UNIVERSITY OF CALIFORNIA,
IRVINE

Retrieval-Enhanced Suggestibility: Experimentation and Meta-Analysis

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Psychological Science

by

Brendon Jerome Butler

Dissertation Committee:
Distinguished Professor Elizabeth F. Loftus, Chair
Professor Linda J. Levine
Assistant Professor of Teaching Amy L. Dent

2020

DEDICATION

To

my parents, family, and friends

Make the right choice, no need for an apology…

Dennis David Coles

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express great appreciation and gratitude to my committee chair, Distinguished Professor Elizabeth F. Loftus, for her unwavering support throughout my graduate career. She is truly the best advisor a graduate student can have. I would also like to thank my committee members, Professor Linda J. Levine and Assistant Professor of Teaching Amy L. Dent, for their support and training throughout my dissertation process. I would also like to thank all the past and present members of Loftus Lab, Levine Lab, and Ditto Lab. Finally, I would like to thank Graduate Division for their support over the years, particularly Daniel Fabrega, Claudia Campos, and Mariela Menendez.

CURRICULUM VITAE

Brendon Jerome Butler

Research interests: misinformation, discrepancy detection, technocognition

Current Position

Doctoral Candidate (ABD) 2014-Present

Psychological Science
University of California, Irvine
Advisor: Dr. Elizabeth F. Loftus

Education

PhD **Psychological Science** Expected June 2020
Specialization: Quantitative Methods
University of California, Irvine

MA **Social Ecology** 2017
University of California, Irvine

BA **Psychology** 2012
University of California, Riverside

Fellowships, Honors, and Awards

National Science Foundation Graduate Research Fellowship 2016-Present
NSF-GRFP

Honorable Mention, Dissertation Fellowship 2019
Ford Foundation

Honorable Mention, Predoctoral Fellowship 2016
Ford Foundation

Graduate Mentoring Award
School of Social Ecology 2016
University of California, Irvine

Graduate Dean's Recruitment Fellowship 2015

University of California, Irvine

Eugene Cota-Robles Diversity Fellowship 2014
University of California, Irvine

Competitive Edge Summer Research Fellowship 2014
University of California, Irvine

Most Dedicated Member Award 2014
Psi Chi International Honor Society in Psychology
University of California, Riverside

Academic Achievement Scholarship 2008
Gamma Zeta Boule Foundation

Publications

Butler, B. J., & Loftus, E. F. (2018). Discrepancy detection in the retrieval-enhanced suggestibility paradigm. *Memory*, *26*(4), 483-492.

Grady, R. H., **Butler, B. J.**, & Loftus, E. F. (2016). What should happen after an officer-involved shooting? Memory concerns in police reporting procedures. *Journal of Applied Research in Memory and Cognition*, *5*(3), 246-251.

Presentations and Invited Lectures

Butler, B. J., Charles, S., & Adams, V. (2018, September). Behavioral Science Recruitment Panel. Invited panel speaker at the Southern California Forum for Diversity in Graduate Education Program Committee.

Butler, B. J. (2016, May). Retrieval-Enhanced Suggestibility. Research presentation given at the University of California, Irvine Psychology and Social Behavior Departmental Colloquium, Irvine, CA.

Butler, B. J., Drake, S., Kizer, J., & Reynaga, C. (2015, July). The Graduate Student Lifecycle. Guest speaker for the University of California, Irvine Summer Research Program, Irvine, CA.

Butler, B. J. (2014, August). Partisan Affiliation and Ripple Effects in Memory. Research proposal presentation given at the University of California, Irvine Graduate Division Summer Research Symposium, Irvine, CA.

Butler, B. J. & Mitchell, M. (2014, July). My Journey to Grad School. Guest speaker for the University of California, Irvine UC-HBCU Summer Research Program, Irvine, CA.

Poster Presentations

Spraggins, E. J. & **Butler, B. J.** (2017, May). Discrete Emotions and Memory. University of California, Irvine Undergraduate Research Opportunity Program Symposium, Irvine, CA.

Jhong, E. & **Butler, B. J.** (2016, May). Retrieval-Enhanced Suggestibility. University of California, Irvine Undergraduate Research Opportunity Program Symposium, Irvine, CA.

Teaching Experience

University of California, Irvine

Teaching Assistant, Industrial/Organizational Psychology	Summer 2018
Teaching Assistant, Psychology Fundamentals	Spring 2016
Teaching Assistant, Human Sexuality	Fall 2015
Teaching Assistant, Error and Bias	Winter 2015

Granite Hill Elementary School 2013-2014
Jurupa Valley, CA
6th Grade Teacher

Professional Training and Workshops

Mentoring Excellence Program
University of California, Irvine, 2016
Mentoring program that prepares graduates students for mentoring diverse groups of students in academia. Topics include: The lifecycle of the mentoring relationship**,** effective interpersonal communication, resilience & conflict resolution
and mentoring across differences.

Professional Affiliations

Association for Psychological Science 2015-Present

Academic, Professional, and Community Service

Reviewer, Journal of Memory and Language

Reviewer, Social Ecology Graduate Review Committee

Served as reviewer for 1st and 2nd year graduate students in the school who were applying to the National Science Foundation Graduate Research Fellowship program. University of California, Irvine, 2018

Panelist, Southern California Forum for Diversity in Graduate Education Program Committee
University of San Diego, 2018

Cohort Liaison, Department of Psychological Science Climate Committee
University of California, Irvine, 2018
Served as cohort for the department's internal climate committee. Met with the Graduate Advisor on a regular basis to address issues concerning grad student well-being.

Graduate Student Lead, UCI Summer Research Program
University of California, Irvine, 2018
Supported undergraduate students from diverse backgrounds who were at UCI conducting research over the summer.

Peer Mentor, Competitive Edge Summer Research Program
University of California, Irvine, 2018
Supported incoming graduate students from diverse backgrounds who plan to pursue a PhD in my department or School.

Peer Writing Reviewer, Competitive Edge Summer Research Program
University of California, Irvine, 2018
Reviewed incoming doctoral students' fellowship applications and provided them with in-depth feedback and support throughout the application process.

Campus Representative, Association for Psychological Science (APS)
University of California, Irvine, 2017-Present
Promoted psychological science amongst my peers, shared information with my department about various APS events and programs for student affiliates throughout the academic year, and networked with APSSC Board members and other APSSC Campus Representatives around the world.

Organizer, Stats 'n' Snacks
Psychology and Social Behavior, University of California, Irvine, 2017-2018
Organized *Stats 'n' Snacks*, a speaker series where knowledgeable statisticians gave lectures on statistics methods to the department.

Peer Mentor, Competitive Edge Summer Research Program
University of California, Irvine, 2017
Supported incoming graduate students from diverse backgrounds who plan to pursue a PhD in my department or School.

Peer Writing Reviewer, Competitive Edge Summer Research Program

University of California, Irvine, 2017
Reviewed incoming doctoral students' fellowship applications and provided them with in-depth feedback and support throughout the application process.

Peer Mentor, Competitive Edge Summer Research Program
University of California, Irvine, 2016
Supported incoming graduate students from diverse backgrounds who plan to pursue a PhD in my department or School.

Recruiter, Competitive Edge Summer Research Program
San Jose State University, 2016
Recruited and provided information for undergraduate students interested in graduate school.

Graduate Mentor, UCI DECADE-PLUS
University of California, Irvine, 2016-2017
Mentored students from first-generation or low-income (FGLI) households, who are the groups that are most at risk of being dismissed from the university to do academic underperformance.

Graduate Mentor, UC-HBCU Initiative
University of California, Irvine, 2015
Mentored students from Historically Black Colleges and Universities; The students were here through the UC-HBCU Initiative, a program that aims to improve the underrepresentation of Blacks in University of California graduate programs through intensive research experiences at a research university.

Peer Mentor, BLAACK
University of California, Riverside, 2010-2012
Through BLAACK (Brothers Leading African Americans Through Consciousness and Knowledge), I mentored freshman and sophomore students at UCR. Helped students with the adjustment to college, financial budgeting, dealing with imposter syndrome, and generally supported the well-being of mentees.

Computer Skills

Programming: R, Python, SQL

Applications: R, STATA, Python, SAS, SPSS

ABSTRACT OF THE DISSERTATION

Retrieval-Enhanced Suggestibility: Experimentation and Meta-Analysis

by

Brendon Jerome Butler

Doctor of Philosophy in Psychological Science

University of California, Irvine, 2020

Distinguished Professor Elizabeth F. Loftus, Chair

Retrieval-enhanced suggestibility (RES) refers to the finding that immediately recalling the details of a witnessed event can increase susceptibility to later misinformation. This finding is particularly surprising considering decades of research on the testing effect, which shows that retrieval practice enhances memory and protects learners against subsequent memory intrusions. Although many researchers have found RES effects in their experiments, many investigations into RES have yielded mixed or null results. Thus, retrieval-enhanced suggestibility is an intriguing finding deserving of empirical review through meta-analysis. The objectives of this dissertation are to: (a) test retrieval-enhanced suggestibility through experimentation in two new contexts; (b) identify the overall size of the retrieval-enhanced suggestibility effect; and (c) identify the methodological factors that moderate the size of the effect.

INTRODUCTION

A century ago, Arthur Irving Gates (1917) showed that immediate testing improves later recall of studied or learned material. Since then, numerous studies have shown this same *testing effect.* Researchers have examined the memorization of word lists (Brewer, Marsh, Meeks, Clark-Foos, & Hicks, 2010; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011; Tulving & Watkins, 1974), picture lists (Wheeler & Roediger, 1992), face-name patterns (Weinstein, McDermott, & Szpunar, 2011), and written narratives (Glover, 1989; Roediger & Karpicke, 2006). For all these different types of tested materials, the results remain consistent: testing enhances learning and memory for practiced material.

Testing is also known to help protect against *retroactive interference*, where newly learned information interferes with memories for previously learned material (e.g. Brewer et al., 2010; Chan & McDermott, 2007; Jang & Huber, 2008; Szpunar, McDermott, & Roediger, 2009; Weinstein et al., 2011). Some researchers believe testing also protects against *proactive interference* by improving source discrimination (Weinstein et al., 2011). According to this view, testing helps individuals isolate discrete sets of information (e.g. specific word lists) from one another, which improves speed and accuracy during retrieval. Alternatively, Pastötter et al. (2011) proposed that initial testing improves the encoding process of the learned material. Because of this enhanced encoding, an individual's memory is more resistant to potential interference from subsequent information.

Despite the extant literature showing that initial testing buffers against proactive interference, there is a growing line of research surrounding an effect known as *retrieval-enhanced suggestibility*, which in some ways can be thought of as a reverse testing effect (Butler

& Loftus, 2018). Retrieval-enhanced suggestibility, or RES, refers to the finding that immediately recalling the details of a witnessed event can increase an individual's susceptibility to later misinformation (Chan, Thomas, & Bulevich, 2009). Some researchers believe that the RES effect is due to initial test questions serving as cues that guide attention to the misinformation (e.g., Gordon, Thomas, and Bulevich, 2015). For example, if a witness is asked, "What color was the robber's hat?" on the initial test, the witness effectively receives a cue that the hat color is important. When the witness is later presented with post-event information, they will pay more attention to information concerning the hat color, which increases the likelihood of them learning the misinformation. Some researchers studying RES have measured increased attention by recording how long participants take to read the misinformation narrative and what is typically found is that subjects who took an initial test spend more time reading sentences that contain misinformation (e.g. Gordon, Thomas, and Bulevich, 2015). Unlike Gordon, Thomas, and Bulevich (2015), Butler and Loftus (2018) did not find that taking an initial test affected the time spent reading sentences containing misinformation.

There are three main types of variables in the RES paradigm: encoding variables (e.g., the type of material subjects initially learned and the type of misinformation); storage variables (e.g., the retention interval between the initial test and misinformation exposure); and retrieval variables (e.g., the format of the retrieval practice). Each of the variable types have been shown to moderate the RES effect. LaPaglia and Chan (2013) found that initial retrieval increased suggestibility when later misinformation (an encoding variable) was presented via an audio narrative, but not when it was presented via misleading questions. The format of the final test (a retrieval variable) has also been shown to produce dissimilar effects in the RES paradigm. For example, an RES effect has commonly been demonstrated using cued recall (e.g., Chan et al.,

2009; Gordon et al., 2015). A testing effect, however, has been found using other types of a final test. For example, Gabbert et al. (2012) found that free recall enhanced memory for the original event and LaPaglia & Chan (2012) found that initial testing increased performance on a lineup identification task.

Chapter 2 of this dissertation focuses on two primary RES experiments that attempt to extend the effect to new contexts. The first explored the RES effect in a classic multiple-choice testing educational context. The second experiment aimed to determine whether retrieval-enhanced suggestibility occurs in a lineup identification context.

Researchers have put forth several hypotheses as to why the RES effect occurs, but empirical support for the hypotheses is inconsistent. Further, although there are many published studies demonstrating an RES effect, researchers have not reached convergence on the conditions under which the RES effect does or does not occur, how generalizable the effect is, and the mechanisms driving the effect. Due to the recent emergence of this phenomenon, now is an ideal time to perform an empirical review through meta-analysis (Chapter 3). Specifically, the objectives of the meta-analysis are to identify: (a) the overall size of the retrieval-enhanced suggestibility effect; (b) the methodological factors that moderate the effect size; (c) the boundary conditions of the effect; and (d) theoretical mechanisms underlying the effect.

