š. (2018). Many Labs 2: investigating variation in replicability across

- samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443–490. <https://doi.org/10.1177/2515245918810225>
- Koriat, A., & Adiv, S. (2012). Confidence in one's social beliefs: implications for belief justification. *Consciousness and Cognition*, 21(4), 1599–1616. <https://doi.org/10.1016/j.concog.2012.08.008>
- Kumkale, G. T., & Albarracín, D. (2004). The sleeper effect in persuasion: a meta-analytic review. *Psychological Bulletin*, 130(1), 143–172. <https://doi.org/10.1037/0033-2909.130.1.143>
- Kurinec, C. A., & Weaver, C. A. (2018). Do memory-focused jury instructions moderate the influence of eyewitness word choice? *Applied Psychology in Criminal Justice*, 14(1), 55–69.
- Kvarven, A., Strømmland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, 4(4), 423–434. <https://doi.org/10.1038/s41562-019-0787-z>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 863. <https://doi.org/10.3389/fpsyg.2013.00863>
- Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: six practical recommendations. *BMC psychology*, 4(1), 24. <https://doi.org/10.1186/s40359-016-0126-3>
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: a tutorial. *Advances in Methods and Practices in*

Psychological Science, 1(2), 259–269.

<https://doi.org/10.1177/2515245918770963>

Larson, H. J. (2018). The biggest pandemic risk? Viral misinformation. *Nature*, 562(7727), 309. <https://doi.org/10.1038/d41586-018-07034-4>

Law, S., Hawkins, S., & Craik, F. (1998). Repetition-induced belief in the elderly: rehabilitating age-related memory deficits. *The Journal of Consumer Research*, 25(2), 91–107. <https://doi.org/10.1086/209529>

Lawrence, M. A. (2016). Ez: easy analysis and visualization of factorial experiments [Computer software]. Retrieved from <https://CRAN.R-project.org/package=eZ>

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094–1096. <https://doi.org/10.1126/science.aao2998>

Ledgerwood, A., Soderberg, C., & Sparks, J. (2017). Designing a study to maximize informational value. In *Toward a More Perfect Psychology: Improving Trust, Accuracy, and Transparency in Research* (pp. 33–58). Washington, DC: American Psychological Association.

Lenth, R. (2020). emmeans: estimated marginal means, aka least-squares means [Computer software]. Retrieved from <https://CRAN.R-project.org/package=emmeans>

Levine, T. R. (2014). Truth-Default Theory (TDT). *Journal of Language and Social Psychology*, 33(4), 378–392. <https://doi.org/10.1177/0261927X14535916>

- Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: understanding and coping with the “post-truth” era. *Journal of Applied Research in Memory and Cognition*, 6(4), 353–369.
<https://doi.org/10.1016/j.jarmac.2017.07.008>
- Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3), 106–131.
<https://doi.org/10.1177/1529100612451018>
- Lewandowsky, S., Stritzke, W. G. K., Oberauer, K., & Morales, M. (2005). Memory for fact, fiction, and misinformation: the Iraq War 2003. *Psychological Science*, 16(3), 190–195. <https://doi.org/10.1111/j.0956-7976.2005.00802.x>
- Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: what could possibly go wrong? *Journal of Experimental Social Psychology*, 79, 328–348. <https://doi.org/10.1016/j.jesp.2018.08.009>
- Loy, A., Hofmann, H., & Cook, D. (2017). Model choice and diagnostics for linear mixed-effects models using statistics on street corners. *Journal of Computational and Graphical Statistics*, 26(3), 478–492.
<https://doi.org/10.1080/10618600.2017.1330207>
- Mahmoudian, M. (2018). Varhandle: functions for robust variable handling [Computer software]. Retrieved from <https://CRAN.R-project.org/package=varhandle>
- Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: how often do they really occur? *Perspectives on Psychological Science*, 7(6), 537–542. <https://doi.org/10.1177/1745691612460688>

- Marsh, E. J., Cantor, A. D., & Brashier, N. M. (2016). Believing that humans swallow spiders in their sleep. In *Psychology of Learning and Motivation* (Vol. 64, pp. 93–132). Elsevier. <https://doi.org/10.1016/bs.plm.2015.09.003>
- Marsh, E. J., & Umanath, S. (2013). Knowledge neglect: failures to notice contradictions with stored knowledge. In D. N. Rapp & J. Braasch (Eds.), *Processing inaccurate information: Theoretical and Applied Perspectives from Cognitive Science and the Educational Sciences*. MIT Press.
- McGlone, M. S., & Tofiqbakhsh, J. (2000). Birds of a feather flock conjointly (?): rhyme as reason in aphorisms. *Psychological Science, 11*(5), 424–428. <https://doi.org/10.1111/1467-9280.00282>
- McGrath, J. E. (1981). Dilemmatics: the study of research choices and dilemmas. *American Behavioral Scientist, 179–210*. <https://doi.org/10.1177/000276428102500205>
- Menegatti, M., & Rubini, M. (2013). Convincing similar and dissimilar others: the power of language abstraction in political communication. *Personality and Social Psychology Bulletin, 39*(5), 596–607. <https://doi.org/10.1177/0146167213479404>
- Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: the use of cognitive heuristics. *Journal of Pragmatics, 59*, 210–220. <https://doi.org/10.1016/j.pragma.2013.07.012>
- Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1), 1. <https://doi.org/10.1186/2046-4053-4-1>

- Morey, R. D., & Rouder, J. N. (2018). BayesFactor: computation of bayes factors for common designs [Computer software]. Retrieved from <https://CRAN.R-project.org/package=BayesFactor>
- Mummolo, J., & Peterson, E. (2018). Demand effects in survey experiments: an empirical assessment. *The American Political Science Review*, *113*(2), 517-529. <https://doi.org/10.1017/S0003055418000837>
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., ... Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021. <https://doi.org/10.1038/s41562-016-0021>
- Nadarevic, L., & Aßfalg, A. (2017). Unveiling the truth: warnings reduce the repetition-based truth effect. *Psychological Research*, *81*(4), 814–826. <https://doi.org/10.1007/s00426-016-0777-y>
- Nadarevic, L., & Erdfelder, E. (2013). Spinoza's error: memory for truth and falsity. *Memory & Cognition*, *41*(2), 176–186. <https://doi.org/10.3758/s13421-012-0251-z>
- Nadarevic, L., & Erdfelder, E. (2014). Initial judgment task and delay of the final validity-rating task moderate the truth effect. *Consciousness and Cognition*, *23*, 74–84. <https://doi.org/10.1016/j.concog.2013.12.002>
- Nadarevic, L., & Erdfelder, E. (2019). [Supplementary material]. *Open Science Framework*. <https://doi.org/10.17605/osf.io/eut35>

- Nadarevic, L., Plier, S., Thielmann, I., & Darancó, S. (2018). Foreign language reduces the longevity of the repetition-based truth effect. *Acta Psychologica, 191*, 149–159. <https://doi.org/10.1016/j.actpsy.2018.08.019>
- Nadarevic, L., Reber, R., Helmecke, A. J., & Köse, D. (2020). Perceived truth of statements and simulated social media postings: an experimental investigation of source credibility, repeated exposure, and presentation format. *Cognitive Research: Principles and Implications, 5*(1), 56. <https://doi.org/10.1186/s41235-020-00251-4>
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511–534. <https://doi.org/10.1146/annurev-psych-122216-011836>
- Newman, E. J., Garry, M., Bernstein, D. M., Kantner, J., & Lindsay, D. S. (2012). Nonprobative photographs (or words) inflate truthiness. *Psychonomic Bulletin & Review, 19*(5), 969–974. <https://doi.org/10.3758/s13423-012-0292-0>
- Newman, E. J., Garry, M., Unkelbach, C., Bernstein, D. M., Lindsay, D. S., & Nash, R. A. (2015). Truthiness and falsiness of trivia claims depend on judgmental contexts. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 41*(5), 1337–1348. <https://doi.org/10.1037/xlm0000099>
- Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. <https://doi.org/10.1037/1089-2680.2.2.175>

- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. <https://doi.org/10.1126/science.aab2374>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, 18(3), e3000691. <https://doi.org/10.1371/journal.pbio.3000691>
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., ... Vazire, S. (2021). Replicability, robustness, and reproducibility in psychological science. PsyArXiv. <https://doi.org/10.31234/osf.io/ksfvq>
- Nosek, B. A., & Lakens, D. (2014). Registered Reports. *Social psychology*, 45(3), 137–141. <https://doi.org/10.1027/1864-9335/a000192>
- Obels, P., Lakens, D., Coles, N. A., & Gottfried, J. (2019). Analysis of open data and computational reproducibility in registered reports in psychology. PsyArXiv. <https://doi.org/10.31234/osf.io/fk8vh>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. <https://doi.org/10.1126/science.aac4716>
- Oppenheimer, D. M. (2004). Spontaneous discounting of availability in frequency judgment tasks. *Psychological Science*, 15(2), 100–105. <https://doi.org/10.1111/j.0963-7214.2004.01502005.x>

- Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, 12(6), 237–241. <https://doi.org/10.1016/j.tics.2008.02.014>
- Orne, M. T. (1962). On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. *American Psychologist*, 17(11), 776–783. <https://doi.org/10.1037/h0043424>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ (Clinical Research Ed.)*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5), 411-419. <https://ssrn.com/abstract=1626226>
- Paschal, O. (2018, August 3). Trump’s tweets and the creation of ‘illusory truth’. Retrieved December 11, 2018, from <https://www.theatlantic.com/politics/archive/2018/08/how-trumps-witch-hunt-tweets-create-an-illusory-truth/566693/>
- Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, 7(6), 531–536. <https://doi.org/10.1177/1745691612463401>
- Paul, C., & Matthew, M. (2016). *The Russian “firehose of falsehood” propaganda model* (pp. 2–7). Rand Corporation. <https://doi.org/10.7249/pe198>
- Peer, E., Samat, S., Brandimarte, L., & Acquisti, A. (2017). Beyond the Turk: an empirical comparison of alternative platforms for crowdsourcing online