Chapter 1: Retrieving Information from Memory

The Testing Effect

Over the years, research has shown that taking a test after learning new information leads to better long-term retention of that information. During the first phase of a typical laboratory experiment examining the testing effect, participants study some material for a set amount of

time. Following this study phase, some subjects take a practice test that assesses their learning of the material they just studied. Other subjects do something else, such as restudy the material, perform another task, or nothing at all (Adesope, Trevisan, & Sundararajan, 2017). In the classic Gates (1917) study, children of various ages studied nonsense syllables and short biographies. Then, Gates instructed some of the children to look away from the material and try to recall the information on their own (self-testing). Later, he tested their retention of the information. He found that the children that self-tested had better retention of the material than the other students. Gates's work, however, was limited because young children may not have the ability to test themselves. More recently, one of the clearest examples of the effects of testing, restudy, or doing neither comes from a study by Nungester and Duchastel (1982). In their study, 97 high school students read passages from a history textbook. The students were then divided into three conditions: students in the *initial test* condition took a test shortly after studying the passage; those in *restudy* condition were allowed to study the passage again, and those in the *control* condition performed a filler task. Two weeks later, students were tested on their retention of the material. Students in the *initial test* condition performed the best, followed by those in the *restudy* condition than those in the *control* condition. That same pattern of results occurred even after five months, when the researchers tested the same group of students again.

Research has also shown that repeated retrieval attempts are more beneficial than only one. For example, Hogan and Kintsch (1971) had subjects study a list of 40 words during the first study trial. One group of subjects proceeded to study the list of words for three more study trials, while the other group took three practice recall tests. All subjects then took a final recall test. They found that subjects who studied and had three recall test trials performed better on the final test than subjects who only had the study trials. Karpicke and Roediger (2008) examined

the effects of repeated study versus repeated retrieval on the learning of Swahili-English word pairs. They found that repeated retrieval enhanced long-term retention of the word pairs, whereas repeated study had no effect.

The time between repeated retrieval attempts also matters. Duchastel (1981) had high school students read study a passage from a history textbook. After studying the material, subjects took either a short-answer test, a multiple-choice test, a free-recall test, or performed filler tasks. All subjects were then tested two weeks later. Duchastel found a testing effect (increased retention due to initial testing) for those who took the short-answer test, but not for the other two types of tests. Experiments by Glover (1989) examined how the number and timing of practice tests affect the benefits of testing. In the first two experiments, subjects studied a 300-word essay and the parts of a flower, respectively. Subjects in the experimental conditions were tested on the studied material, while those in the control conditions were not. Four days after studying the material, all subjects then were given a final retention test. As expected, a testing effect was observed: subjects in the experimental conditions performed better on the final tests than subjects in the control condition. The fourth experiment consisted of two conditions: *spaced test* and *massed test.* Subjects in the *massed test* condition took an intervening test immediately after studying the material, while subjects in the *spaced test* condition took the intervening test two days after studying the material. Glover found that subjects performed best when they were given an intervening test two days after studying. In the final experiment, Glover examined the number of practice tests following study. On Day 1, all subjects were given a drawing of a flower and were instructed to study its different parts. Subjects were assigned to one of three experimental conditions: *one intervening test*, *two intervening tests (massed)*, *two intervening tests (spaced)*. The next day (Day 2), subjects in the *one intervening test* condition were tested on

the different parts of the flower, and subjects in the *massed* condition took the intervening test twice (the second right after the first). Subjects in the *spaced* condition took the intervening test once on Day 2 and again on Day 3. On Day 5, all groups (including a control condition) took the final retention test. Glover found a standard testing effect: subjects who took at least one test performed better than control subjects on the final test. Subjects who took two spaced intervening tests performed best, and there was no statistically significant difference in performance between the *massed* and *one test* groups.

Wheeler, Ewers, and Buonanno (2003) studied the effects of repeated testing and repeated studying, with particular interest in their differential effects on the rate of memory decay. In their first experiment, subjects studied a list of 40 words. Subjects in the *repeated test* condition took three recall tests (spaced roughly three minutes apart) before taking a final recall test either five minutes or 48 hours after the final test. Subjects in the *repeated study* condition had three additional study sessions before taking a final recall either five minutes or 48 hours after the final test. Wheeler and colleagues found that after five minutes, subjects in the *repeated study* condition were able to recall more words ($M \approx 14$) on the final test than those in the *repeated test* condition ($M \approx 9$). At 48 hours, there was no difference between the two groups. The second experiment was very similar to the first except for two changes: the final recall test was now after seven days (instead of 48 hours), and each group had an extra study or test trial (depending on condition). Like in Experiment 1, *repeated study* subjects performed better ($M \approx 21$) after five minutes than *repeated test* subjects ($M \approx 12$). Seven days later, however, the *repeated test* subjects performed better ($M \approx 9$) than those with multiple study sessions ($M \approx 6$). Together, these results demonstrate two interesting findings. First, after a very brief retention interval, multiple study sessions can be more beneficial than multiple test sessions. Second, after

a longer retention interval, there was either no difference between groups (after 48 hours) or multiple test sessions led to better performance than multiple study sessions (after seven days). Similar results were found by Roediger and Karpicke (2006). They had subjects study a scientific passage and then either write down as much as they could about the passage (free recall test) or restudy the passage. Subjects were then given a free recall test five minutes, two days, and seven days later. Like Wheeler and colleagues (2003) they found that after a very short delay (five minutes), the restudy group performed better than the retest group. However, two and seven days later, the trend reversed: subjects who were tested after studying the passage performed better.

Providing learners with the correct answer following a retrieval attempt has been shown to enhance the testing effect (Roediger and Butler, 2011). Pashler and colleagues (2005) presented subjects with Luganda-English word pairs. After two presentations, subjects were tested twice and given corrective feedback for both correct and incorrect responses. The researchers found that providing corrective feedback following incorrect responses dramatically improved final retention of the word pairs. Corrective feedback is believed to enhance learning by helping the test-taker correct the errors that were made on the test (Roediger and Butler, 2011).

While there has not been much debate whether feedback is beneficial to learning, researchers do hold different views about the *timing* of the feedback. Kulik and Kulik (1988) conducted a meta-analysis on the timing of feedback on verbal learning. They found immediate feedback to be more beneficial than delayed feedback, concluding, "to delay feedback is to hinder learning." (Kulik and Kulik, 1988, p. 94). More recent research, however, has shown that delayed feedback can be beneficial for learning. For example, Butler, Karpicke, and Roediger

(2007) examined the type and timing of feedback on learning from multiple-choice tests. Subjects were given feedback immediately or after a 10-minute distractor task. The type of feedback was varied: some subjects received standard corrective feedback, while other subjects were allowed to keep answering until they chose the question correct (answer-until-correct-feedback). Subjects took a final test the following day. Although they did not find feedback type to make a significant difference, they did find that subjects who received delayed feedback performed better on the final test than those who received immediate feedback.

Theoretical explanations. There are several classes of theories that have been put forth to account for the testing effect: *overlearning/overexposure, effortful retrieval, elaborative retrieval,* and *transfer appropriate processing* (Rowland, 2014). The bifurcation model, a framework for interpreting testing effect results, is also discussed.

Overlearning/overexposure. Overlearning/overexposure refers to the study of newly-learned information beyond the point of initial mastery. Some researchers have posited that the testing effect is merely the result of overexposure to the studied and later tested material (e.g., Thompson et al., 1978). Slamecka and Katsaiti (1988) had subjects study a list of paired associates, and subjects in the experimental condition were tested on subsets of the list over multiple sessions. They found no significant differences in the rate of memory decay between subjects who were tested and subjects who were not. This suggests that testing itself does not lead to better retention, rather, it is the overlearning of a portion of the material (Roediger & Karpicke, 2006).

Effortful retrieval. Proponents of the theory of effortful retrieval (or retrieval difficulty) posit that the testing effect is directly related to how much effort/difficulty was involved in the retrieval attempt (Rowland, 2014). Some researchers (e.g. Jacoby, 1978; Karpicke & Roediger,

2007) suggest that as the difficulty of the retrieval attempt increases, so should the size of the testing effect. One way that researchers have demonstrated effortful retrieval is by increasing the interval between testing and retrieval (e.g. Rowland, 2014; Whitten & Bjork, 1977). For example, Pyc and Rawson (2009) manipulated both the intervals between study/test phases, as well as the number of items subjects had to correctly retrieve. Their results supported the retrieval effort hypothesis: as the difficulty of retrieval increased (either by longer retention intervals or by a greater number of items needed to be recalled), so did the magnitude of the testing effect.

Although studies have demonstrated that effortful retrieval can lead to larger testing effects, the specific causal mechanism has not been determined. Some researchers believe that retrieval increases the number of retrieval routes that one can use when searching for information (Rowland, 2014). For example, McDaniel & Masson (1985) found that cued recall strengthened the memory's representation and increased its variability (the number of ways to access the memory). Similar results have been found in other research (e.g. Cuddy & Jacoby, 1982; Glenberg, 1979).

Elaborative retrieval. The elaborative retrieval hypothesis proposes that the act of retrieval results in the elaboration of a memory trace, which leads to a testing effect (Rowland, 2014). Due to elaboration, the memory is more easily retrieved. Carpenter (2009) had subjects study cue-target pairs that were either strongly associated (e.g., Toast-Bread) or weakly associated (e.g., Basket-Bread), followed by either a cued-recall test or restudy. She found that retention for restudied items were similar regardless of whether they were weakly or strongly associated with each other. However, tested items that were weakly associated were best retained over time. These results support the elaborative retrieval hypothesis; weakly-associated items,

which required more elaborative retrieval processes, showed greater improvement as a result of initial testing than strongly associated items which were already well-recalled.

Transfer-appropriate processing. The final theory of the testing effect to consider is the theory of transfer appropriate processing *(TAP)*, which suggests that the testing effect is a result of processing similarities that result from the overlap in the initial and final tests (Rowland, 2014). According to TAP theory, the degree of similarity between the two tests should moderate the strength of the testing effect (i.e. the more similar the two tests, the greater the testing effect) (Rowland, 2014).

Although transfer appropriate processing is a very intuitive theory to explain the testing effect, in practice the results are mixed (Rowland, 2014). As discussed previously, Duchastel and Nungester (1982) had students read passages, take a test, and then take a final recall test after a two-week delay. The final test contained both multiple-choice and short-answer questions. The researchers found that students who took an initial multiple-choice test performed better on the multiple-choice portion of the final test. Johnson and Mayer (2009) had subjects study a narrated animation, followed by either restudy or two types of practice tests. They found that subjects who took a practice-retention test after watching the animation performed the best on the final retention test (which was of the same format), supporting a transfer appropriate processing theory of the testing effect. On the other hand, some researchers have failed to find support for TAP theory. For example, Carpenter and DeLosh (2006) employed a fully-crossed design to investigate the TAP and elaborative retrieval theories of the testing effect. After studying 12 lists of nouns, subjects were tested on four of the lists via free recall, four via cued recall, and four via recognition. The remaining four lists were given an additional study period. Subjects were randomly assigned to one of three final test conditions: free recall, cued recall, or recognition.

The researchers found that subjects who had matched intervening and final tests did not perform better than those who did not, failing to find support for TAP theory.

In summary, the support for TAP theory is inconsistent. However, it is possible that TAP theory can be useful for understanding the effects of retrieval practice on suggestibility.

Bifurcation framework. In this framework, test and study condition items rest on independent probability distributions (Rowland, 2014). After restudy, all items generally increase in memory strength. After testing, however, there is a bifurcation such that successfully-retrieved items increase in memory strength, while unsuccessfully-retrieved items remain the same. According to the bifurcation framework, the larger the proportion of successfully-retrieved items during an initial test, the larger the testing effect should be (Rowland, 2014). Unsuccessfully-retrieved items do not see any benefit of initial testing, leaving their memory strength static.

Summary. In a recent meta-analysis, Rowland (2014) investigated the effect of testing versus restudy on retention. As expected, testing led to better retention than restudy. He did not find support for the TAP theory of the testing effect: the effect size for studies with matching initial and final tests did not significantly differ from the effect size where the two tests were mismatched. He did, however, find support for the retrieval effort theories of the testing effect. As discussed previously, according to retrieval effort theories, the magnitude of the testing effect is positively associated with how difficult or effortful the test is (e.g. Pyc and Rawson, 2009). Generally, free-recall tests are considered the most difficult, followed by cued-recall, and finally recognition. Rowland (2014) found that the magnitude of effect was indeed correlated with final test difficulty; free-recall tests produced the largest effect size ($g = 0.82$), followed by cued-recall ($g = 0.72$), and finally recognition ($g = 0.36$). In summation, the testing effect has been shown to

be reliable and robust. Even when varying the learning context, study materials, and test type, the result remains the same; after learning, testing leads to better retention than restudy (or nothing).

Retrieval Enhanced Suggestibility

Given the extensive literature showing that testing leads to better memory performance, it is surprising that testing can lead to worse memory performance in certain conditions. Specifically, after encoding some information, taking an initial test can cause individuals to be MORE susceptible to misinformation than those who were not tested. This finding is known as *retrieval-enhanced suggestibility* (RES; Figure 1; Chan & Langley, 2011). In the initial study on RES (Chan, Thomas, and Bulevich, 2009), subjects watched a video of a television show that depicted criminal activity. Immediately after watching the video, half of the subjects were tested on details from the video, while the other half completed an alternate task. All subjects were later exposed to misinformation contained in a post-event narrative that summarized some of the details in the crime video. Then they took a final test on the details of the video. Chan and colleagues found that the subjects who took the initial test were more susceptible to misinformation, performing worse on the final test than those who did not take the initial test.

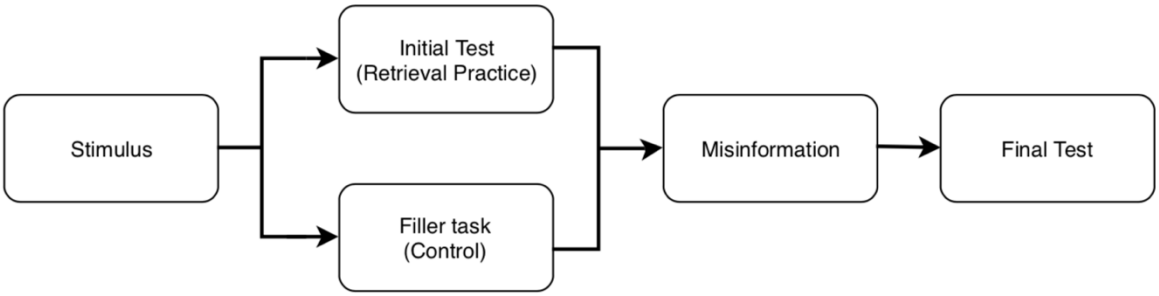
Theoretical explanations. There are three main theoretical explanations that researchers have put forth in an effort to explain (or, at least, partially explain) the RES effect: *test-potentiated learning (TPL)*, *memory reconsolidation*, and *misinformation acceptance*.

***Test-potentiated learning*.** Research has shown than that testing can potentiate the learning of subsequent learning of new information (Pastotter and Bauml, 2014). For example, Wissman, Rawson, and Pyc (2011) had subjects study a three-section text describing government intervention in the United States labor market. Subjects in the interim test group took intervening free-recall tests after reading each of the three sections. Subjects in the control condition also

took a free-recall test after reading section three, followed by free-recall tests for sections one and two. The researchers found that subjects in the interim test group performed better than the control condition on the free-recall test for section three, although both groups took the test immediately after reading the section. Thus, the authors found a forward effect of testing such that interim testing facilitates the subsequent learning of new information (Wissman et al., 2011). In the context of RES, after retrieval practice the to-be-learned "new information" is the misinformation, which would result in an RES effect.

Weinstein, McDermott, and Szpunar (2011) found a forward effect of testing when using face-name pairs. Subjects studied four lists of face-name pairs. After studying the first three lists, subjects in the *tested* condition took a cued-recall test on pairs from lists one through three, while subjects in the control condition performed math problems. After the fourth list, both groups were tested on face-name pairs from that list. Following that test, all subjects were given a cumulative test for all face-name pairs that they had studied. The researchers found a forward effect of testing; when tested on the fourth list, subjects in the *tested* condition performed better than control subjects, despite the fact that both groups took the test immediately after studying the list. Further, subjects in the *tested* condition outperformed than control subjects in the cumulative test as well.

Figure 1. Retrieval-enhanced suggestibility procedure



Why does testing facilitate the learning of subsequent learning of new information? According the *retrieval explanation*, being tested on sets of information increases set differentiation, where memory traces for each set are enhanced and made more distinct, which improves recall and reduces memory intrusions (Pastotter and Bauml, 2014). Evidence for this is found in the aforementioned study by Wissman and colleagues (2011). While being tested on the fourth list, subjects in the control condition experienced approximately 12 times more prior list intrusions than subjects who were previously tested on lists one through three.

According to the *encoding explanation*, testing facilitates subsequent learning by improving the encoding of the new material (Pastotter and Bauml, 2014). Pastotter et al. (2011) have suggested that after studying multiple lists, intermittent testing leads to a "reset" of the encoding process, which makes encoding of later lists as effective as the encoding of earlier ones. Looking at EEG data, they found that when subjects studied multiple lists without retrieval between lists, there was a buildup of power in the alpha band, which led to poor recall. Other research has supported this view that buildup in alpha power can lead to poorer encoding of items (e.g. Sederberg et al., 2006). In contrast, intermittent retrieval attempts break up this buildup, which resets the encoding process, allowing subsequent lists to be learned as efficiently as prior ones (Pastotter et al., 2011).

Test-potentiated learning appears to be an intuitive explanation for retrieval-enhanced suggestibility. After encoding the initial event the initial test facilitates learning of subsequent information, which in the RES paradigm is misinformation. However, test-potentiated learning in the RES paradigm seems to be at odds with some of the other benefits of initial test, such as increased list discrimination (e.g. Chan & McDermott, 2007) and a reduction in interference between lists (e.g. Szpunar et al., 2008). If the initial witnessed event and later misinformation are thought of as two separate lists of information, the intervening test should increase memory for both while also increasing a differentiation between the two sets of information. On the final test, subjects are typically instructed to answer questions based on what they saw during the initial witnessed event (e.g. Butler and Loftus, 2018). Thus, the intervening test should prevent memory intrusions from the misinformation narrative.

Memory reconsolidation. Once a memory trace is acquired, it stabilizes and moves into long-term memory. This process is known as consolidation (Dudai, 2004). Some research has shown that memories are not consolidated shortly after being acquired (e.g. McGaugh, 1966). Instead, the consolidation process has been shown to occur slowly over a span of several hours. During that time, our memories are susceptible both impairment and enhancement (McGaugh, 1966). For example, Mondadori and Ducret (1991) demonstrated that the use of nootropic CGS 5649B improved retention in mice for up to 24 hours after initial encoding. After a memory trace has been stored and subsequently activated, such as during a retrieval attempt, it must go through another round of consolidation (*reconsolidation*; Przybyslawski & Sara, 1997). According to the *reconsolidation hypothesis*, once stored memories are activated, they are highly susceptible to interference (Nader, Schafe, & Le Doux, 2000). Although the initial finding occurred in mice via

the loss of a previously-formed fear memory, it has been replicated in animals and in humans in a variety of contexts (for a review see Besnard, Caboche, & Laroche, 2012).

Applying the reconsolidation hypothesis to the retrieval-enhanced suggestibility paradigm, testing after encoding the initial event can leave the memory trace more susceptible to interference. While the memory is in this labile state, it is possible that the exposure of misinformation can disrupt the reconsolidation of the original memory trace, either leading to consolidation of the misinformation or misinformation intrusions at the time of recall. The reconsolidation hypothesis cannot (at least fully) account for the RES effect, considering consolidation is a process that occurs over a protracted period of time (McGaugh, 1966). In a typical RES study, subjects are tested immediately after encoding the initial event (e.g. Chan et al., 2009). Because of this narrow timeline, it is unlikely that memories have gone through complete consolidation and were subsequently reactivated by the initial test. Further, if newly acquired memories are susceptible to interference (Nader & Hardt, 2009), then everyone in an RES study should be more suggestible, not only the subjects who took an initial test.

Misinformation acceptance. Although research has shown that misleading post-event information can impair memory for an event (for an overview, see Loftus, 2005), some researchers have argued that the misinformation effect is not evidence for memory impairment (e.g. McCloskey & Zaragoza, 1985). For example, Belli (1989) posited that the misinformation effect is attributed, at least in part, to misinformation acceptance. It is possible that subjects view the misinformation narrative as corrective feedback, thus updating the memories with this new, accurate information (Chan, Manley, and Lang, 2017). It is also possible that subjects simply accept the details in the misinformation narrative to be true, or they fail to notice discrepancies between what they witnessed originally and misinformation in the post-event narrative. Despite

being an important determinant of misinformation acceptance, measures of discrepancy detection are seldom-used in misinformation research (Butler & Loftus, 2018). Loftus (1979) found that when subjects are presented with blatantly contradictory information, they were more likely to reject it and were more resistant to other pieces of misinformation. Researchers have also found that reading misinformation narratives more slowly is associated with increased scrutiny, which leads to a greater likelihood of detecting discrepancies and resisting misinformation (Tousignant, Hall, & Loftus, 1986).

This finding seems relevant to – and perhaps at odds with – some RES findings, particularly to the results found by Gordon et al. (2015). Tousignant et al. found slower reading times to be associated with discrepancy detection and increased resistance to misinformation, while Gordon and colleagues found that slower reading times lead to increased learning of misinformation. A relevant study is one by Butler and Loftus (2018, Experiment 1) who measured discrepancy detection in the retrieval-enhanced suggestibility paradigm. We found that discrepancy detection was a key predictor of who endorsed misinformation on the final test: those who noticed discrepancies between the original event and the misleading post-event narrative were less likely to endorse misinformation. Although the theory of misinformation acceptance appears to be a plausible explanation for misinformation endorsement *in general*, it cannot explain why misinformation endorsement rates are higher for participants that engaged in retrieval practice.

Summary. Several theories have been put forth to explain the retrieval-enhanced suggestibility effect, but none of them can completely explain the exact causal mechanisms behind the finding. Like the many theories relating to the testing effect, it is possible that they are

all contributing someway to the overall RES effect. In the following section, I discuss two primary research studies where I try to extend the RES effect to different contexts.

Chapter 2: Experimental Work

This chapter reports the results of two new RES studies; *Multiple-choice testing* (Butler, 2018 Experiment 1) and *Lineup Identification* (Butler, 2019 Experiment 1).

Primary Study 1: Multiple-choice testing (Butler, 2018 Experiment 1)

It is possible that the intensity of the RES effect — or whether it occurs at all — is contingent on the type of information being encoded initially; the vast majority of RES studies have been conducted in an eyewitness scenario, using mock crime videos (e.g. Wilford, Chan, & Tuhn, 2014) or slideshows (e.g. Butler & Loftus, 2018). However, the implications of retrieval-enhanced suggestibility might extend to a variety of contexts. Education, for example, is a domain where retrieval practice has perhaps the most prominent role. Pre-tests, practice tests, pop quizzes, and self-testing are prominent fixtures in Western schooling (Adesope et al., 2017; Rowland, 2014). The goal of the present research is to examine whether retrieval practice enhances suggestibility in a multiple-choice testing context.

Multiple-choice testing is used heavily throughout all levels of education. Multiple-choice tests are an attractive option for instructors because they are easier to grade and less subjective than other test formats (Butler, 2018). Further, students prefer multiple-choice questions over other formats such as short-answer and essay-type questions (Struyven, Dochy, & Janssens, 2005). Despite multiple-choice tests having positive memorial benefits, there are negative consequences of multiple-choice testing. For example, as the number of lures increases, so does the likelihood that a learner will select the incorrect answer on a later test (Roediger &

Marsh, 2005). Further, if a learner erroneously endorses a lure it can lead them to acquire false knowledge (Butler & Roediger, 2008). In addition to acquiring misinformation from the test itself, learners are at risk of acquiring faulty information from outside sources, such as study guides. Typically, study guides enhance learning by increasing engagement with the material and by presenting learners with the most important salient information (Mafinejad et al., 2014). However, it is possible for erroneous information in study guides or outside discussions to be learned as factual knowledge. The goal of the present research was to examine whether retrieval-enhanced suggestibility occurs in a common testing paradigm. In both experiments, students studied 12 prose passages relating to either historical or scientific information. Half of the students immediately took a multiple-choice test after studying, while the other half performed filler tasks. Later, all participants read summaries for each of the 12 passages. Unbeknownst to them, four of the summaries contained erroneous information. Finally, all students took a final multiple-choice test that covered details from each of the 12 passages.

The testing effect literature has shown that retrieval practice enhances memory for learned information and protects learners from retroactive interference. Conversely, the retrieval-enhanced suggestibility and test-potentiated learning frameworks suggest that retrieval-practice will enhance the learning of subsequent misinformation. In reconciliation of these seemingly competing ideas, I hypothesized that positive and potentially negative effects of testing would counteract each other such that retrieval practice would ultimately not enhance susceptibility to misinformation.

Method

91 undergraduate students (74 female, 16 male, 1 other) from the University of California, Irvine participated in this experiment in exchange for course credit (*Mage* = 23.57,

SD = 6.75). Participants were randomly assigned to one of the two conditions: control (*N* = 46) or retrieval practice (*N* = 45).

The stimuli used in this experiment were 12 prose passages used in previous testing effect research (e.g. Butler & Roediger, 2008; Roediger & Marsh, 2005). The passages covered historical topics or persons (e.g. the history of basketball, the history of education, humanitarian Dorothea Dix's life story). Each passage consisted of approximately 300 words, broken into four paragraphs. The 12 passages were presented to participants in random order. Post-event information was presented to participants in the form of 12 short summaries which ostensibly reiterated four "key points" from each of the 12 passages. However, there was a "misinformation" version of each narrative that consisted of three accurate statements and one inaccurate statement (the misinformation). For example, the original passage would correctly state that Dorothea Dix was born in *Maine*; the accurate version of the summary would state, "Dorothea Dix was born in *Maine*", while the misinformation version would state, "Dorothea Dix was born in *Delaware*". Eight of the 12 passage summaries presented to participants were accurate, while four of the summaries were misinformation versions. In order to minimize any specific item effects, the computer program randomly selected which of the four summaries would be presented to participants in their misinformation version.