- behavioral research. *Journal of Experimental Social Psychology*, (70), 153–163. <https://doi.org/10.2139/ssrn.2594183>
- Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12), 1865–1880. <https://doi.org/10.1037/xge0000465>
- Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*. 25(5), 388-402 <https://doi.org/10.1016/j.tics.2021.02.007>
- Polage, D. C. (2012). Making up history: false memories of fake news stories. *Europe's Journal of Psychology*, 8(2), 245–250. <https://doi.org/10.5964/ejop.v8i2.456>
- R Core Team. (2018). R: A language and environment for statistical computing [Computer software]. Retrieved from <https://www.R-project.org/>
- R Core Team. (2019). R: A language and environment for statistical computing [Computer software]. Retrieved from <https://www.R-project.org/>
- Reber, R., Fazendeiro, T., & Winkielman, P. (2002). Processing fluency as the source of experiences at the fringe of consciousness: commentary on Mangan. *Psyche: An Interdisciplinary Journal of Research on Consciousness*, 8(10), 1–21.
- Reber, R., & Greifeneder, R. (2017). Processing fluency in education: how metacognitive feelings shape learning, belief formation, and affect. *Educational Psychologist*, 52(2), 84–103. <https://doi.org/10.1080/00461520.2016.1258173>

- Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, 8(3), 338–342.
<https://doi.org/10.1006/ccog.1999.0386>
- Reber, R., & Unkelbach, C. (2010). The epistemic status of processing fluency as source for judgments of truth. *Review of Philosophy and Psychology*, 1(4), 563–581. <https://doi.org/10.1007/s13164-010-0039-7>
- Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, 49(4), 243–256. <https://doi.org/10.1026/1618-3169.49.4.243>
- Renkewitz, F., & Keiner, M. (2018). How to detect publication bias in psychological research. *Zeitschrift für Psychologie*, 227(4), 261-279
<https://doi.org/10.1027/2151-2604/a000386>
- Resnick, B. (2017). The science behind why fake news is so hard to wipe out - Vox.
Retrieved October 24, 2017, from <https://www.vox.com/science-and-health/2017/10/5/16410912/illusory-truth-fake-news-las-vegas-google-facebook>
- Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., ... PRISMA-S Group. (2021). PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews*, 10(1), 39. <https://doi.org/10.1186/s13643-020-01542-z>
- Riege, A., & Reber, R. (2022). Availability. In R. F. Pohl (Eds.), *Cognitive illusions: intriguing phenomena in judgement, thinking and memory* (3rd ed.).
Routledge

- Rife, S. C., Nuijten, M. B., & Epskamp, S. (2016). statcheck: Extract statistics from articles and recompute p-values [web application]. Retrieved from <http://statcheck.io>
- Rodd, J. (2019). How to maintain data quality when you can't see your participants. *APS Observer*, 32(3). Retrieved from <https://www.psychologicalscience.org/observer/how-to-maintain-data-quality-when-you-cant-see-your-participants>
- Roggeveen, A. L., & Johar, G. V. (2002). Perceived source variability versus familiarity: testing competing explanations for the truth effect. *Journal of Consumer Psychology*, 12(2), 81–91. <https://doi.org/10.1207/153276602760078622>
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Scheel, A. M., Schijen, M., & Lakens, D. (2020). An excess of positive results: comparing the standard Psychology literature with Registered Reports. 4(2), 1–12 <https://doi.org/10.1177/25152459211007467>
- Schmid, J., & Fiedler, K. (1996). Language and implicit attributions in the Nuremberg trials analyzing prosecutors' and defense attorneys' closing speeches. *Human Communication Research*, 22(3), 371–398. <https://doi.org/10.1111/j.1468-2958.1996.tb00372.x>
- Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13(2), 90–100. <https://doi.org/10.1037/a0015108>

- Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija*, 43(4), 441–464.
<https://doi.org/10.2298/PSI1004441S>
- Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(1), 82–102.
<https://doi.org/10.1037/0278-7393.9.1.82>
- Schwartz, M. (1982). Repetition and rated truth value of statements. *American Journal of Psychology*, 95(3), 393–407. <https://doi.org/10.2307/1422132>
- Schwarz, N. (2012). Feelings-as-Information Theory. In *Handbook of theories of social psychology: volume 1* (pp. 289–308). SAGE Publications Ltd.
<https://doi.org/10.4135/9781446249215.n15>
- Schwarz, N. (2015). Metacognition. In M. Mikulincer, P. R. Shaver, E. Borgida, & J. A. Bargh (eds.), *APA handbook of personality and social psychology, Volume 1: Attitudes and social cognition*. (pp. 203–229). Washington: American Psychological Association. <https://doi.org/10.1037/14341-006>
- Schwarz, N., & Jalbert, M. (2020). When (fake) news feels true: Intuitions of truth and the acceptance and correction of misinformation. In R. Greifeneder, M. Jaffé, E. J. Newman, & N. Schwarz (eds.), *The psychology of fake news: Accepting, sharing, and correcting misinformation*. (pp. 73-89). London, UK: Routledge. <https://doi.org/10.4324/9780429295379-7>
- Schwarz, N., Jalbert, M., Noah, T., & Zhang, L. (2021). Metacognitive experiences as information: processing fluency in consumer judgment and decision

making. *Consumer Psychology Review*, 4(1), 4–25.

<https://doi.org/10.1002/arcp.1067>

Semin, G. R. (2000a). Agenda 2000? communication: language as an

implementational device for cognition. *European Journal of Social*

Psychology, 30(5), 595–612. <https://doi.org/10.1002/1099->

0992(200009/10)30:5<595::AID-EJSP23>3.0.CO;2-A

Semin, G. R. (2000b). Language as a cognitive and behavioral structuring resource:

question—answer exchanges. *European Review of Social Psychology*, 11(1),

75–104. <https://doi.org/10.1080/14792772043000004>

Semin, G. R., & Fiedler, K. (1988). The cognitive functions of linguistic categories

in describing persons: social cognition and language. *Journal of Personality*

and Social Psychology, 54(4), 558–568. <https://doi.org/10.1037//0022->

3514.54.4.558

Semin, G. R., & Fiedler, K. (1991). The linguistic category model, its bases,

applications and range. *European Review of Social Psychology*, 2(1), 1–30.

<https://doi.org/10.1080/14792779143000006>

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-*

experimental designs for generalized causal inference. (pp. 623). Boston:

Houghton Mifflin. *Social Service Review*, 76(3), 510–514.

<https://doi.org/10.1086/345281>

Shamseer, L., Moher, D., Clarke, M., Gherzi, D., Liberati, A., Petticrew, M., ...

PRISMA-P Group. (2015). Preferred reporting items for systematic review

and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation.

BMJ (Clinical Research Ed.), 350, g7647. <https://doi.org/10.1136/bmj.g7647>

- Shidlovski, D., Schul, Y., & Mayo, R. (2014). If I imagine it, then it happened: the implicit truth value of imaginary representations. *Cognition*, *133*(3), 517–529. <https://doi.org/10.1016/j.cognition.2014.08.005>
- Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology*, *70*, 747–770. <https://doi.org/10.1146/annurev-psych-010418-102803>
- Silva, R. R., Chrobot, N., Newman, E., Schwarz, N., & Topolinski, S. (2017). Make it short and easy: username complexity determines trustworthiness above and beyond objective reputation. *Frontiers in Psychology*, *8*, 2200. <https://doi.org/10.3389/fpsyg.2017.02200>
- Silva, R. R., Garcia-Marques, T., & Reber, R. (2017). The informative value of type of repetition: perceptual and conceptual fluency influences on judgments of truth. *Consciousness and Cognition*, *51*, 53–67. <https://doi.org/10.1016/j.concog.2017.02.016>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2160588>
- Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80. <https://doi.org/10.1177/1745691613514755>

- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): a proposed addition to all empirical papers. *Psychological Science, 12*(6), 1123–1128. <https://doi.org/10.1177/1745691617708630>
- Skurnik, I., Yoon, C., Park, D., & Schwarz, N. (2005). How warnings about false claims become recommendations. *The Journal of Consumer Research, 31*(4), 713–724. <https://doi.org/10.1086/426605>
- Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J. G., Singleton Thorn, F., Vazire, S., ... Nosek, B. A. (2020). Research quality of registered reports compared to the traditional publishing model. MetaArXiv. <https://doi.org/10.31222/osf.io/7x9vy>
- Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science, 10*(6), 886–899. <https://doi.org/10.1177/1745691615609918>
- Stanley, M. L., Yang, B. W., & Marsh, E. J. (2018). When the unlikely becomes likely: qualifying language does not influence later truth judgments. *Journal of Applied Research in Memory and Cognition, 8*(1), 118–129. <https://doi.org/10.1016/j.jarmac.2018.08.004>
- Sterling, Rosenbaum, & Weinkam. (1995). Publication decisions revisited: the Effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49*(1), 108-112 <https://doi.org/10.1080/00031305.1995.10476125>
- Stott, P. (2011). Don't sound like a liar at work. Retrieved from <https://www.newsday.com/news/new-york/don-t-sound-like-a-liar-at-work-1.3031372>