Participants' memory for the passages was tested using a 48-item multiple-choice test. The test was divided into 12 sections, one for each of the 12 passages. Each question corresponded to one of the key points from the corresponding passage and narrative summary. When answering each question, participants could select one of six possible responses: the accurate "target" response, the misinformation lure, or one of four neutral lures.

Participants were told that they would be studying a series of passages that they would later be tested on. They were instructed to study the passages carefully and that they would have 90 seconds to study each passage. After 90 seconds elapsed, the page automatically advanced to the next passage. Participants were then randomly assigned to either the *retrieval practice* or *control* condition. Participants in the retrieval practice condition took the 48-item multiple-choice test immediately after studying the final passage, while those in the control condition completed a filler task (health and lifestyle surveys). Following either the initial test or filler task, all participants performed a series of unrelated distractor tasks (behavior and attitude questionnaires) to fill a 20-minute retention interval. After completing the distractor tasks, participants were told that they would be presented with summaries of the 12 passages they read earlier. The summaries were presented one at a time, and they were given a maximum of 15 seconds to read each summary before the computer automatically advanced to the next summary. As detailed in Section 2.1.3, eight of the summaries were completely factual and four contained misinformation. After reading the final summary, participants took a final, 48-item multiple-choice test (which was identical to the initial test taken earlier by participants in the retrieval practice condition).

Results and discussion

The primary conclusions reported here are based on the results from Bayesian statistical analyses. In lieu of *p* values, which have a number of significant limitations (see Wagenmakers, 2007; Nuzzo, 2014), I instead report Bayes factors. The Bayes factor (BF) quantifies and compares the predictive fitness of two competing statistical models (Wagenmakers et al., 2018). In the present experiments, the two competing models are the null (H_0) and alternative (H_1) hypotheses. As an example, imagine an outcome where the Bayes factor $BF_{10} = 5$. This

indicates that the data are 5 times more likely under H_1 than under H_0. Similarly, the Bayes factor $BF_{01} = 5$ indicates that the data are 5 times more likely under H_0 than under H_1.

Bayesian hypothesis testing and parameter estimation has many benefits in comparison to traditional null hypothesis significance testing (for a detailed review, see Wagenmakers et al., 2018). Most notably, the Bayes factor quantifies the amount of evidence for the null vs. the alternative hypothesis. In contrast to merely "rejecting" the null hypothesis based on $p > .05$, the Bayes factor can quantify the support that the data provide the null hypothesis. All analyses reported in this paper were performed using the *R* (R Core Team, 2018) package *BayesFactor* (Morey & Rouder, 2018) and JASP (JASP Team, 2018) using the default Cauchy prior distribution parameters.

As a reminder, only participants in the retrieval practice condition (from here on referred to as "RP participants") took an initial multiple-choice test immediately after studying the 12 prose passages. Performance on the initial test was quite low ($M = 0.50$, $SD = 0.20$). Participants endorsed misinformation lures at rates similar to chance ($M = 0.16$, $SD = 0.14$). This was expected, as participants had not yet been presented with the misinformation summaries.

As predicted by the testing effect, retrieval practice participants performed much better on the second test in comparison to the first ($M_{diff} = 0.25$, $BF_{10} = 1.76e+10$). A Bayesian repeated-measures analysis of variance indicated extreme evidence of a standard misinformation effect. Overall, participants performed much better on consistent trials ($M = 0.75$) than on misinformation trials ($M = 0.18$), $BF_{10} = 1.74e+37$. However, participants in the retrieval practice condition ($M = 0.28$) were not more likely to endorse misinformation on the final test than control participants ($M = 0.39$). A Bayesian independent samples t-test resulted in the Bayes factor $BF_{01} = 1.86$, which indicates that the data are 1.86 times more likely under the null

hypothesis that retrieval does not enhance suggestibility. In other words, the data did not provide evidence of an RES effect. The data did not support the hypothesis that initial testing would reduce retroactive interference, leading to a reduction in suggestibility. Retrieval practice (M = 0.15) and control (M = 0.21) participants performed similarly on misinformation trials, BF_{01} = 2.12.

Participants endorsed lures on the final test 30% of the time. For participants in the retrieval practice condition, the odds of endorsing a lure on the final test were 3.65 times higher if they also endorsed a lure for the corresponding question on the initial test, OR = 3.65, 95% CI [2.76, 4.88].

In summary, retrieval practice did not result in an increase in suggestibility.

Primary Study 2: Lineup Identification (Butler, 2019 Experiment 1)

As mentioned previously, the majority of RES research occurs in the mock-crime paradigm. Despite being a large focus of misinformation research, there is a noticeable lack of lineup identification tasks in retrieval-enhanced suggestibility research. The lone exception would be in LaPaglia & Chan (2012); they had participants go through a traditional RES experiment, but the final test was a target-present lineup identification task. During the misinformation phase, the participants in the retrieval practice condition were given misleading details about the appearance of the suspect. The researchers found misinformation reduced target identifications from 23% to 12%. In an effort to extend the findings of this study, I designed an experiment to test whether subjects would be more suggestible when making lineup identifications after being shown a misleading suspect photograph.

Method

137 undergraduate students from the University of California, Irvine participated in this experiment in exchange for course credit. Participants were randomly assigned to one of the two conditions: control ($N = 69$) or retrieval practice ($N = 68$).

The stimulus used in this experiment was a 50-second mock-crime video used in Murphy and Greene (2016). The video depicts a young woman entering an office and stealing various items and then leaving. Unbeknownst to the perpetrator, a man witnessed the crime by looking through a window. Immediately following the video, participants in the retrieval practice condition took cued-recall test based on the events that occurred in the video. Following the recall test, retrieval-practice participants performed two lineup identification tasks for the thief and the witness, respectively. The lineups were both target-present, containing the thief/witness and four foils.

Post-event information was presented to participants in the form of a written misinformation narrative that ostensibly summarized what happened in the video. Accompanying the misinformation narrative were photos of the thief and the witness. However, the photo of the "thief" was randomized between either the actual target or one of the four foil photographs. Finally, all participants took a final recall test and completed the same lineup identification tasks that the retrieval-practice participants completed previously.

Results and discussion

Participants in the retrieval-practice condition endorsed text-based misinformation at a higher rate than participants in the control condition — an RES effect ($M_{RP} = 0.57$, $M_{C} = 0.33$). A Bayesian independent samples t-test produced a bayes factor of 7.70, indicating moderate to strong evidence for the alternative hypothesis that retrieval does enhance suggestibility.

When looking at the results for the thief lineup identification task, participants in the retrieval-practice group endorsed the foil (misinformation) at slightly higher rates than those in the control condition. A Bayesian independent samples t-test produced a bayes factor of 1.43, indicating weak evidence for the null hypothesis that there is no difference in suggestibility between the retrieval practice (M = 0.69) and control (M = 0.58) conditions.

In summary, this experiment demonstrated that retrieval can enhance suggestibility for some types of information. Specifically, retrieval enhanced suggestibility for information where misinformation was delivered via a narrative. When misinformation was presented in the form of an erroneous suspect photograph, there was not enough evidence to support the claim that retrieval enhances suggestibility.

The two primary studies reported here, in conjunction with the mixed results from Butler and Loftus (2018), led me to believe that obtaining or replicating the RES effect was not straight-forward. Further, when I did observe an RES effect, the effect size was often smaller than what other studies have reported. As a result, I chose to conduct a meta-analysis in order to obtain a pooled estimate of the RES effect size as well as in order to identify the variables that moderate the effect.

Chapter 3: Meta-Analysis

In Chapter 1 I discussed the inconsistent findings in the retrieval-enhanced suggestibility literature. In Chapter 2 I described two primary studies that attempted to extend RES to new contexts. The next step in my investigation was meta-analysis. Instead of drawing conclusions from a collection of individual studies qualitatively, the meta-analysis allowed me to synthesize and summarize the data from the available studies quantitatively.

Moderator Variables

One of the primary goals of this meta-analysis was to identify the factors that moderate the RES effect. There are three main classes of theoretical moderators in the RES paradigm: encoding variables (e.g., the type of material subjects initially learned and the type of misinformation); storage variables (e.g., the retention interval between the initial test and misinformation exposure); and retrieval variables (e.g., the format of the retrieval practice). Additionally, I included two methodological moderators — *research group* and *publication status*. Table 2 contains the list of moderator variables and corresponding hypotheses.

Report characteristics.

Publication status. Research has shown that studies with statistically significant findings are more likely to be accepted for publication (see Dickersin, 2005). Researchers that find small or nonsignificant effects tend to simply store the results of their research in their "file drawers", which leads to the scientific community only being aware of statistically significant effects (Rosenthal, 1979). In an attempt to combat this publication bias, I contacted researchers for unpublished RES studies. Publication status will be included as a categorical moderator with two levels: published and unpublished. Theses, dissertations, and other research reports obtained from researchers were coded as unpublished. Due to psychological science's tendency to prefer statistically significant findings (e.g., Ferguson & Brannick, 2012; Rosenthal, 1979), I hypothesized that published research reports will produce larger effect sizes than unpublished reports.

Research group. A noticeable proportion of retrieval-enhanced suggestibility research has been conducted by the author that discovered the effect (Chan) and their colleagues (e.g. Bulevich, Gordon, LaPaglia, Thomas, and Wilford). Research group will be used as a moderator to identify whether those researchers produce studies with larger RES effects than the rest of the

field. I hypothesized that the Chan & colleagues research group will produce studies with larger effect sizes than other researchers. This hypothesis is solely driven from my observation that Chan and colleagues appear to produce reports with larger RES effect sizes than other researchers.

Encoding variables.

Stimulus type. The most common stimulus used in RES research is the depiction of a mock crime. Chan and colleagues (2009) showed participants an episode of the television show 24, which depicted a bank robbery taking place. They found an RES effect, similar to other researchers who have used mock-crime videos (e.g., Chan et al., 2012; Wilford et al., 2014). RES effects have also been found in studies where participants viewed mock-crime slideshows (e.g., Butler & Loftus, 2018; Rindal et al., 2016), learned factual information (e.g., Mullet et al., 2014), and studied word pairs (Weber, 2012). In contrast, some studies have failed to find an RES effect. Huff et al. (2016), for example, found that retrieval practice reduced suggestibility. In the study, participants viewed a series of scenes that many common household scenes (e.g., bathroom, kitchen) and were told that their memories for items in the scenes would be tested. After completing filler tasks, participants in the testing conditions took free-recall tests for each scene that they viewed. They were then required to review six recall tests that were ostensibly completed by previous test-takers but were actually researcher-created and contained misinformation. Immediately after reviewing the tests, participants took a final free-recall test that was identical to the first. The researchers found a protective benefit of testing, such that participants in the testing conditions were less likely to endorse misinformation on the final test. In other words, the authors found the opposite of an RES effect. Stimulus type will be coded categorically: mock-crime video; mock-crime slideshow; factual knowledge; word pairs/lists;

and other. Research has shown that different types of stimuli can affect processing and perception of information. For example, Sundar (2000) showed that participants' abilities to correctly recall news information differed based on the type of multimedia enhancements accompanied by it. I did not have a specific prediction about how stimulus type would moderate the RES effect; to my knowledge, past research has not examined the effect of different types of stimuli on misinformation endorsement. Thus, stimulus type was examined exploratorily.

Misinformation type. Misinformation is typically given to participants in the form of a narrative that is ostensibly an accurate summary the contents of the stimulus. Although subjects typically read the narrative, it is sometimes delivered auditorily (e.g., Butler & Loftus, 2018 Experiment 3; LaPaglia & Chan, 2013). Although narratives have typically facilitated an RES effect, other methods of delivering misinformation have not been as successful. For example, Pansky and Tenenboim (2011) did not find an RES effect when misinformation was presented to participants via misleading questions. Similarly, LaPaglia and Chan (2013) found that retrieval enhanced suggestibility when misinformation was delivered via a narrative but decreased suggestibility when it was delivered via misleading questions. Misinformation type will be coded categorically: written narrative; audio narrative; misleading questions; and other. The other category is reserved for studies that do not present misinformation in one of the above methods. Memon et al. (2010), for example, had participants forcibly fabricate information after a Cognitive Interview. Consistent with previous research, I hypothesized that misinformation delivered via a narrative (either audio or written) would produce larger RES effects than misinformation delivered via misleading questions or other methods. When misinformation is delivered via a narrative, it is possible that participants view the information as corrective feedback, passively updating their memories with the new information. In contrast, at the time of

test (where performing well is important), subjects might be more discerning with regards to the information they are being given.

Stimulus and misinformation exposure duration. The most common stimulus used in RES research is a mock-crime video, usually either 20 or 40 minutes in length. RES effects are less common in studies where participants are presented with shorter stimuli (and corresponding misinformation narratives). Butler and Loftus (2018) found inconsistent RES effects after presenting participants with a pair of three-minute slideshows. Huff et al. (2016) showed participants six color images of common household scenes (each presented for 15 seconds) and found that retrieval practice reduced suggestibility. LaPaglia and Chan (2012) showed participants a 45-second mock-crime video and Gabbert et al. (2012) showed participants a two-minute mock-crime video, both of which led to a reduction in suggestibility.