- Street, C. N. H. (2015). ALIED: Humans as adaptive lie detectors. *Journal of Applied Research in Memory and Cognition*, 4(4), 335–343.
<https://doi.org/10.1016/j.jarmac.2015.06.002>
- Street, C. N. H., & Kingstone, A. (2017). Aligning Spinoza with Descartes: an informed Cartesian account of the truth bias. *British Journal of Psychology*, 108(3), 453–466. <https://doi.org/10.1111/bjop.12210>
- Street, C. N. H., & Richardson, D. C. (2015). Descartes versus Spinoza: truth, uncertainty, and bias. *Social Cognition*, 33(3), 227–239.
<https://doi.org/10.1521/soco.2015.33.2.2>
- Stroebe, W. (2019). What can we learn from Many Labs replications? *Basic and Applied Social Psychology*, 41(2), 91–103.
<https://doi.org/10.1080/01973533.2019.1577736>
- StudySwap. (2018). OSF | StudySwap: A platform for interlab replication, collaboration, and research resource exchange. Retrieved February 28, 2018, from <https://osf.io/view/studyswap/>
- Sundar, A., Kardes, F. R., & Wright, S. A. (2015). The influence of repetitive health messages and sensitivity to fluency on the truth effect in advertising. *Journal of Advertising*, 44(4), 375–387.
<https://doi.org/10.1080/00913367.2015.1045154>
- Sundar, S. S., Knobloch-Westerwick, S., & Hastall, M. R. (2007). News cues: information scent and cognitive heuristics. *Journal of the American Society for Information Science and Technology*, 58(3), 366–378.
<https://doi.org/10.1002/asi.20511>

- Sungur, H., Hartmann, T., & van Koningsbruggen, G. M. (2016). Abstract mindsets increase believability of spatially distant online messages. *Frontiers in Psychology*, 7, 1056. <https://doi.org/10.3389/fpsyg.2016.01056>
- Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. H. (2017). Processing political misinformation: comprehending the Trump phenomenon. *Royal Society Open Science*, 4(3), 160802. <https://doi.org/10.1098/rsos.160802>
- Tell it like it is. (2020). *Nature Human Behaviour*, 4(1), 1. <https://doi.org/10.1038/s41562-020-0818-9>
- Topor, M., Pickering, J. S., Barbosa Mendes, A., Bishop, D. V. M., Büttner, F. C., Henderson, E. L., ... Westwood, S. J. (2020). An integrative framework for planning and conducting non-interventional, reproducible, and open systematic reviews (NIRO-SR). MetaArXiv. <https://doi.org/10.31222/osf.io/8gu5z>
- Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 207–232. [https://doi.org/10.1016/0010-0285\(73\)90033-9](https://doi.org/10.1016/0010-0285(73)90033-9)
- Unkelbach, C. (2007). Reversing the truth effect: learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 33(1), 219–230. <https://doi.org/10.1037/0278-7393.33.1.219>
- Unkelbach, C., & Greifeneder, R. (2018). Experiential fluency and declarative advice jointly inform judgments of truth. *Journal of Experimental Social Psychology*, 79, 78–86. <https://doi.org/10.1016/j.jesp.2018.06.010>

- Unkelbach, C., Koch, A., Silva, R. R., & Garcia-Marques, T. (2019). Truth by repetition: explanations and implications. *Current Directions in Psychological Science*, 28(3), 247–253. <https://doi.org/10.1177/0963721419827854>
- Unkelbach, C., & Rom, S. C. (2017). A referential theory of the repetition-induced truth effect. *Cognition*, 160, 110–126. <https://doi.org/10.1016/j.cognition.2016.12.016>
- Unkelbach, C., & Speckmann, F. (2021). Mere repetition increases belief in factually true CoViD-19-related information. *Journal of Applied Research in Memory and Cognition*. <https://doi.org/10.1016/j.jarmac.2021.02.001>
- Unkelbach, C., & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition*, 18(1), 22–38. <https://doi.org/10.1016/j.concog.2008.09.006>
- Uygun-Tunç, D., & Tunç, M. N. (2020). A falsificationist treatment of auxiliary hypotheses in social and behavioral sciences: systematic replications rramework. PsyArXiv. <https://doi.org/10.31234/osf.io/pdm7y>
- van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, 67, 2–12. <https://doi.org/10.1016/j.jesp.2016.03.004>
- Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology*, 3(1), 1. <https://doi.org/10.1525/collabra.74>

- Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, 13(4), 411–417. <https://doi.org/10.1177/1745691617751884>
- Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2020). Credibility beyond replicability: improving the four validities in psychological science. PsyArXiv. <https://doi.org/10.31234/osf.io/bu4d3>
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5), 779–804. <https://doi.org/10.3758/BF03194105>
- Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part I: theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, 25(1), 35–57. <https://doi.org/10.3758/s13423-017-1343-3>
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6), 632–638. <https://doi.org/10.1177/1745691612463078>
- Wanke, M., & Hansen, J. (2015). Relative processing fluency. *Current Directions in Psychological Science*, 24(3), 195–199. <https://doi.org/10.1177/0963721414561766>

- Weigold, A., Weigold, I. K., & Russell, E. J. (2013). Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods. *Psychological Methods, 18*(1), 53–70.
<https://doi.org/10.1037/a0031607>
- West, W. C., & Holcomb, P. J. (2000). Imaginal, semantic, and surface-level processing of concrete and abstract words: an electrophysiological investigation. *Journal of Cognitive Neuroscience, 12*(6), 1024–1037.
<https://doi.org/10.1162/08989290051137558>
- Whittlesea, B. W. A. (1993). Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(6), 1235–1253.
<https://doi.org/10.1037//0278-7393.19.6.1235>
- Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software, 40*(1). <https://doi.org/10.18637/jss.v040.i01>
- Wickham, H. (2017). Tidyverse: easily install and load the 'tidyverse'. [Computer software]. Retrieved from: <https://CRAN.R-project.org/package=tidyverse>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., ... Yutani, H. (2019). Welcome to the tidyverse [Computer software]. *The Journal of Open Source Software, 4*(43), 1686.
<https://doi.org/10.21105/joss.01686>
- Wolffe, T. A. M., Whaley, P., Halsall, C., Rooney, A. A., & Walker, V. R. (2019). Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management. *Environment International, 130*, 104871. <https://doi.org/10.1016/j.envint.2019.05.065>

Xie, Y. (2015). Dynamic documents with R and knitr [Computer software].

Retrieved from <https://yihui.name/knitr/>

Yarkoni, T. (2019). The generalizability crisis. PsyArXiv.

<https://doi.org/10.31234/osf.io/jqw35>

Zhu, H. (2018). KableExtra: Construct complex table with 'kable' and pipe syntax

[Computer software]. Retrieved from [https://CRAN.R-](https://CRAN.R-project.org/package=kableExtra)

[project.org/package=kableExtra](https://CRAN.R-project.org/package=kableExtra)

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, e120.

<https://doi.org/10.1017/S0140525X17001972>

Online Supplemental Materials

Chapter 2

- Paper <https://doi.org/10.1525/collabra.192>
- Project <https://osf.io/s2389/>
- Preregistration <https://osf.io/dtcg9>
- Materials <https://osf.io/thp6v/>
- Data <https://osf.io/dypax/>
- Analysis code <https://osf.io/gbv2h/>

Chapter 3

- Paper <https://doi.org/10.3758/s13423-021-01995-w>
- Project <https://osf.io/dm9yx/>
- Preregistration <https://osf.io/ar4hm>
- Search record <https://osf.io/xsnhm/>
- Coding scheme <https://osf.io/a9mfq/>
- Data & analysis code <https://osf.io/ebnm5/>
- Systematic map database <https://osf.io/37xma/>

Chapter 4

- Paper <http://doi.org/10.5334/joc.161>
- Project <https://osf.io/nvugt/>
- Power analysis <https://osf.io/3d7sf/>
- Preregistration <https://osf.io/9mncq>
- Materials <https://osf.io/24ygh/>
- Video of procedure <https://osf.io/zgkpc/>
- Data <https://osf.io/upqb9/>
- Analysis code <https://osf.io/vr35h/>

Appendix A: Chapter 3 Benchmark List

1. Begg, I., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, 121, 446-458.
2. Brown, A. S., & Nix, L. A. (1996). Turning lies into truths: Referential validation of falsehoods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1088-1100.
3. Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2009). Mix Me a list: Context moderates the truth effect and the mere exposure effect. *Journal of Experimental Social Psychology*, 45, 1117-1122.
4. Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, 14, 238-257.
5. Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144, 993-1002.
6. Garcia-Marques, T., Silva, R. R., & Mello, J. (2017). Asking simultaneously about truth and familiarity may disrupt truth effects. *Análise Psicológica*, 35, 61-71.
7. Hawkins, S. A., & Hoch, S. J. (1992). Low-involvement learning: Memory without evaluation. *Journal of Consumer Research*, 19, 212-225.
8. Hawkins, S. A., Hoch, S. J., & Meyers Levy, J. (2001). Low-involvement learning: Repetition and coherence in familiarity and belief. *Journal of Consumer Psychology*, 11, 1-11.
9. Law, S., Hawkins, S. A., & Craik, F. I. M. (1998). Repetition-induced belief in the elderly: Rehabilitating age-related memory deficits. *Journal of Consumer Research*, 25, 91-107.
10. Mitchell, J. P., Dodson, C. S., & Schacter, D. L. (2005). fMRI evidence for the role of recollection in suppressing misattribution errors: The illusory truth effect. *Journal of Cognitive Neuroscience*, 17, 800-810.

11. Mitchell, J. P., Sullivan, A. L., Schacter, D. L., & Budson, A. E. (2006). Misattribution errors in Alzheimer's disease: The illusory truth effect. *Neuropsychology*, 20, 185-192.
12. Mutter, S. A., Lindsey, S. E., & Pliske, R. M. (1995). Aging and credibility judgment. *Aging and Cognition*, 2, 89-107.
13. Nadarevic, L., & Erdfelder, E. (2014). Initial judgment task and delay of the final validity-rating task moderate the truth effect. *Consciousness and Cognition*, 23, 74-84.
14. Nadarevic, L., Plier, S., Thielmann, I., & Darancó, S. (2018). Foreign language reduces the longevity of the repetition-based truth effect. *Acta psychologica*, 191, 149-159.
15. Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147, 1865–1880.
16. Roggeveen, A. L., & Johar, G. V. (2002). Perceived source variability versus familiarity: Testing competing explanations for the truth effect. *Journal of Consumer Psychology*, 12, 81-91.
17. Scholl, S. G., Greifeneder, R., & Bless, H. (2014). When fluency signals truth: Prior successful reliance on fluency moderates the impact of fluency on truth judgments. *Journal of Behavioral Decision Making*, 27, 268-280.
18. Schwartz, M. (1982). Repetition and rated truth value of statements. *American Journal of Psychology*, 95, 393-407.
19. Silva, R. R., Garcia-Marques, T., & Mello, J. (2016). The differential effects of fluency due to repetition and fluency due to color contrast on judgments of truth. *Psychological Research*, 80, 821-837.
20. Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 219-230.

Appendix B: Chapter 3 Bibliographic Database and Grey Literature Searches

Bibliographic database searches were conducted with “apply equivalent subjects/map term to subject heading” de-selected.

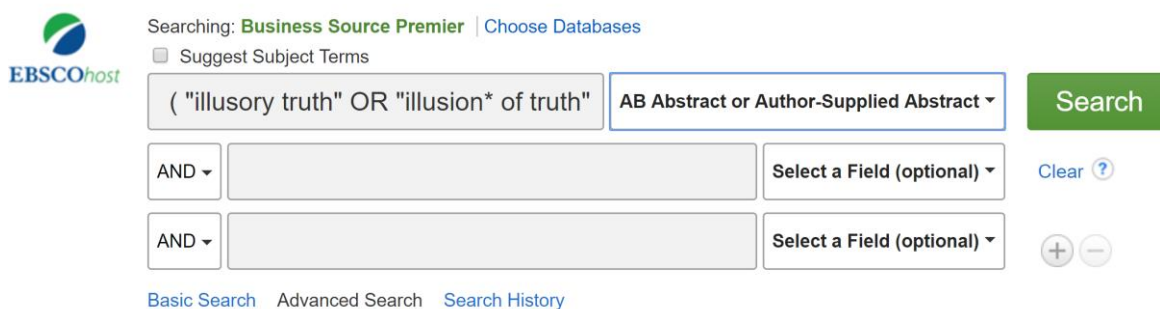
Business Source Premier

Field: “Abstract or author-supplied abstract”

Using “Advanced Search”

Search string: ("illusory truth" OR "illusion* of truth" OR "induced truth effect" OR "reiteration effect" OR "tainted truth effect" OR “repetition based truth effect” OR “repetition induced increases” OR repeat OR repeated OR repeating OR repetition OR "prior exposure") AND AB (true* OR truth OR "truth effect*" OR belief) AND AB (statement* OR items OR stimulus OR stimuli OR claim* OR judgment* OR judgement* OR rating* OR "subjective truth" OR "truth value" OR “judged validity” OR “validity ratings” OR "processing fluency" OR "fluency effect*" OR "perceptual fluency")

Search modes - Boolean/Phrase



Searching: **Business Source Premier** | [Choose Databases](#)

Suggest Subject Terms

("illusory truth" OR "illusion* of truth") **AB Abstract or Author-Supplied Abstract** ▼ **Search**

AND ▼ **Select a Field (optional)** ▼ [Clear](#) ?

AND ▼ **Select a Field (optional)** ▼ + -

[Basic Search](#) [Advanced Search](#) [Search History](#)

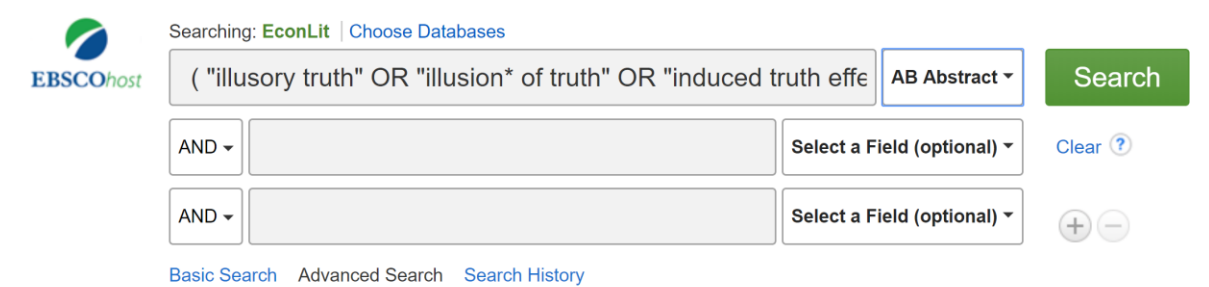
EconLit

Field: “Abstract”

Using “Advanced Search”

Search string: ("illusory truth" OR "illusion* of truth" OR "induced truth effect" OR "reiteration effect" OR "tainted truth effect" OR “repetition based truth effect” OR “repetition induced increases” OR repeat OR repeated OR repeating OR repetition OR "prior exposure") AND AB (true* OR truth OR "truth effect*" OR belief) AND AB (statement* OR items OR stimulus OR stimuli OR claim* OR judgment* OR judgement* OR rating* OR "subjective truth" OR "truth value" OR “judged validity” OR “validity ratings” OR "processing fluency" OR "fluency effect*" OR "perceptual fluency")

Search modes - Boolean/Phrase



Searching: [EconLit](#) | [Choose Databases](#)

EBSCOhost

("illusory truth" OR "illusion* of truth" OR "induced truth effe

AB Abstract ▾

Search

AND ▾

Select a Field (optional) ▾

Clear ?

AND ▾

Select a Field (optional) ▾

+ -

[Basic Search](#) [Advanced Search](#) [Search History](#)

ERIC

Field: “Abstract”

Using “Advanced Search”

Search string: ("illusory truth" OR "illusion* of truth" OR "induced truth effect" OR "reiteration effect" OR "tainted truth effect" OR “repetition based truth effect” OR “repetition induced increases” OR repeat OR repeated OR repeating OR repetition OR "prior exposure") AND AB (true* OR truth OR "truth effect*" OR belief) AND AB (statement* OR items OR stimulus OR stimuli OR claim* OR judgment* OR judgement* OR rating* OR "subjective truth" OR "truth value" OR “judged validity” OR “validity ratings” OR "processing fluency" OR "fluency effect*" OR "perceptual fluency")

Search modes - Boolean/Phrase

EBSCOhost Searching: ERIC | [Choose Databases](#)

("illusory truth" OR "illusion* of truth" OR "induced truth effe AB Abstract ▾ Search

AND ▾ Select a Field (optional) ▾ Clear ?

AND ▾ Select a Field (optional) ▾ + -

[Basic Search](#) [Advanced Search](#) [Search History](#)

PsycINFO

Field: "Abstracts"

Using "Advanced Search"

Search string: (("illusory truth" OR "illusion* of truth" OR "induced truth effect" OR "reiteration effect" OR "tainted truth effect" OR "repetition based truth effect" OR "repetition induced increases" OR repeat OR repeated OR repeating OR repetition OR "prior exposure") AND (true* OR truth OR "truth effect*" OR belief) AND (statement* OR items OR stimulus OR stimuli OR claim* OR judgment* OR judgement* OR rating* OR "subjective truth" OR "truth value" OR "judged validity" OR "validity ratings" OR "processing fluency" OR "fluency effect*" OR "perceptual fluency"))

[Basic Search](#) | [Find Citation](#) | [Search Tools](#) | [Search Fields](#) | **[Advanced Search](#)** | [Multi-Field Search](#)

1 Resource selected | [Hide](#) | [Change](#)

PsycINFO 1806 to October Week 1 2019

Enter keyword or phrase
(* or \$ for truncation)

Keyword Author Title Journal

(("illusory truth" OR "illusion* of truth" OR "induced truth effect" OR "reiteration effect" OR "tainted truth effect" OR "repetition based truth effect" OR "repetition induced increases" OR repeat OR repeated OR repeating OR repetition OR "prior exposure") AND (true* OR truth OR "truth effect*" OR belief) AND (statement* OR items OR stimulus OR stimuli OR claim* OR judgment* OR judgement* OR rating* OR "subjective truth" OR "truth value" OR "judged validity" OR "validity ratings" OR "processing fluency" OR "fluency effect*" OR "perceptual fluency"))

▼ Limits (close)

Include Multimedia Map Term to Subject Heading

- | | | |
|---|--|---|
| <input type="checkbox"/> Full Text | <input type="checkbox"/> PsycARTICLES Journals | <input type="checkbox"/> All Journals |
| <input type="checkbox"/> Latest Update | <input type="checkbox"/> Human | <input type="checkbox"/> English Language |
| <input checked="" type="checkbox"/> Abstracts | <input type="checkbox"/> Test DOI | <input type="checkbox"/> Open Access |
| <input type="checkbox"/> Impact Statement | | |

Publication Year - - - -

[Additional Limits](#) [Edit Limits](#)

PubMed

Field: “Title/Abstract”

Using “Advanced Search Builder”

Search string: "illusory truth" OR "illusion* of truth" OR "induced truth effect" OR "reiteration effect" OR "tainted truth effect" OR "repetition based truth effect" OR "repetition induced increases" OR repeat OR repeated OR repeating OR repetition OR "prior exposure"