Research has shown that longer exposure time is associated with enhanced recognition accuracy (Palmer et al., 2013). So why, then, have studies with longer stimuli produced larger RES effects? The answer might be related to the length of the misinformation narrative and not the stimuli itself. Research has shown that receiving increasing amounts of misinformation leads to increased misinformation endorsement (Pena et al., 2017). The researchers hypothesized that this is due to increased cognitive demands participants face when being presented with a large amount of misinformation, which reduces their ability to detect the inaccuracies. Less misinformation is presented in shorter narratives, which might enable participants to more-easily detect inaccuracies in the information. Thus, I hypothesized that the size of the RES effect will be stronger with longer misinformation exposure duration. Exposure duration was coded in minutes as a continuous variable.

Storage variables.

Delay between stimulus and retrieval practice; Delay between initial retrieval practice and misinformation; and Delay between misinformation and final test. Researchers have found RES effects whether retrieval practice is immediate or not (e.g., Chan and LaPaglia, 2013). Chan and Langley (2011) found that retrieval enhanced suggestibility when misinformation was delivered shortly after the initial test (Experiment 1) and after a one-week delay (Experiment 2). In contrast, Chan and LaPaglia (2013) found that the RES effect was reduced when the misinformation delivery was delayed by 48 hours. Research has shown that longer retention intervals lead to a reduction in eyewitness identification and recognition accuracy, due to the tendency for memories to fade and become weaker over time (e.g. Deffenbacher et al., 2008; Ebbinghaus, 1964; Palmer et al., 2013; Sauer et al., 2010). Thus, I hypothesized that longer retention intervals will be positively associated with the size of the RES effect. Each of the delay variables will be coded continuously as seconds.

Retrieval variables.

Format of retrieval practice and final test. The prevailing form of retrieval practice in RES research is the cued-recall test. However, free-recall tests (e.g., Wilford et al., 2014) and cognitive interviews (e.g., LaPaglia et al., 2014) have also been shown to produce RES effects. Surprisingly, recognition tests have rarely been used as a form of retrieval practice in RES research. Other forms of retrieval practice have resulted in inconsistent findings. For example, Holliday (2003) and Gabbert et al. (2012) Cognitive Interviews did not increase susceptibility to misinformation. In fact, Self-Administered Cognitive Interviews in Gabbert et al. (2012) significantly decreased suggestibility. Retrieval practice type will be coded categorically as free recall, cued recall, recognition/AFC, modified-modified free recall, cognitive interview, and other.

RES effects have been observed in most final test types: cued recall (e.g. Chan et al., 2009); free recall (e.g., Wilford et al., 2014); recognition (e.g. Butler & Loftus, 2018); and modified-modified free recall (e.g., Gordon & Thomas, 2014). Despite RES effects occurring in various test types, not all researchers have found RES effects. The final test in Gabbert et al. (2012) was free recall, and they found a reduction in suggestibility. Rindal et al. (2016) failed to find an RES effect when the final test was recognition (although the misleading option was not presented, just a new lure). Final test type was coded categorically: free recall; cued recall; recognition/AFC; modified-modified free recall; cognitive interview; and other.

As discussed earlier, the effortful retrieval hypothesis states that more difficult retrieval attempts lead to better retention of the tested material (e.g. Jacoby, 1978). Free-recall tests are usually considered the most difficult, followed by cognitive interview, cued-recall tests, and finally recognition tests. In the RES paradigm, effortful retrieval (which leads to better recognition) should enhance memory performance and reduce suggestibility. Thus, I hypothesized that more difficult retrieval attempts will lead to smaller RES effects (in order of difficulty: free recall; cognitive interview; cued recall; and recognition).

Retrieval practice – final test match. This variable will be coded categorically as same if the retrieval practice and final test are the same format; and different if they are not. According to the theory of transfer-appropriate processing, the strength of the testing effect is positively associated with the similarity between retrieval practice and final test (Rowland, 2014). As a result, I hypothesized that a match in the format of the retrieval practice and the final test will lead to increased memory performance, which will result in smaller RES effects.

Warning. In misinformation research, warning participants that what they heard/read/saw significantly reduces the misinformation effect (for a meta-analysis on post-warning studies, see

Blank & Launay, 2014). A similar trend has been observed in RES research. In Thomas et al. (2010), after listening to the misleading audio narrative, some participants were told that the veracity of the narrative could not be verified. The authors found that this warning eliminated the RES effect. In Butler and Loftus (2018), participants were required to judge the accuracy of the misinformation narrative as they read or listened to it. The authors found that warning reduced the size of the RES effect. When participants were warned that there might be erroneous information in the narratives, they were more likely to detect the discrepancies and resist acceptance of the misinformation. I hypothesized that studies in which participants are warned about the veracity of the misinformation would produce smaller RES effects than studies without a warning. Several theories have been proposed to explain the RES effect, but none of them completely explain the exact causal mechanisms behind the finding. Further, methodological variation across studies has produced inconsistent and sometimes contradictory findings.

Given the inconsistent findings discussed and demonstrated in Chapters 1 and 2, the objectives of this meta-analysis were to identify: (a) the overall size of the retrieval-enhanced suggestibility effect; (b) the methodological factors that moderate this effect size; (c) the boundary conditions of the effect; and (d) the theoretical mechanisms underlying the effect.

Method

Search Strategy and Coding Procedure

This meta-analysis followed reporting guidelines outlined by The PRISMA Group (Moher, Liberati, Tezlaff, Altman, & The PRISMA Group, 2009). The flowchart of the search and screening process can be seen in Figure 2. The official literature search was conducted via ProQuest's electronic database search on August 5, 2019 and returned 733 results. The three databases searched were *PsycINFO & PsycArticles* (n = 547) and *ProQuest Dissertations and*

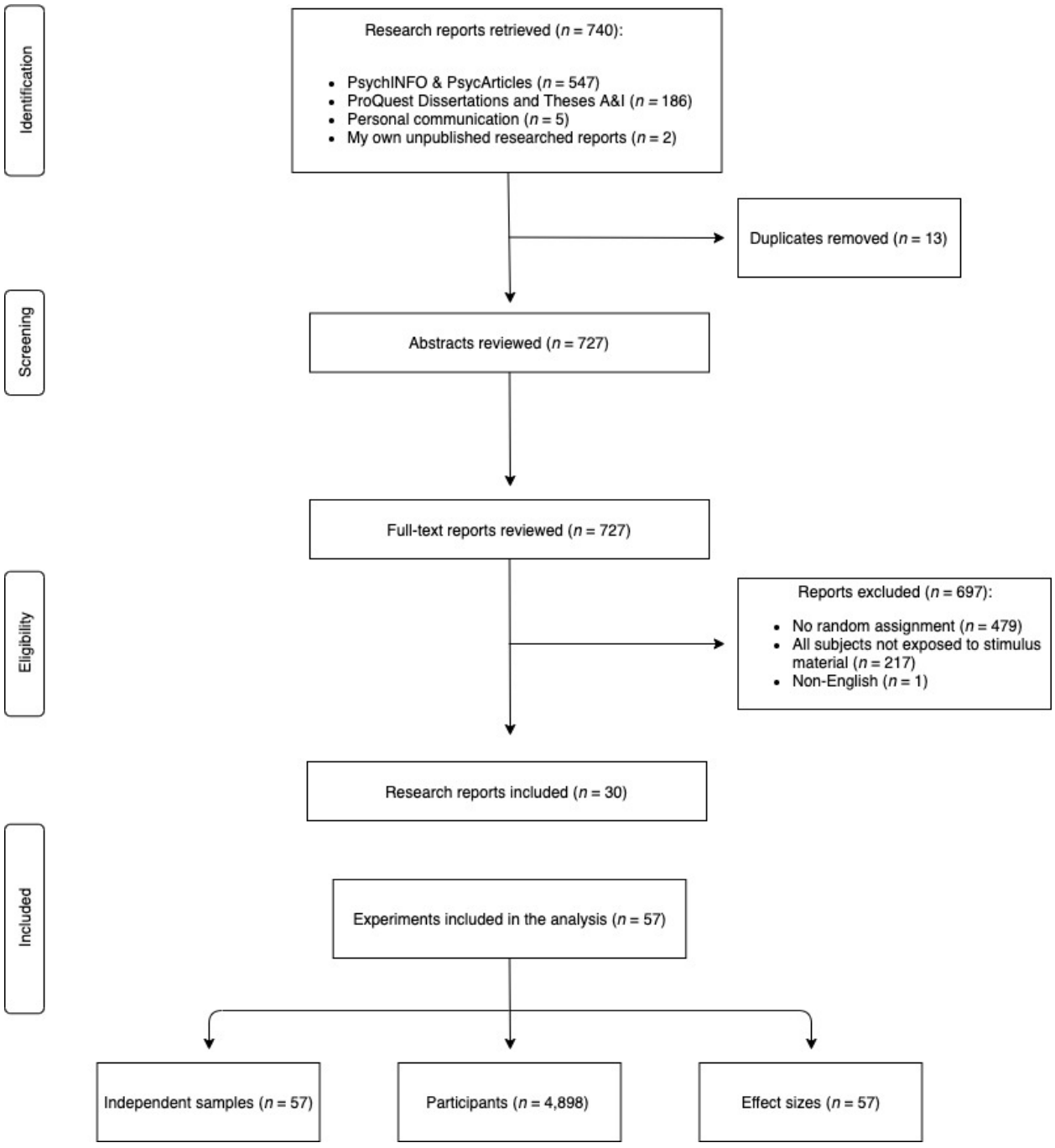
Theses A&I (k = 186). An overview of the terms and parameters for this literature search can be seen in **Table 1**. The search terms were: “retrieval enhanced suggestibility”; eyewitness AND suggestibility; eyewitness AND susceptibility; test* AND misinformation suggestibility; test* AND misinformation susceptibility; retrieval AND misinformation; “retrieval practice” AND misinformation; test* AND misinformation; "revers* test* effect"; and “RES” AND misinformation. Following the electronic database search, I also performed forward and reverse citation search on all papers included in the meta-analysis.

In August 2019 I made a call for papers on the Psychonomic Society’s listserv and asked leading RES researchers (Chan, Thomas, Gordon, Bulevich, and LaPaglia) for any unpublished RES research; These efforts resulted in five research reports. Additionally, I contributed two of my own unpublished research reports.

Table 1. Overview of literature search procedure for electronic databases

Search date	Search terms	Electronic databases	Documents retrieved
Monday, August 5, 2019	“retrieval enhanced suggestibility” *or* eyewitness AND suggestibility *or* eyewitness AND susceptibility *or* test* AND misinformation suggestibility *or* test* AND misinformation susceptibility *or* retrieval AND misinformation *or* “retrieval practice” AND misinformation *or* test* AND misinformation *or* "revers* test* effect" *or* “RES” AND misinformation	*PsycINFO* and *ProQuest Dissertations and Theses A&I* (2)	733

Figure 2. Flowchart of search and screening process



Inclusion criteria. All studies in the present meta-analysis followed the general RES procedure (Figure 1). Specifically, the five necessary criteria for a study's inclusion were: (1) all subjects were presented some type of stimulus material (e.g., mock crime video; 211 studies

removed); (2) random assignment (479 studies removed); all subjects exposed to stimulus material 217 studies removed); (4) all subjects were exposed to misinformation; (5) all subjects were tested on their memory for details in the originally-encoded stimulus (0 studies removed); and (6) the methods and results were reported in English (1 study removed).

Six main types of information were recorded from each study: (1) report characteristics; (2) encoding variables; (3) storage variables; (4) retrieval variables; (5) participant characteristics; and (6) effect size. A complete list of information that was coded can and the coding protocol can be found in Appendix A. I coded all studies and two trained independent researchers each coded a random 25% of studies. Interrater reliability was $k = 0.91$ and all discrepancies were resolved through in-person discussion.

Table 2. Moderators, hypotheses, and study characteristics

Moderators	*n*	*k*	Hypothesis
Research Group	57	29	Research reports from Chan & colleagues will produce larger RES effects than reports from other researchers
Chan & colleagues	36		
Other researchers	21		
Publication Status	57	29	Published research reports will have larger RES effects than unpublished reports
Published	52		
Unpublished	5		
Stimulus Type	56	28	
Videos and slideshows	49		No prediction
Word lists/pairs	2		
Pictures	5		
Factual knowledge	1		
Stimulus Length	56	29	Longer exposure duration times will lead to smaller RES effects
Misinformation Type	57	29	
Audio narrative	30		Misinformation delivered via a narrative (either audio or written) will produce larger RES effects than misinformation delivered via misleading questions or other methods
Misleading Questions	1		
Written narrative	16		
Other	10		
Misinformation Length	28	13	Longer exposure duration times will lead to larger RES effects
Retention Interval 1	49	24	Longer retention intervals will lead to larger RES effects.
Retention Interval 2	49	24	Longer retention intervals will lead to larger RES effects.
Retention Interval 3	42	22	Longer retention intervals will lead to larger RES effects.
Retrieval Practice Format	56	29	
Cognitive Interview	5		More difficult retrieval attempts (i.e., cognitive interview, free recall) will lead to smaller RES effects
Cued Recall	39		
Free Recall	8		
Recognition	2		
Other	2		
Final Test Format	56	29	
Cued Recall	42		More difficult retrieval attempts (i.e., cognitive interview, free recall) will lead to smaller RES effects
Free Recall	6		
Recognition	7		
Other	1		
Test Match	56	29	A match in the format of the retrieval practice and the final test will lead to increased memory performance, which will result in smaller RES effects
Match	44		
No match	12		
Warning	56	29	Studies in which participants are warned about the veracity of the misinformation will produce smaller RES effects than studies without a warning
Warning	4		
No warning	###		

Note. *n* - number of effect sizes; *k* - number of independent samples.