AND true* OR truth OR "truth effect*" OR belief

AND statement* OR items OR stimulus OR stimuli OR claim* OR judgment* OR judgement* OR rating* OR "subjective truth" OR "truth value" OR "judged validity" OR "validity ratings" OR "processing fluency" OR "fluency effect*" OR "perceptual fluency"

```
((("illusory truth"[Title/Abstract] OR "illusion* of truth"[Title/Abstract] OR "induced truth effect"[Title/Abstract] OR "reiteration effect"[Title/Abstract] OR "tainted truth effect"[Title/Abstract] OR "repetition based truth effect"[Title/Abstract] OR "repetition induced increases"[Title/Abstract] OR repeat[Title/Abstract] OR repeated[Title/Abstract] OR repeating[Title/Abstract] OR repetition[Title/Abstract] OR "prior exposure"[Title/Abstract])) AND (true*[Title/Abstract] OR truth[Title/Abstract] OR "truth effect"[Title/Abstract] OR belief[Title/Abstract])) AND (statement*[Title/Abstract] OR items[Title/Abstract] OR stimulus[Title/Abstract] OR stimuli[Title/Abstract] OR claim*[Title/Abstract] OR judgment*[Title/Abstract] OR judgement*[Title/Abstract] OR rating*[Title/Abstract] OR "subjective truth"[Title/Abstract] OR "truth value"[Title/Abstract] OR "judged validity"[Title/Abstract] OR "validity ratings"[Title/Abstract] OR "processing fluency"[Title/Abstract] OR "fluency effect"[Title/Abstract] OR "perceptual fluency"[Title/Abstract])
```

[Edit](#)

[Clear](#)

Builder

Title/Abstract	"illusory truth" OR "illusion* of truth" OR "induced truth effect" OR "reiteration effect" OR "ta	Show index list
AND	Title/Abstract true* OR truth OR "truth effect*" OR belief	Show index list
AND	Title/Abstract statement* OR items OR stimulus OR stimuli OR claim* OR judgment* OR judgement* OR	Show index list
AND	All Fields	Show index list

[Search](#) or [Add to history](#)

Scopus

Field: “Article title, Abstract, Keywords”

Using “Advanced Search”

Search string: TITLE-ABS-KEY((((("illusory truth" OR "illusion* of truth" OR "induced truth effect" OR "reiteration effect" OR "tainted truth effect" OR “repetition based truth

effect" OR "repetition induced increases" OR repeat OR repeated OR repeating OR repetition OR "prior exposure") AND (true* OR truth OR "truth effect*" OR belief) AND (statement* OR items OR stimulus OR stimuli OR claim* OR judgment* OR judgement* OR rating* OR "subjective truth" OR "truth value" OR "judged validity" OR "validity ratings" OR "processing fluency" OR "fluency effect*" OR "perceptual fluency"))))

Documents Authors Affiliations Advanced

[Search tips ?](#)

Enter query string

TITLE-ABS-KEY((((("illusory truth" OR "illusion* of truth" OR "induced truth effect" OR "reiteration effect" OR "tainted truth effect" OR "repetition based truth effect" OR "repetition induced increases" OR repeat OR repeated OR repeating OR repetition OR "prior exposure") AND (true* OR truth OR "truth effect*" OR belief) AND (statement* OR items OR stimulus OR stimuli OR claim* OR judgment* OR judgement* OR rating* OR "subjective truth" OR "truth value" OR "judged validity" OR "validity ratings" OR "processing fluency" OR "fluency effect*" OR "perceptual fluency"))))

[Outline query](#) [Add Author name / Affiliation](#) [Clear form](#)

[Search Q](#)

Web of Science

Field: "Topic"

Using "Basic Search"

Search string: (((("illusory truth" OR "illusion* of truth" OR "induced truth effect" OR "reiteration effect" OR "tainted truth effect" OR "repetition based truth effect" OR "repetition induced increases" OR repeat OR repeated OR repeating OR repetition OR "prior exposure") AND (true* OR truth OR "truth effect*" OR belief) AND (statement* OR items OR stimulus OR stimuli OR claim* OR judgment* OR judgement* OR rating* OR "subjective truth" OR "truth value" OR "judged validity" OR "validity ratings" OR "processing fluency" OR "fluency effect*" OR "perceptual fluency"))))

Timespan: All years. **Databases:** WOS, BCI, BIOSIS, KJD, MEDLINE, RSCI, SCIELO.

Search language=Auto

Select a database

Basic Search Cited Reference Search Advanced Search

+ Add row | Reset

Timespan

Google Scholar (via Harzing's Publish or Perish)

Google Scholar searches will be combined and deduplicated before being added to the master spreadsheet.

Search 1:

Title words "illusory truth"

Google Scholar search	
Authors:	<input type="text"/>
Publication name:	<input type="text"/>
Title words:	"illusory truth"
Keywords:	<input type="text"/>

Search 2:

Title words "truth effect"

Google Scholar search	
Authors:	<input type="text"/>
Publication name:	<input type="text"/>
Title words:	"truth effect"
Keywords:	<input type="text"/>

Search 3:

Title words "truth judgement"

Google Scholar search	
Authors:	<input type="text"/>
Publication name:	<input type="text"/>
Title words:	"truth judgement"
Keywords:	<input type="text"/>

Search 4:

Title words "truth judgment"

Google Scholar search	
Authors:	<input type="text"/>
Publication name:	<input type="text"/>
Title words:	"truth judgment"
Keywords:	<input type="text"/>

Search 5:

Keywords "illusory truth"

Google Scholar search	
Authors:	<input type="text"/>
Publication name:	<input type="text"/>
Title words:	<input type="text"/>
Keywords:	<input type="text" value="illusory truth"/>

Grey Literature Databases: OpenGrey, PsyArXiv, Curate Science, PsychFileDrawer,

DART-Europe, EthOS, ProQuest Dissertation & Theses Global, Thesis Commons

Search 1: “illusory truth“

Search 2: “truth effect”

Search 3: “truth judgement”

Search 4: “truth judgment”

Appendix C: Chapter 3 References Included in Full-text Database

	Author	Year	Title
01	Arkes, H.R., Hackett, C., Boehm, L.	1989a	The generality of the relation between familiarity and judged validity
02	Arkes, H. R., Nash, J. G., & Joyner, C. A.	1989b	Solving a word puzzle makes subsequent statements containing the word seem more valid
03	Arkes, Hal R.; Boehm, Lawrence E; Xu, Gang	1991	Determinants of judged validity
04	Arkes, H. R., Nash, J. G., & Joyner, C. A.	1993	Replication of solving a word puzzle makes subsequent statements containing the word seem more valid
05	Bacon, F.T.	1979	Credibility of repeated statements: Memory for trivia
06	Bechkoff, J. R.	2008	Proprioception and the truth effect: A case in favor of the cartesian model of information processing
07	Begg, I., Armour, V., & Kerr, T.	1985	On believing what we remember
08	Begg, I., & Armour, V.	1991	Repetition and the ring of truth: Biasing comments
09	Begg, I., Anas, A., & Farinacci, S.	1992	Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth
10	Béna, J., Carreras, O., Terrier, P.	2019a	Attention division and the truth effect: A case of moderation by source credibility manipulation
11	Béna, J., Carreras, O., Terrier, P.	2019b	Delay between exposure and truth judgement decreases the truth effect in a one judgement procedure
12	Béna, J., Carreras, O., Terrier, P.	2020	Does delay between exposure and truth judgement decrease the truth effect through a recollection impairment? A Remember/Know study
13	Boehm, L. E.	1994	The validity effect: A search for mediating variables
14	Brashier, N. M., Umanath, S., Cabeza, R., & Marsh, E. J.	2017	Competing cues: Older adults rely on knowledge in the face of fluency

15	Brashier, N. M., Eliseev, E. D., & Marsh, E. J.	2020	An initial accuracy focus prevents illusory truth
16	Brown, A. S.; Nix, L. A.	1996	Turning lies into truths: Referential validation of falsehoods
17	Calio, F.	2019	Untersuchungen zur zeitlichen stabilität und zur vermeidbarkeit der wahrheitsillusion [Investigations into the temporal stability and the avoidability of the illusion of truth]
18	Chang, Y.	2019	Is the plausibility account of the illusion of truth effect plausible?
19	Corneille, O., Mierop, A., & Unkelbach, C.	2020	Repetition increases both the perceived truth and fakeness of information: An ecological account
20	De keersmaecker, J.; Dunning, D.; Pennycook, G.; Rand, D.G.; Sanchez, C.; Unkelbach, C.; Roets, A.	2020	Investigating the robustness of the illusory truth effect across individual differences in cognitive ability, need for cognitive closure, and cognitive style.
21	Dechêne, A., Stahl, C., Hansen, J., & Wänke, M.	2009	Mix me a list: Context moderates the truth effect and the mere-exposure effect
22	DiFonzo, N., Beckstead, J. W., Stupak, N., & Walders, K.	2016	Validity judgments of rumors heard multiple times: The shape of the truth effect
23	Doland, C. A.	1999	Repeating is believing: an investigation of the illusory truth effect
24	Ecker, U., Lewandowsky, S., & Chadwick, M.	2020	Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect
25	Effron, D. A., & Raj, M.	2020	Misinformation and morality: encountering fake-news headlines makes them seem less unethical to publish and share
26	Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J.	2015	Knowledge does not protect against illusory truth
27	Fazio, L., & Sherry, C.	2019a	The effect of repetition on truth judgments across development
28	Fazio, Lisa K; Rand, David G; Pennycook, Gordon	2019b	Repetition increases perceived truth equally for plausible and implausible statements
29	Fazio, L. K.,	2020a	Repetition increases perceived truth even for known falsehoods
30	Fazio, L. K.	2020b	Preventing the illusory truth effect: When repetition does not increase perceived truth

31	Frances, C.; Costa, A.; Baus, C.	2018	On the effects of regional accents on memory and credibility
32	Garcia-Marques, T., Silva, R. R., Reber, R., & Unkelbach, C.	2015	Hearing a statement now and believing the opposite later
33	Garcia-Marques, T., Silva, R. R., & Mello, J.	2016a	Judging the truth-value of a statement in and out of a deep processing context
34	Garcia-Marques, T., Prada, M., & Mackie, D. M.	2016b	Familiarity increases subjective positive affect even in non-affective and non-evaluative contexts
35	Garcia-Marques, T.; Silva, R.R.; Mello, J.	2017	Asking simultaneously about truth and familiarity may disrupt truth effects
36	Garcia-Marques, T., Silva, R. R., Mello, J., & Hansen, J.	2019	Relative to what? Dynamic updating of fluency standards and between-participants illusions of truth
37	Gigerenzer, G.	1984	External validity of laboratory experiments: The frequency-validity relationship
38	Hasher, L., Goldstein, D., & Toppino, T.	1977	Frequency and the confidence of referential validity
39	Hawkins, S. A., & Hoch, S. J.	1992	Low-involvement learning: Memory without evaluation
40	Hawkins, S. A., Hoch, S. J., & Meyers-Levy, J.	2001	Low-involvement learning: Repetition and coherence in familiarity and belief
41	Hernández Vera, A. V.	2020	El efecto de la repetición sobre la percepción de veracidad de frases: un estudio sobre la ilusión de verdad.
42	Jackson, D. R.	2018	Ethics in fake news: combatting the illusory truth effect with corrections
43	Jalbert M., Newman E., & Schwarz N.	2016	Trivia claim truth effect
44	Jalbert M., Newman E., & Schwarz N.	under review	Only half of what I'll tell you is true: How experimental procedures lead to an underestimation of the truth effect
45	Kim, C.	2002	The role of individual differences in general skepticism in the illusory truth effect
46	Ladowsky-Brooks, R. L.	2010	The truth effect in relation to neuropsychological functioning in traumatic brain injury
47	Law, S., & Hawkins, S. A.	1997	Advertising repetition and consumer beliefs: The role of source memory

48	Law, S.	1998a	Do we believe what we remember or, do we remember what we believe?
49	Law, S.	1998b	Investigating the truth effect in young and elderly consumers: The role of recognition and source memory
50	Law, S., Hawkins, S. A., & Craik, F. I.	1998c	Repetition-induced belief in the elderly: Rehabilitating age-related memory deficits
51	Lindsey, S.	1994	Aging and the truth effect in validity judgment
52	Mitchell, J. P.	2003	Asymmetries in the processing of true and false information
53	Mitchell, J. P., Dodson, C. S., & Schacter, D. L.	2005	fMRI evidence for the role of recollection in suppressing misattribution errors: The illusory truth effect
54	Mitchell, J. P., Sullivan, A. L., Schacter, D. L., & Budson, A. E.	2006	Misattribution errors in Alzheimer's disease: The illusory truth effect
55	Moritz, S., Köther, U., Woodward, T. S., Veckenstedt, R., Dechêne, A., & Stahl, C.	2012	Repetition is good? An internet trial on the illusory truth effect in schizophrenia and nonclinical participants
56	Murray, S., Stanley, M., McPhetres, J., Pennycook, G., & Seli, P.	2020	"I've said it before and I will say it again": Repeating statements made by Donald Trump increases perceived truthfulness for individuals across the political spectrum
57	Mutter, S. A., Lindsey, S. E., & Pliske, R. M.	1995	Aging and credibility judgment
58	Nadarevic, L.	2007	A failed replication of the truth effect
59	Nadarevic, L. & Rinnewitz, L.	2011	Judgment mode instructions do not moderate the truth effect
60	Nadarevic, L.; Meckler, D.; Schmidt, A.	2012	Are there interindividual differences of the truth effect? An investigation of different personality traits
61	Nadarevic, L.; Erdfelder, E.	2014	Initial judgment task and delay of the final validity-rating task moderate the truth effect
62	Nadarevic, L., & Aßfalg, A.	2017	Unveiling the truth: warnings reduce the repetition-based truth effect
63	Nadarevic, L., Plier, S., Thielmann, I., & Darancó, S.	2018	Foreign language reduces the longevity of the repetition-based truth effect
64	Newman, E. J., Jalbert, M. C., Schwarz, N., & Ly, D. P.	2020	Truthiness, the illusory truth effect, and the role of need for cognition

65	Oğuz Taşbaş, E. H., Unkelbach, C.,	2020b	Repetition effect and decision making
66	Pennycook, G., & Rand, D. G.	2017	The illusory truth effect for fake news is similar regardless of format
67	Pennycook, G., Cannon, T. D., & Rand, D. G.	2018	Prior exposure increases perceived accuracy of fake news
68	Polage, D.C.	2012	Making up history: False memories of fake news stories
69	Reyes de Luna, B.	2018	La formación de nuestras creencias: Efecto ilusorio de la verdad [The Formation of our beliefs: The illusory truth effect]
70	Roggeveen, A. L., & Johar, G. V.	2002	Perceived source variability versus familiarity: Testing competing explanations for the truth effect
71	Schwartz, Marian	1982	Repetition and rated truth value of statements
72	Silva, R. R., Garcia-Marques, T., & Mello, J.	2016	The differential effects of fluency due to repetition and fluency due to color contrast on judgments of truth
73	Silva, R. R., Garcia-Marques, T., & Reber, R.	2017	The informative value of type of repetition: Perceptual and conceptual fluency influences on judgments of truth
74	Sim, R.	2010	Memory mistakes and aging: How susceptibility to false recognition and the illusory truth effect changes across the lifespan
75	Skurnik, I. W.	1998	Metacognition and the illusion of truth
76	Skurnik, I., Yoon, C., Park, D. C., & Schwarz, N.	2005	How warnings about false claims become recommendations
77	Srull, T. K.	1983	The role of prior knowledge in the acquisition, retention, and use of new information
78	Stanley, M. L., Yang, B. W., & Marsh, E. J.	2019	When the unlikely becomes likely: Qualifying language does not influence later truth judgments
79	Sundar, A.; Kardes, F.R.; Wright, S.A.	2015	The influence of repetitive health messages and sensitivity to fluency on the truth effect in advertising
80	Toppino, T. C., Robertshaw, W., Hasher, L., & Goldstein, D.	1977	Frequency of occurrence and judgments of truth and falsity

81	Toppino, T.C.; Ann Brochin, H.	1989	Learning from tests: The case of true-false examinations
82	Toppino, T.C.; Luipersbeck, S.M.	1993	Generality of the negative suggestion effect in objective tests
83	Ulenaers, W., Pieters, R., & Warlop, L.	2000	Felt expertise and the truth effect
84	Unkelbach, C.	2007	Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth
85	Unkelbach, C., & Stahl, C.	2009	A multinomial modeling approach to dissociate different components of the truth effect
86	Unkelbach, C., Bayer, M., Alves, H., Koch, A., & Stahl, C.	2011	Fluency and positivity as possible causes of the truth effect
87	Unkelbach, C., & Rom, S. C.	2017	A referential theory of the repetition-induced truth effect
88	Unkelbach, C.; Greifeneder, R.	2018	Experiential fluency and declarative advice jointly inform judgments of truth
89	Unkelbach, C., & Speckmann, F.	2020a	Acting on illusory truth despite knowing better
90	Unkelbach	2020b	Knowledge does partially protect against illusory truth: The case of information related to the Corona crisis
91	Vicari, S. M.	2016	Overcoming the illusory truth effect: The influence of contextual details on memory monitoring
92	Vogel, Tobias; Silva, Rita R; Thomas, Aurelia; Wanke, Michaela	2020	Truth is in the mind, but beauty is in the eye: Fluency effects are moderated by a match between fluency source and judgment dimension
93	Wang, W. C., Brashier, N. M., Wing, E. A., Marsh, E. J., & Cabeza, R.	2016	On known unknowns: Fluency and the neural mechanisms of illusory truth

Appendix D: Chapter 3 Summary of Statcheck Issues

N errors	N papers	% papers	N decision errors	N papers	% papers
0	31	33.3	0	53	57
1	12	12.9	1	2	2.2
2	8	8.66	3	2	2.2
3	2	2.2			
4	1	1.1			
5	2	2.2			
9	1	1.1			
Unable to read	36	38.7			

Note. Statcheck recomputes p-values and compares them to those reported in the text. Inconsistent p-values are recorded as an “error”. If the reported result is significant and the recomputed result is not, or vice versa, the result is recorded as a “decision error”.

Appendix E: Chapter 4 Supplemental Analyses

Note. The analyses reported in the supplement are exploratory: We did not test our hypotheses with these analyses.

Supplement 1: Reports an ANOVA comparable to the analysis typically used in the literature. Note this is for illustrative purposes only and we do not draw any conclusions from this analysis.

Supplement 2: Reports the results of the funnel debriefing questions.

Supplement 3: Reports the results of the optional future research questions.

Supplement 4: Reports the results of an exploratory analysis investigating whether the illusory truth effect varies with statement topic.

Supplement 5: Reports the results of an exploratory analysis investigating whether the illusory truth effect varies with statement length.

E.1 Analyses using ANOVA

The illusory truth has typically been analysed using ANOVAs. The ANOVA analysis treats stimuli as fixed effects, and ordinal data violates the assumption of a continuous dependent variable with variance proportional to the mean, so we provide this ANOVA only for illustrative purposes, and only draw our conclusions from the results of the cumulative link mixed model analysis in the manuscript.