Effect Size Calculations

In meta-analysis, effect sizes are used to measure the consistency of an effect across studies and to calculate a summary effect (the estimated mean of our sample of effect sizes; Borenstein et al., 2009). An *effect size* is a measure that quantifies the size of a relation between two variables or the difference between them (Coe, 2002). In this meta-analysis, the effect size quantifies the magnitude and direction of the difference in misinformation endorsement between the retrieval-practice and control condition. The most common effect size used in retrieval-enhanced suggestibility research is Cohen's *d*, which represents the standardized difference between means from two independent groups. It is defined as:

$$d = \frac{M_{RP} - M_C}{SD_{pooled}}, \text{ (Equation 1)}$$

where M_{RP} and M_C represent the mean misinformation endorsement for the retrieval-practice and control condition, respectively. In the denominator, the pooled standard deviation is defined as:

$$SD_{pooled} = \sqrt{\frac{(n_{RP}-1)S_{RP}^2 + (n_C-1)S_C^2}{n_{RP} + n_C - 2}}, \text{ (Equation 2)}$$

where n_{RP} and n_c are the sample sizes from the two groups and S_{RP} and S_C are their standard deviations (Borenstein et al., 2009). Because Cohen's *d* tends to overestimate the true effect size in small samples, a correction factor *J* is used to reduce this bias (Borenstein et al., 2011). It is defined as:

$$J = 1 - \frac{3}{4df - 1}, \text{ (Equation 3)}$$

where *df* is the degrees of freedom used to estimate the pooled standard deviation $(n_{RP} + n_C - 2)$. The correction factor is then multiplied by *d*, which results in the unbiased effect size estimate, Hedges' *g*:

$$g = J * D. \quad \textit{(Equation 4)}$$

All analyses were conducted using Hedges' *g*, which is interpreted the same way as Cohen's *d* (Rowland, 2014). Positive values for *g* indicate higher misinformation endorsement in the *retrieval practice* group, and negative values indicate higher misinformation endorsement in the *control* group.

Statistical Approach

I chose a random-effects model (instead of a fixed-effect model) for the present meta-analysis. In a fixed-effect model, we assume that the estimated effects come from a single homogenous population (Schwarzer, Carpenter, & Rucker, 2015) Under this theoretical assumption, all of the factors that could influence the summary effect are identical across studies, with the only reason they find different effect sizes is due to sampling (i.e., estimation) error (Borenstein, et al., 2009). In contrast, a random-effects model estimates the *mean* of a distribution of true[1] effect sizes (Borenstein et al., 2009). Under this theoretical assumption we acknowledge that the studies included are similar enough to be meta-analyzed, however they are not functionally identical. Therefore, we would expect effect sizes to differ as a result of different population characteristics, stimulus material, and methodological differences.

[1] *True* effect size refers to the underlying effect size in the population that would be observed if the sample size was infinitely large. In contrast, *observed* effect size refers to the effect size that was observed and reported in a research study.

I used random-effects meta-regression to estimate all effects in this meta-analysis. Meta-regression is meta-analytic equivalent of traditional regression, where the model predicts an outcome $\hat{Y}$ (the RES effect) as a function *f(x)* of the explanatory variables (the moderators). I estimated the overall RES effect size using an intercept-only meta-regression model. In an intercept-only model, the coefficient is interpreted as the estimated mean effect size and the associated *t*-statistic is used to test the null hypothesis that the effect size is equal to zero. In models with continuous moderators, the meta-regression coefficient represents the estimated amount of change in the RES effect given a one-unit increase in the moderator. In these models, the *t*-statistic is used to test the null hypothesis that the relationship (slope) between the moderator and the RES effect is equal to zero. A statistically significant *p*-value on this test indicates that we can reject the null hypothesis, concluding that the moderator is significantly associated with the outcome. In meta-regression models with categorical moderators, the intercept coefficient represents the estimated mean effect size for the reference group and the associated *t*-statistic is used to test the null hypothesis that the effect size is equal to zero. The coefficients for the other subgroups in the analysis represent the estimated difference in effect size between each subgroup and the reference group, and the associated *t*-statistics are used to test whether each difference between means is equal to zero. I used a Wald test to determine whether there is a statistically significant difference between all levels of the categorical moderators (Rubio-Aparicio et al., 2019).

For each meta-regression, I used hierarchical robust variance estimation in order to produce robust estimates of the variance-covariance matrix of the model coefficients. Robust variance estimation (RVE) is used to account for the non-independence of effect sizes in meta-analysis (Tipton, 2015). Although each effect size in the present meta-analysis is ostensibly

independent, there is clear non-independence that can be accounted for — specifically, many of the effect sizes included in the meta-analysis come from the same research report, with participants coming from the same pool of students at the same university. Another advantage of RVE is that it tends to perform better when the number of studies is relatively small (Tipton, 2015). This is particularly useful for the current analyses as several moderators were not well-represented in the dataset. In the hierarchical model, the RVE weight is defined as

$$w_{ij} = \frac{1}{V_{ij} + \tau^2 + \omega^2}, \text{ (Equation 5)}$$

where V_{ij} is the within-study variance for the *i*th effect size in the *j*th study. Hierarchical meta-regressions using RVE calculate two measures of variability: tau-squared (τ^2), the between-cluster variance component of the model; and omega-squared (ω^2), the within-cluster variance component.

All moderator analyses were conducted using the R packages robumeta and clubSandiwch (R Core Team, 2017; Fisher, Tipton & Zhipeng, 2017; Pustejovsky, 2020).

Results

Description of Dataset

In total, 57 independent effect sizes were collected from 30 research reports totaling 4,898 participants. The breakdown of effect sizes and independent samples per moderator can be found in Table 2.

Overall Effect

A forest plot showing the distribution of observed effect sizes and the summary effect size can be seen in Figure 3. Overall, the mean weighted effect size for retrieval-enhanced

suggestibility was 0.212 (*SE* = 0.05, *t*(23.9) = 3.91, $p < 0.001$, 95% CI[0.12, 0.30]), indicating a small but positive RES effect. In many RES studies, the number of participants was determined by researchers performing power analyses based on medium and large effect sizes — the result here suggests that those studies likely have been underpowered.

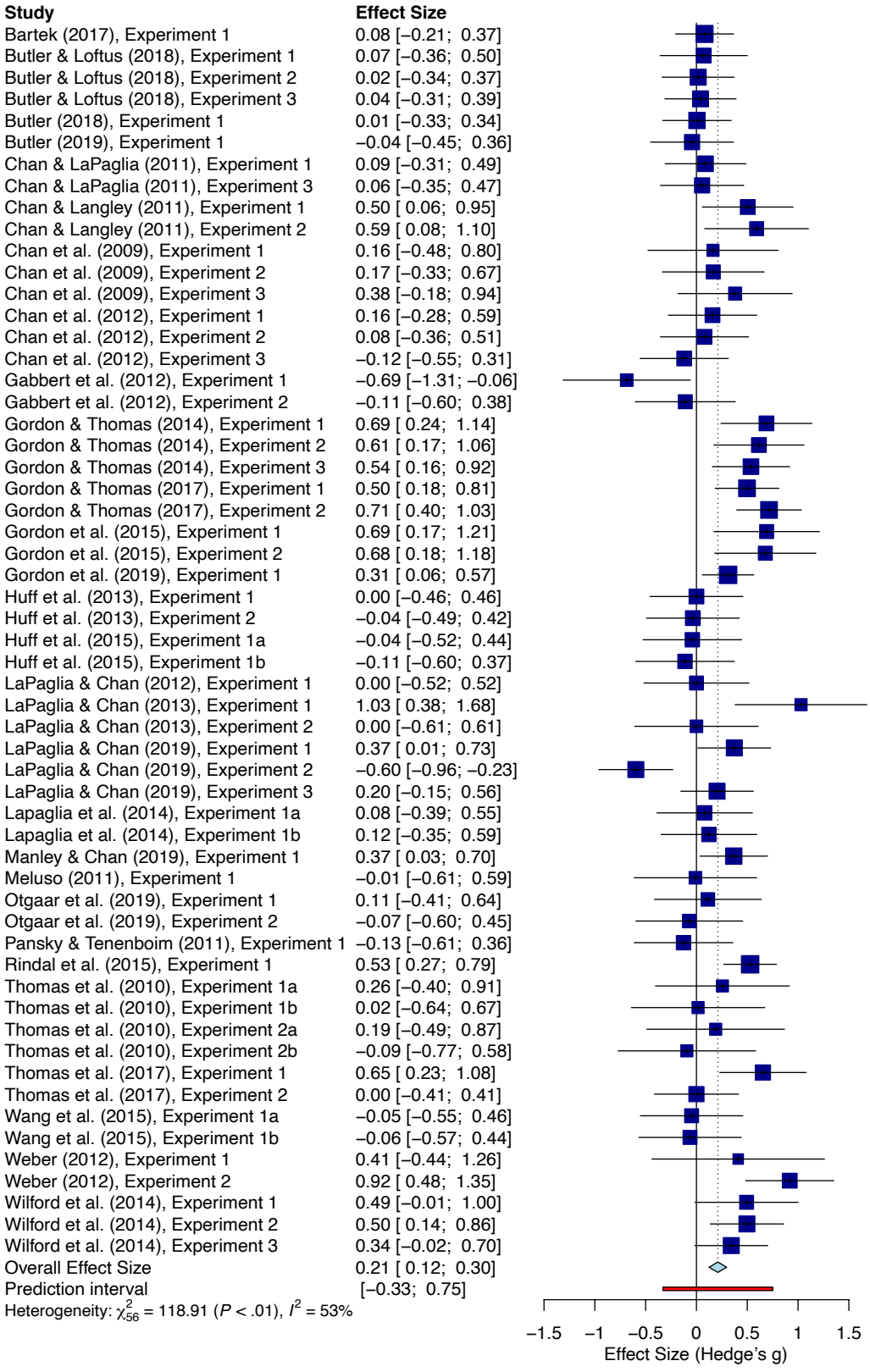
Moderator Analyses

Results of the moderator meta-regressions can be found in Tables 3 through 6. For each moderator, I performed two separate meta-regression analyses. The first meta-regression model is specified as the RES effect predicted by the moderator alone[2]. The second meta-regression model is specified as the RES effect predicted by the moderator plus *Research Group* and *Publication Status* — study level moderators — as covariates[3].

[2] I will be using the terms *moderator-only* model/meta-regression and *individual* model/meta-regression interchangeably.

[3] Note: *Research Group* was also included in a meta-regression model by itself as it was the first moderator meta-regression conducted.

Figure 3. Forest plot of effect sizes



I chose these two moderators as covariates due to their likely ability to explain part of the variability in the RES effect sizes. In meta-regression, the results with degrees of freedom below four should not be trusted because the Type I error rate increases dramatically (Tipton, 2015). Due to low degrees of freedom, I collapsed three moderators into dichotomous variables: *Stimulus Type* (Videos/Slideshows vs. Other), *Retrieval Practice Format* (Cued Recall vs. Other), and *Final Test Format* (Cued Recall vs. Other). The results for the moderator-only meta-regression analyses can be found in Table 3, and the results for the meta-regression analyses with covariates can be found in Table 4. For completeness, Table 5 contains the individual meta-regression analyses without collapsed moderators and Table 6 contains the covariate meta-regression analyses without collapsed moderators.

Report characteristics.

Research group. I hypothesized that research reports from Chan and colleagues would produce larger RES effects than reports from other researchers. Research group was a significant predictor of RES effect size. As hypothesized, studies conducted by Chan and colleagues produced larger effect sizes (g = 0.30, b = 0.24, SE = 0.10, t(20.2) = 2.48, p = 0.02, 95% CI[0.03,0.45]) than reports from other researchers (g = 0.06).

Publication status. I hypothesized that published research reports would produce larger RES effects than unpublished reports. This hypothesis was not supported, with publication status not a significant predictor of RES effect size in neither the individual model (b = -0.02, SE = 0.21 , t(3.59) = -0.09, p = 0.93 , 95% CI[-0.63,0.60]) nor or the covariate model (b = -0.22, SE = 0.21 , t(5.07) = -1.05, p = 0.34 , 95% CI[-0.76,0.32]).