We conducted a 2 (repetition: new vs. repeated) x 4 (retention interval: immediately vs. 1 day vs. 1 week vs. 1 month) repeated measures ANOVA using participants' mean truth ratings as the dependent variable. The analysis was performed in R 3.6.2 (R Core Team, 2019) using the ezANOVA function from the R package ez version 4.4-0 (Lawrence, 2016). There was a significant main effect of both repetition $F(1, 510) = 511.95, p < .001$, and interval (with Greenhouse-Geisser correction for sphericity) $F(2.77, 1413.03) = 55.70, p < .001$. The repetition-by-interval interaction was also significant (with Greenhouse-Geisser correction for sphericity) $F(2.53, 1289.53) = 77.14, p < .001$.

```

$ANOVA
      Effect DFn  DFd      F      p p<.05      ges
2      repetition    1  510 511.94852 5.286861e-79 * 0.09899875
3      interval     3 1530  55.69935 3.604501e-34 * 0.02498054
4 repetition:interval  3 1530  77.13552 1.827937e-46 * 0.02884505

$`Mauchly's Test for Sphericity`
      Effect      W      p p<.05
3      interval 0.8849437 4.301330e-12 *
4 repetition:interval 0.7720104 1.071216e-26 *

$`Sphericity Corrections`
      Effect      GGe      p[GG] p[GG]<.05      HFe      p[HF]
3      interval 0.9235465 9.113467e-32 * 0.9290976 6.097952e-32
4 repetition:interval 0.8428288 1.305286e-39 * 0.8473754 8.267529e-40
p[HF]<.05
3      *
4      *

```

E.2 Funnel debrief

At the end of the phase 4 study we used three funnel debrief questions to ascertain whether participants had guessed the purpose of the overall study. Anonymised responses to all questions are available at <https://osf.io/upqb9/>. We asked participants the open-ended question “Do you have an idea about what we were testing in this study? (If yes, please describe what you think we were testing.)”. If the response included reference to repetition and truth (or their synonyms), we coded the answer as “yes” (even though the main hypothesis relates to the effect over time). If the response did not include both elements, we coded the answer as “no”. If the response was ambiguous, we coded it as “maybe”. Using this criterion, 20 (3.9%) guessed that the study related to the illusory truth effect, 478 (94.3%) did not guess, and 9 (1.8%) were maybe.

Next we asked participants the multiple choice question “Did you notice that some statements were repeated during the study?”. Twenty five (4.9%) noticed that statements had been repeated in all phases, 322 (63.5%) noticed that statements had been repeated in some phases, and 160 (31.6%) did not notice the repetition. A caveat to these results is that it became apparent that the question was ambiguously worded; some participants interpreted it as asking whether they noticed that statements were repeated within a study phase, rather than repeated from phase 1.

Finally, we asked participants the open-ended question “Why do you think we repeated some statements?”. We coded the answers in the same way as the first open-ended question. Using that criterion, and now that participants had received information about repetition, 23 (4.5%) guessed that the study related to the illusory truth effect, 473 (93.3%) did not guess, and 11 (2.2%) were coded as maybe.

E.3 Participants' views on future research

After participants had been debriefed at the end of phase 4, we asked for their views about future research on the illusory truth effect. The two questions were optional. First, we asked participants the open-ended question “What questions do you think researchers in this area should focus on answering? (There are no wrong answers. This question is *optional*)”. Anonymised responses to this question (n = 160), and to the following question (n = 472), are available at <https://osf.io/upqb9/>. We redacted the response from one participant because it contained hate speech.

Second, we asked “Below are five potential areas for further research on the illusory truth effect. Please rank them in order of importance with 1 being the most important and 5 being the least important (This question is *optional*)”. Participants ranked the five options in the following order of importance: 1) *How could the effect could be used to overcome misinformation?*, 2) *Can people overcome or avoid the effect if they know about it (i.e., can they avoid judging repeated information as truer)?*, 3) *Why do people seem to equate repetition with truth?*, 4) *How does the effect work in different contexts e.g. marketing, political campaigns?*, 5) *What are the boundaries of the effect in terms of number of repetitions over which time period (i.e., do 10 repetitions produce a stronger effect than 5 repetitions)?*.

E.4 Association between statement topic and size of illusory truth effect

During the exposure phase participants categorised statements into one of six categories (see Table 1). Based on a participant comment, we were interested to see whether there was a relationship between statement topic and the illusory truth effect. For example, statements relating to science, nature and technology might be considered truer than those relating to arts and entertainment.

Table 1

Mean and SDs for the Illusory Truth Effect by Statement Category

category	mean	sd	N
art & entertainment	0.46	0.28	19
language	0.42	0.19	19
science, nature & technology	0.39	0.21	54
geography	0.35	0.18	30
sports	0.33	0.19	7
history & politics	0.29	0.16	41

Note that some statements could be assigned to two categories (e.g., “Lubaantun is a ruined Mayan city in Belize” could be assigned to geography, or history & politics). Out of 128 statements, 84 could not be uniquely assigned to a single category. These were removed from the subsequent analysis. Stimulus category explained only about 5.0% of the variance in the size of the illusory truth effect across stimuli. The effect of stimulus category on the size of the illusory truth effect was not significant, $F(5,80) = 1.89$, $p = .106$.


```

Call:
lm(formula = it_eff ~ category, data = stim_it2)

Residuals:
    Min       1Q   Median       3Q      Max
-0.47087 -0.15509 -0.00292  0.10391  0.60688

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.52835    0.05974   8.844 1.78e-13 ***
categorygeography -0.18488    0.07903  -2.339  0.0218 *
categoryhistory & politics -0.21958    0.07712  -2.847  0.0056 **
categorylanguage -0.10714    0.10347  -1.035  0.3036
categoryscience, nature & technology -0.11759    0.07103  -1.655  0.1018
categorysports -0.16291    0.11015  -1.479  0.1431
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2069 on 80 degrees of freedom
Multiple R-squared:  0.1056, Adjusted R-squared:  0.04965
F-statistic: 1.888 on 5 and 80 DF,  p-value: 0.1056

```

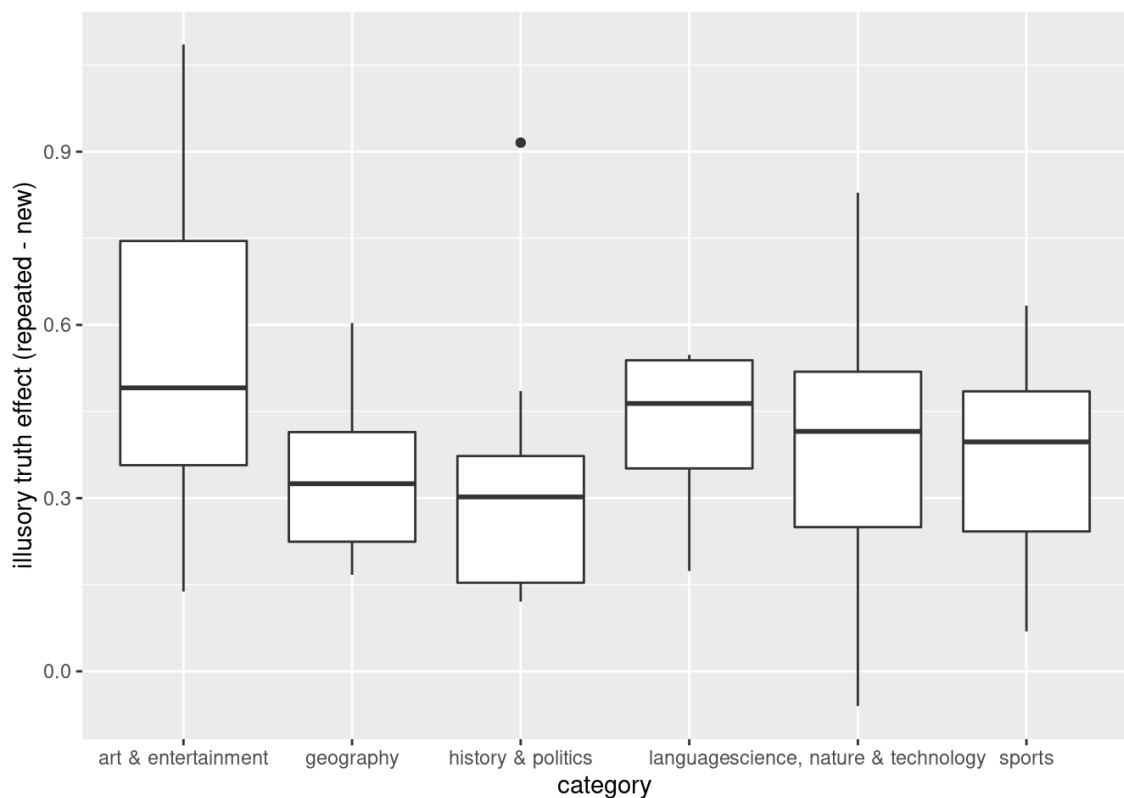


Figure 1. *Boxplot of the illusory truth effect by statement category.*

E.5 Association between statement length and size of illusory truth effect

We investigated whether there was a relationship between the illusory truth effect and statement length. Shorter statements are typically less complex and may have been easier to remember. Stimulus length explained only about 0.2% of the variance in the size of the illusory truth effect across stimuli. The effect of stimulus length on the size of the illusory truth effect was not significant, $F(1,126) = 0.77$, $p = .383$.

```
Call:
lm(formula = it_eff ~ n_words, data = stimlen2)

Residuals:
    Min       1Q   Median       3Q      Max
-0.42786 -0.16795 -0.02021  0.09447  0.71155

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.316459   0.072230   4.381 2.46e-05 ***
n_words      0.006411   0.007319   0.876  0.383
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.207 on 126 degrees of freedom
Multiple R-squared:  0.006053,    Adjusted R-squared:  -0.001836
F-statistic: 0.7673 on 1 and 126 DF,  p-value: 0.3827
```

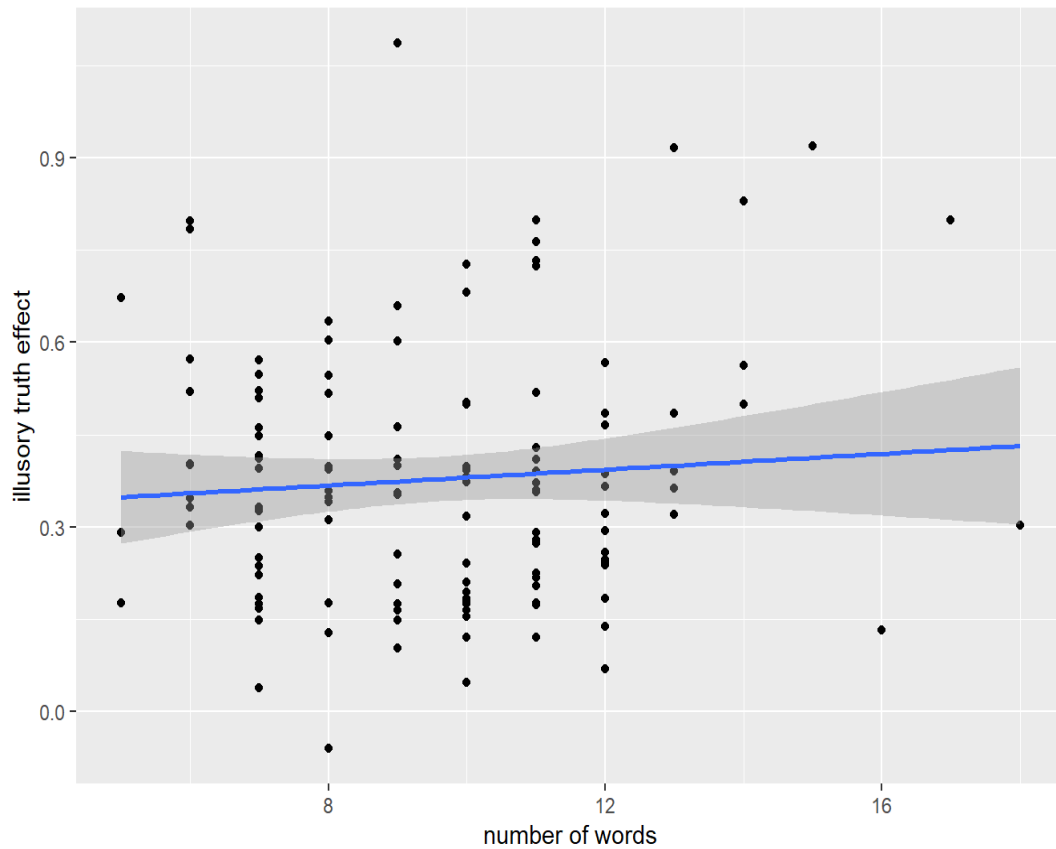


Figure 2. *Scatterplot of the illusory truth effect by statement length.*

References

Lawrence, M. A. (2016). ez: Easy Analysis and Visualization of Factorial Experiments, R package version 4.4-0, <https://CRAN.R-project.org/package=ez>.

R Core Team (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria. <https://www.R-project.org>.

Appendix F: Chapter 4 Amended Figure Showing Distribution of Participants' Age

The below plot visualising the distribution of participant ages has been adjusted and now uses one bin per year. The equivalent plot in Chapter 4 used the ggplot2 default bin settings that were not ideal for this plot.

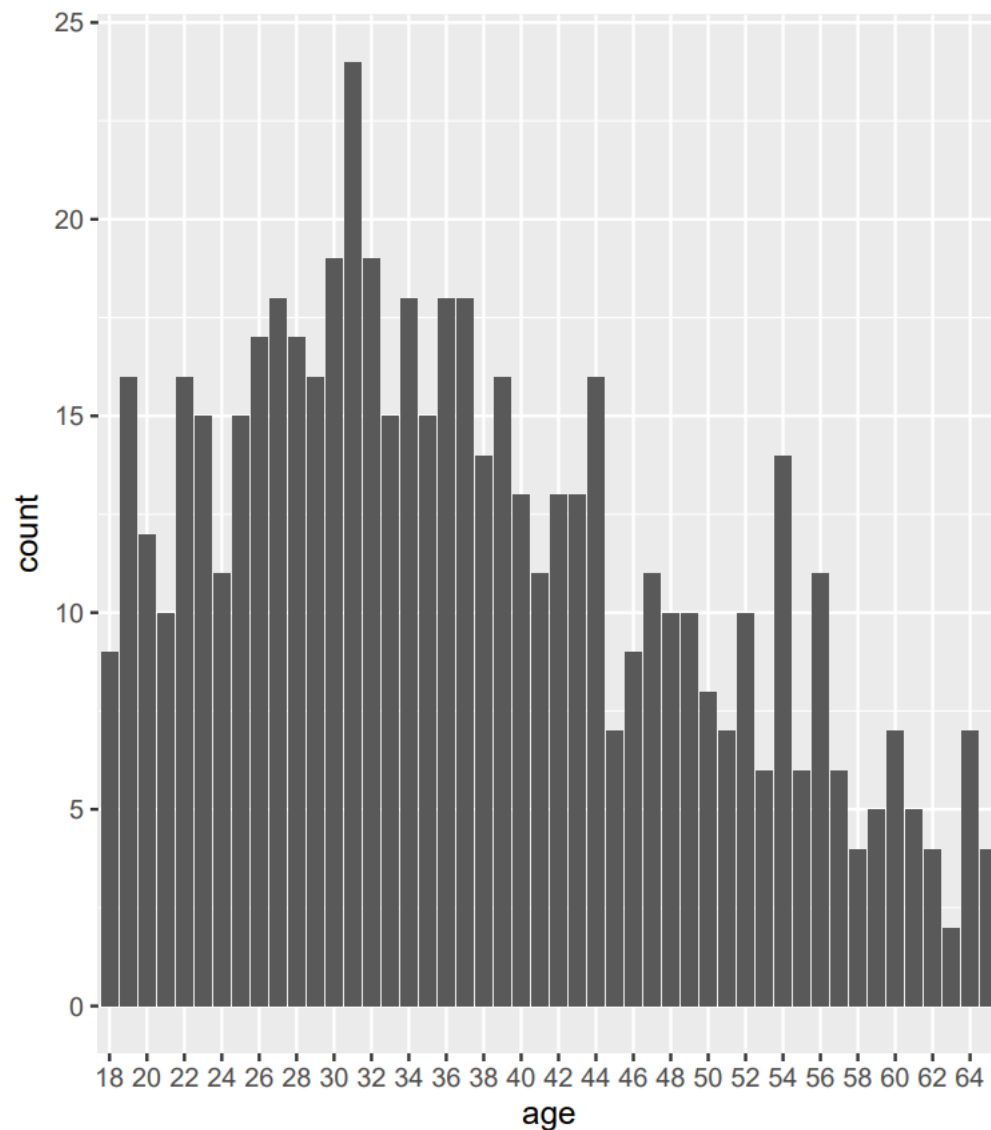


Figure 6. *Distribution of participants' age.*

Appendix G: Chapter 2 Effect Size Calculations and Explanations

In Chapter 2, two effect sizes are reported based on the results of Experiment 1 from Hansen & Wänke (2010).

Results from Hansen & Wänke (2010)

Abstract language: $M = 3.45$, $SD = 0.404$

Concrete language: $M = 3.57$, $SD = 0.448$

Mean difference score = 0.121, SD of the difference scores = 0.254

d_z : Standardized difference scores

Using the above values, $d_z =$

mean diff / SDdiff =

$0.121 / 0.254 = 0.48$

d_r : Residual standard deviation in the denominator

The BayesFactor packages uses d_r rather than d_z or d .

Using the above values, $d_r =$

(mean diff / SDdiff) / sqrt2 =

$(0.121 / 0.254) / \text{sqrt}2 = 0.34$

As the original Hansen and Wänke (2010) effect size was not precise ($d_z = 0.48$; 95% CI [0.19, 0.78]), I did not have a strong prior that the effect size would be precisely $d_r = 0.34$ (the equivalent of $d_z = 0.48$). I therefore used 0.34 as the scale parameter to determine the width of the alternative distribution, rather than testing against a location parameter of 0.34 (an exact spike alternative). Setting the scale parameter to $d_r = 0.34$ generates an alternative hypothesis that the effect is likely to be in a range with a median at 0.34 and equally likely to be above or below that (with some spread).

Appendix H: Ethics

From: [Redacted]
Date: 5 March 2018 at 09:54:14 GMT
To: "Henderson, Emma L" <[Redacted]>
Subject: **FREC 18 21 A Sense of Truth: Concreteness, Psychological Distance and the Illusory Truth Effect**

Dear Emma

FREC 18 21 A Sense of Truth: Concreteness, Psychological Distance and the Illusory Truth Effect – Outcome

Further to the submission of your project entitled '*A Sense of Truth: Concreteness, Psychological Distance and the Illusory Truth Effect*' the Faculty Research Ethics Committee have given approval for you to proceed.

Best wishes

Bereni

[Redacted]

Research Operations Manager

Kingston Business School

Kingston University

Notice of Approval: New Submission

March 20, 2018

Principal Investigator	Daniel Simons
Protocol Title	The effects of concrete and abstract wording on judgments of truth
Protocol Number	18672
Funding Source	Unfunded
Review Type	Exempt 2
Status	Active
Risk Determination	No more than minimal risk
Approval Date	03/20/2018

This letter authorizes the use of human subjects in the above protocol. The University of Illinois at Urbana-Champaign Institutional Review Board (IRB) has reviewed and approved the research study as described.

Exempt protocols are approved for a five year period from their original approval date, after which they will be closed and archived. Researchers may contact our office if the study will continue past five years.

The Principal Investigator of this study is responsible for:

- Conducting research in a manner consistent with the requirements of the University and federal regulations found at 45 CFR 46.
- Requesting approval from the IRB prior to implementing modifications.
- Notifying OPRS of any problems involving human subjects, including unanticipated events, participant complaints, or protocol deviations.
- Notifying OPRS of the completion of the study.

Office for the Protection of Research Subjects
University of Illinois at Urbana-Champaign

[Redacted]