Table 3. Moderator-only meta-regression results

Moderators	*F*	Coeff	*SE*	*t*	*df*	*p*	95% CI	*n*	*k*	ω^2	τ^2
Report Characteristics											
Research Group								57	29	0.010	0.04
Chan & colleagues		**0.24**	0.10	2.43	20.20	0.02	[0.03,0.45]				
Other researchers[a]		0.60	0.07	0.81	10.10	0.44	[-0.10,0.23]				
Publication Status								57	29	0.010	0.05
Published		-0.02	0.21	-0.09	3.59	0.93	[-0.63,0.60]				
Unpublished[a]		0.23	0.20	1.13	2.96	0.34	[-0.42,0.88]				
Encoding Variables											
Stimulus Type								57	29	0.008	0.05
Videos and slideshows		**0.09**	0.16	0.58	5.11	0.06	[-0.03,0.51]				
Other[a]		0.13	0.15	0.95	3.82	0.44	[-0.30,0.57]				
Stimulus Length		**0.01**	0.00	2.57	16.70	0.02	[0.00,0.01]	56	29	0.012	0.03
Misinformation Type	**13.90**				10.90	<0.001		57	29	0.000	0.05
Audio narrative		0.28	0.18	1.59	7.37	0.15	[-0.13,0.69]				
Misleading Questions		-0.14	0.16	-0.86	4.54	0.43	[-0.57,-0.29]				
Written narrative		0.19	0.20	0.91	8.52	0.39	[-0.28,0.65]				
Other[a]		0.01	0.16	0.09	4.54	0.94	[-0.42,0.45]				
Misinformation Length		0.05	0.03	1.89	5.99	0.11	[-0.01,0.11]	28	13	0.000	0.00
Storage Variables											
Retention Interval 1[b]		**-0.05**	0.02	-2.94	2.65	0.07	[-0.11,-0.01]	49	24	0.004	0.03
Retention Interval 2[c]		0.00	0.00	-2.68	1.62	0.14	[-0.00,0.00]	49	24	0.020	0.03
Retention Interval 3[d]		0.00	0.00	0.09	1.29	0.94	[-0.00,0.00]	42	22	0.003	0.06
Retrieval Variables											
Retrieval Practice Format								56	29	0.015	0.03
Cued Recall		**0.26**	0.10	2.52	13.66	0.02	[0.04,0.48]				
Other[a]		0.03	0.08	0.34	7.45	-0.16	[-0.16,0.22]				
Final Test Format								56	29	0.011	0.03
Cued Recall		**0.28**	0.07	3.78	8.66	<0.01	[0.11,0.49]				
Other[a]		<0.01	0.05	0.05	5.35	0.96	[-0.13,0.14]				
Test Match								56	29	51.190	0.05
Match		**0.25**	0.10	2.57	7.29	0.04	[0.02,0.47]				
No match[a]		0.00	0.08	0.29	4.69	0.79	[-0.18,0.23]				
Warning								56	29	0.008	0.05
Warning		-0.23	0.07	-3.34	1.15	0.16	[-0.86,0.41]				
No warning[a]		**0.23**	0.06	3.98	23.61	<.001	[0.11,0.35]				

Note. *n* - number of effect sizes; *k* - number of independent samples; I^2 - percentage of true heterogeneity to variance across the observed effect sizes; tau^2 - between-study variance in study-average effect sizes; [a] - reference group (intercept); [b] Delay between stimulus and retrieval practice; [c] Delay between retrieval practice and misinfomation; [d] Delay between misinforation and final test

Table 4. Moderator meta-regression results with covariates

Moderators	*F*	Coeff	*SE*	*t*	*df*	*p*	95% CI	*n*	*k*	ω^2	τ^2
Report Characteristics											
Research Group								57	29	0.012	0.04
Chan & colleagues		**0.24**	0.10	2.43	20.20	0.02	[0.03,0.45]				
Other researchers[a]		0.60	0.07	0.81	10.10	0.44	[-0.10,0.23]				
Publication Status								57	29	0.020	0.03
Published		-0.22	0.21	-1.05	5.07	0.34	[-0.76,0.32]				
Unpublished[a]		0.23	0.20	1.13	2.96	0.34	[-0.42,0.88]				
Encoding Variables											
Stimulus Type								56	28	0.014	0.03
Videos and slideshows		-0.03	0.13	-0.21	3.16	0.84	[-0.41,0.37]				
Other[a]		0.23	0.22	1.08	3.05	0.35	[-0.46,0.91]				
Stimulus Length		<.01	<.01	1.14	8.81	0.29	[0.00,0.01]	56	29	0.020	0.03
Misinformation Type	0.89				5.66	0.50		57	29	0.000	0.05
Audio narrative		0.13	0.26	0.50	7.69	0.63	[-0.47,-0.72]				
Misleading Questions		-0.05	0.14	-0.39	4.29	0.71	[-0.43,0.32]				
Written narrative		0.14	0.23	0.59	8.37	0.57	[-0.39,0.66]				
Other[a]		0.13	0.37	0.35	3.90	0.75	[-0.91,1.17]				
Misinformation Length		0.02	0.03	0.61	2.85	0.59	[-0.09,0.14]	28	13	0.000	0.00
Storage Variables											
Retention Interval 1[b]		-0.02	0.02	-1.18	2.92	0.33	[-0.09,-0.04]	49	24	0.019	0.02
Retention Interval 2[c]		<.01	<.01	-1.73	1.86	0.24	[-0.00,0.00]	49	24	0.030	0.02
Retention Interval 3[d]		<.01	<.01	0.47	1.41	0.70	[-0.00,0.00]	42	22	0.030	0.03
Retrieval Variables											
Retrieval Practice Format								56	29	0.015	0.03
Cued Recall		0.17	0.10	1.79	8.85	0.11	[-0.05,0.40]				
Other[a]		0.04	0.22	0.21	6.71	8.38	[-0.48,0.58]				
Final Test Format								56	29	0.011	0.03
Cued Recall		0.15	0.10	1.58	8.84	0.15	[-0.07,0.38]				
Other[a]		0.07	0.22	0.31	7.93	0.76	[-0.44,0.58]				
Format Match								56	29	0.012	0.05
Match		0.11	0.11	1.00	6.83	0.35	[-0.15,0.37]				
No match[a]		0.12	0.23	0.51	7.49	0.63	[-0.42,0.65]				
Warning								56	29	0.016	0.03
Warning		-0.11	0.19	-0.60	1.25	0.64	[-1.60,1.37]				
No warning[a]		0.23	0.20	1.13	2.96	0.34	[-0.42,0.87]				

Note. *n* - number of effect sizes; *k* - number of independent samples; I^2 - percentage of true heterogeneity to variance across the observed effect sizes; tau^2 - between-study variance in study-average effect sizes; [a] - reference group (intercept); [b] Delay between stimulus and retrieval practice; [c] Delay between retrieval practice and misinfomation; [d] Delay between misinforation and final test; **Research Group* was run as an individual model, *Publication Status* was run with *Research Group* as a covariate, and all subsequent models were run with both as covariates.

Encoding variables.

Stimulus type. I did not have a hypothesis for this moderator. In the individual model, stimulus type was not a significant predictor of the RES effect (g = 0.23, b = 0.09, SE = .16, t(5.11) = 0.58, p = 0.06, 95% CI[-0.03,0.51]). Studies using videos/slides (g = 0.23) and word lists (g = 0.80) produced larger RES effects than those using pictures (g = 0.01), although due to low degrees of freedom the results should not be trusted. In the model with covariates, stimulus type was not a significant moderator of the RES effect size (b = -0.03, SE = 0.13, t(3.16) = -0.21, p = 0.84, 95% CI[-0.41,0.37]).

Stimulus length. I hypothesized that longer exposure duration times would lead to smaller RES effects. Contrary to this hypothesis, longer stimuli led to larger RES effects in the individual model, although the effect was small (b = 0.01, SE < 0.01, t(16.70) = 2.57, p = 0.02, 95% CI[0.00,0.01]). In the covariate model, there was no statistically significant relationship *(*b < 0.01, SE < 0.01, t(8.81) = 0.29, p = 0.06, 95% CI[-0.03,0.51]) .

Misinformation type. I hypothesized that misinformation delivered via a narrative (either audio or written) would produce larger RES effects than misinformation delivered via misleading questions or other methods. In the individual model, misinformation type significantly moderated the size of the RES effect (F(10.90) = 13.90, p <0.001, ω^2 < 0.001, τ^2 = 0.05). However, unsupportive of my hypothesis, misinformation delivered via audio narratives (g = 0.30, b = 0.25, SE = 0.16, t(7.64) < 1.61, p = 0.15, 95% CI[-0.11,0.63]) did not result in significantly larger RES effects than written narratives *(*g = 0.19). In the covariate model controlling for research group and publication status, misinformation type did not significantly moderate the size of the RES effect (F(5.66) = 0.89, p = 0.50, ω^2 < 0.001, τ^2 = 0.05). Additionally, misinformation delivered via narrative (b = 0.14, SE = 0.21, t(8.80) = 0.66, p =

0.75, 95% CI[-0.34,0.62]) did not result in significantly larger RES effects than other types of misinformation.

Misinformation length. I hypothesized that longer exposure duration times would lead to larger RES effects. This hypothesis was not supported; misinformation length was not a significant moderator of the RES effect in the individual model (b = 0.05, SE = 0.03, t(5.99) = 1.89, p = 0.11, 95% CI[-0.11,0.11]) nor the covariate model (b = 0.02, SE = 0.03, t(2.85) = 0.61, p = 0.59, 95% CI[-0.09,0.14]).

Storage variables.

Retention interval 1. Delay between stimulus and retrieval practice. I hypothesized that longer retention intervals would be associated with larger RES effects. In the individual model, the moderator approached significance, but the low degrees of freedom indicate that the result should not be trusted (b = -0.05, SE = 0.02, t(2.65) = -2.94, p = 0.07, 95% CI[-0.11,0.01]). Retention interval did not appear to moderate the RES effect in the covariate model (b = -0.02, SE = 0.02, t(2.92) = -1.18, p = 0.33, 95% CI[-0.09,-0.04]).

Retention interval 2. Delay between stimulus and retrieval practice. I hypothesized that longer retention intervals would be associated with larger RES effects. This hypothesis was not supported, with the second retention interval length not significantly moderating the RES effect in neither the individual model (b = < 0.01, SE < 0.01, t(1.62) = -2.68, p = 0.14 , 95% CI[-0.01,0.00]) nor the covariate model (b = < 0.01, SE = < 0.01, t(1.86) = -1.73, p = 0.24 , 95% CI[-0.00,0.00]).

Retention interval 3: Delay between stimulus and retrieval practice. I hypothesized that longer retention intervals would be associated with larger RES effects. This hypothesis was not supported, with the third retention interval length not significantly moderating the RES effect in

neither the individual model (b = < 0.01, SE < 0.01, t(0.94) = 1.29, p = 0.94 , 95% CI[-0.00,0.00]) nor the covariate model (b = < 0.01, SE = < 0.01, t(1.41) = 0.47, p = 0.70 , 95% CI[-0.00,0.00]).

Retrieval variables.

Retrieval practice format. I hypothesized that more difficult retrieval attempts (i.e., cognitive interview, free recall) would lead to smaller RES effects. Due to low degrees of freedom, I collapsed the initial moderator categories into a dichotomous variable: cued recall vs. other. In the individual model, studies that employed cued recall as the retrieval practice format produced larger effect sizes (b = 0.26, SE = 0.10, t(13.67) = 2.52, p = 0.04, 95% CI[0.04, 0.48]) than other formats. Cued recall, however, was not a significant moderator in the covariate model (b = 0.28, p = 0.17).

Final test format. I hypothesized that more difficult retrieval attempts (i.e., cognitive interview, free recall) would lead to smaller RES effects. Due to low degrees of freedom, I collapsed the initial moderator categories into a dichotomous variable: cued recall vs. other. In the individual model, cued recall (g = 0.28, b = 0.28, SE = 0.07, t(8.66) = 3.78, p < 0.01, 95% CI[0.11,0.49]) resulted in larger RES effects than other retrieval practice formats such as free recall (g = -0.03) and recognition (g = 0.03). However, as was the case with retrieval practice format, final test format was not a significant moderator of the RES effect when controlling for research group and publication status (b = 0.07, SE = 0.22, t(8.84) = 1.58, p = 0.15, 95% CI[-0.07,0.38]).

Retrieval practice – final test format match. I hypothesized that a match in the format of the retrieval practice and the final test would lead to increased memory performance, which would thus result in smaller RES effects. Contrary to this hypothesis, test format matching

significantly moderated the size of the RES effect (g = 0.27, b = 0.25, SE = 0.07, t(7.29) = 2.57, p = 0.04, 95% CI[0.02,0.47]). When the retrieval practice and final test formats did not match, the RES effect was much smaller (g = 0.04). Test format match was not a significant moderator in the covariate model (b = 0.11, SE = 0.11, t(6.83) = 1.00, p = 0.35, 95% CI[-0.15,0.37]).

Warning. I hypothesized that studies in which participants are warned about the veracity of the misinformation would produce smaller RES effects than studies without a warning. Studies with participants were warned produced significantly smaller effect sizes (g = 0.01, b = 0.23, SE = 0.06, t(23.61) = 3.98, p < 0.001, 95% CI[0.11,0.35]) than studies without such a warning (g = 0.23). However, when controlling for covariates, warning was no longer a significant moderator (b = -0.11, SE = 0.19, t(1.25) = -0.60, p = 0.64, 95% CI[-1.60,1.37]).

Publication Bias

Whether a study is published is largely based on the statistical significance of the results (Dickerson, 2005). In meta-analysis, a clear bias can occur from combining effect sizes only from published studies, often leading to overestimation of the effectiveness of an intervention (Montori et al., 2000). In this section, I assess potential publication bias several different ways.

Fail-safe N.

Rosenthal's Fail-safe *N*, also called the "file drawer" analysis, is a statistical approach to estimate how many studies in the proverbial file drawer would need to be included in the analysis for the *p*-value of the meta-analysis to rise above 0.05 (Borenstein et al., 2009). Similarly, Orwin's Fail-safe *N* determines how many studies would need to be included in order to reduce the effect size to a level of "substantive importance", in the case of the current analysis, that target *p*-value is 0.05. Rosenthal's Fail-safe *N* was 1,040, suggesting that over 1,000 studies with a mean effect size of zero would need to be added in order for the summary effect size

found in this meta-analysis to become non-significant. Orwin's Fail-safe *N* was 57, suggesting that 57 studies with a mean effect size of zero would need to be added in order for the *p* value for the summary statistic to rise above 0.1015. I ran a meta-regression with *publication status* as a predictor which revealed that publication status was not a significant predictor of the RES effect in neither the individual model (*b* = -0.02, SE = 0.21 , *t*(3.59) = -0.09, *p* = 0.93 , 95% CI[-0.63,0.60]) nor the covariate model (*b* = -0.22, SE = 0.21 , t(5.07) = -1.05, *p* = 0.34 , 95% CI[-0.76,0.32]).

Funnel plots.

Funnel plots are the most common way to assess publication bias visually. Study effect sizes are plotted on the x-axis with their corresponding standard errors on the y-axis. If there is no publication bias, effect sizes from individual studies should be distributed symmetrically around the summary effect size roughly in the shape of a funnel (Borenstein et al., 2009). If the plot is asymmetrical, for example if studies with large standard errors seem to be systematically missing on one side of the summary effect, publication bias might be present. The funnel plot for the RES effect can be seen in Figure 4. As the direction of the effect is to the right, the is a visible gap on the bottom left portion of the plot which is likely where nonsignificant studies would reside if they were included in the analysis. Additionally, I performed Egger's regression test to evaluate funnel plot asymmetry (Sterne and Egger, 2005). This test did not reach statistical significance, indicating that there is not significant asymmetry in effect sizes (*z* = -1.02, *p* = 0.31).

Figure 4. Funnel plot

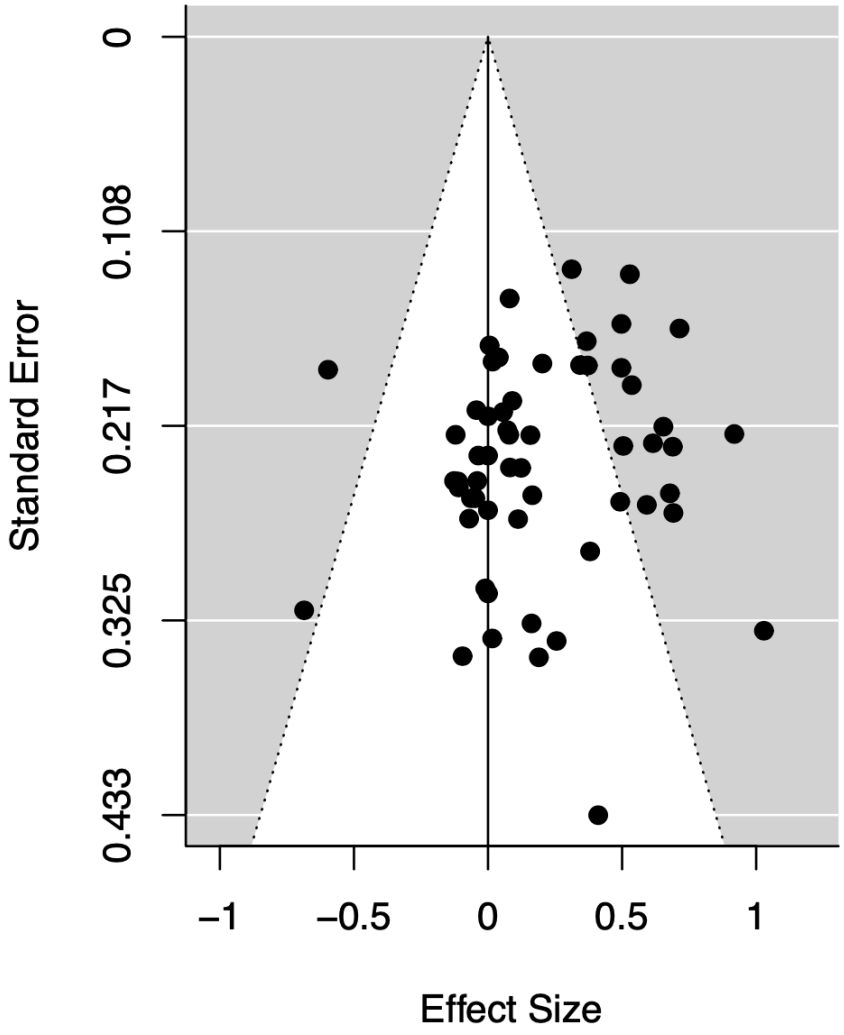


Table 5. Moderator-only meta-regression results with un-collapsed moderators

Moderators	*F*	Coeff	*SE*	*t*	*df*	*p*	95% CI	*n*	*k*	ω^2	τ^2
Report Characteristics											
Research Group								57	29	0.010	0.04
Chan & colleagues		**0.24**	0.10	2.43	20.20	0.02	[0.03,0.45]				
Other researchers[a]		0.60	0.07	0.81	10.10	0.44	[-0.10,0.23]				
Publication Status								57	29	0.010	0.05
Published		-0.02	0.21	-0.09	3.59	0.93	[-0.63,0.60]				
Unpublished[a]		0.23	0.20	1.13	2.96	0.34	[-0.42,0.88]				
Encoding Variables											
Stimulus Type	29.80				1.87	0.04		56	28	0.005	0.04
Videos and slideshows		0.24	0.07	3.32	2.41	0.06	[-0.03,0.50]				
Factual Knowledge		-0.03	0.04	-0.79	1.97	0.52	[-0.21,0.15]				
Word lists/pairs		0.78	0.08	9.54	2.88	<0.01	[0.52,1.05]				
Pictures[a]		-0.01	0.04	-0.27	1.97	0.81	[-0.19,0.17]				
Stimulus Length		**0.01**	0.00	2.57	16.70	0.02	[0.00,0.01]	56	29	0.012	0.03
Misinformation Type	**13.90**				10.90	<0.001		57	29	0.000	0.05
Audio narrative		0.28	0.18	1.59	7.37	0.15	[-0.13,0.69]				
Misleading Questions		-0.14	0.16	-0.86	4.54	0.43	[-0.57,-0.29]				
Written narrative		0.19	0.20	0.91	8.52	0.39	[-0.28,0.65]				
Other[a]		0.01	0.16	0.09	4.54	0.94	[-0.42,0.45]				
Misinformation Length		0.05	0.03	1.89	5.99	0.11	[-0.01,0.11]	28	13	0.000	0.00
Storage Variables											
Retention Interval 1[b]		-0.05	0.02	-2.94	2.65	0.07	[-0.11,-0.01]	49	24	0.004	0.03
Retention Interval 2[c]		0.00	0.00	-2.68	1.62	0.14	[-0.00,0.00]	49	24	0.020	0.03
Retention Interval 3[d]		0.00	0.00	0.09	1.29	0.94	[-0.00,0.00]	42	22	0.003	0.06
Retrieval Variables											
Retrieval Practice Format	1.83				1.83	0.37		56	29	0.009	0.04
Cognitive Interview		-0.03	0.16	-0.19	1.85	0.87	[-0.79,0.73]				
Cued Recall		0.35	0.09	3.86	1.09	0.15	[-0.61,1.32]				
Free Recall		0.18	0.14	1.22	1.47	0.38	[-0.72,1.07]				
Recognition		0.11	0.06	1.78	1.00	0.33	[-0.69,0.92]				
Other[a]		-0.07	0.06	-1.04	1.00	0.49	[-0.87,0.74]				
Final Test Format	12.00				3.66	0.02		56	29	0.012	0.04
Cued Recall		**0.28**	0.06	4.48	19.31	<0.001	[0.15,0.41]				
Free Recall		-0.04	0.12	-0.30	2.56	0.79	[-0.47,0.40]				
Recognition		0.04	0.01	5.70	1.67	0.04	[0.00,0.07]				
Other[a]		0.00	0.00	0.00	16.95	1.00	[-0.00,0.00]				
Test Match								56	29	51.190	0.05
Match		**0.25**	0.10	2.57	7.29	0.04	[0.02,0.47]				
No match[a]		0.00	0.08	0.29	4.69	0.79	[-0.18,0.23]				
Warning								56	29	0.008	0.05
Warning		-0.23	0.07	-3.34	1.15	0.16	[-0.86,0.41]				
No warning[a]		**0.23**	0.06	3.98	23.61	<.001	[0.11,0.35]				

Note. *n* - number of effect sizes; *k* - number of independent samples; I^2 - percentage of true heterogeneity to variance across the observed effect sizes; tau^2 - between-study variance in study-average effect sizes; [a] - reference group (intercept); [b] Delay between stimulus and retrieval practice; [c] Delay between retrieval practice and misinfomation; [d] Delay between misinforation and final test

Table 6. Meta-regression moderator results with covariates and un-collapsed moderators

Moderators	*F*	Coeff	*SE*	*t*	*df*	*p*	95% CI	*n*	*k*	ω^2	τ^2
Report Characteristics											
Research Group								57	29	0.012	0.04
Chan & colleagues		**0.24**	0.10	2.43	20.20	0.02	[0.03,0.45]				
Other researchers[a]		0.60	0.07	0.81	10.10	0.44	[-0.10,0.23]				
Publication Status								57	29	0.020	0.03
Published		-0.22	0.21	-1.05	5.07	0.34	[-0.76,0.32]				
Unpublished[a]		0.23	0.20	1.13	2.96	0.34	[-0.42,0.88]				
Encoding Variables											
Stimulus Type	13.70				2.33	0.05		56	28	0.014	0.03
Videos and slideshows		0.04	0.08	0.48	2.74	0.67	[-0.24,0.32]				
Factual Knowledge		-0.07	0.07	-0.09	1.24	0.50	[-0.68,0.54]				
Word lists/pairs		0.75	0.10	7.24	1.63	0.03	[0.19,1.31]				
Pictures[a]		0.03	0.08	0.35	1.24	0.78	[-0.58,0.63]				
Stimulus Length		<.01	<.01	1.14	8.81	0.29	[0.00,0.01]	56	29	0.020	0.03
Misinformation Type	0.89				5.66	0.50		57	29	0.000	0.05
Audio narrative		0.13	0.26	0.50	7.69	0.63	[-0.47,-0.72]				
Misleading Questions		-0.05	0.14	-0.39	4.29	0.71	[-0.43,0.32]				
Written narrative		0.14	0.23	0.59	8.37	0.57	[-0.39,0.66]				
Other[a]		0.13	0.37	0.35	3.90	0.75	[-0.91,1.17]				
Misinformation Length		0.02	0.03	0.61	2.85	0.59	[-0.09,0.14]	28	13	0.000	0.00
Storage Variables											
Retention Interval 1[b]		-0.02	0.02	-1.18	2.92	0.33	[-0.09,-0.04]	49	24	0.019	0.02
Retention Interval 2[c]		<.01	<.01	-1.73	1.86	0.24	[-0.00,0.00]	49	24	0.030	0.02
Retention Interval 3[d]		<.01	<.01	0.47	1.41	0.70	[-0.00,0.00]	42	22	0.030	0.03
Retrieval Variables											
Retrieval Practice Format	2.03				2.59	0.32		56	29	0.013	0.03
Cognitive Interview		0.03	0.15	0.22	1.88	0.85	[-0.63,0.70]				
Cued Recall		0.28	0.09	3.09	1.15	0.17	[-0.59,1.16]				
Free Recall		0.19	0.10	1.96	1.46	0.23	[-0.42,0.80]				
Recognition		-0.01	0.08	-0.17	1.11	0.89	[-0.83,0.80]				
Other[a]		-0.06	0.22	-0.27	1.90	0.82	[-1.06,0.94]				
Final Test Format	**5.05**				5.82	0.05		56	29	0.011	0.04
Cued Recall		**0.33**	0.07	4.43	12.86	<.001	[0.17,0.49]				
Free Recall		0.16	0.14	1.13	4.33	0.32	[-0.21,0.53]				
Recognition		0.24	0.13	1.77	3.77	0.16	[-0.14,0.62]				
Other[a]		-0.10	0.21	-0.48	3.97	0.66	[-0.70,0.49]				
Format Match								56	29	0.012	0.05
Match		0.11	0.11	1.00	6.83	0.35	[-0.15,0.37]				
No match[a]		0.12	0.23	0.51	7.49	0.63	[-0.42,0.65]				
Warning								56	29	0.016	0.03
Warning		-0.11	0.19	-0.60	1.25	0.64	[-1.60,1.37]				
No warning[a]		0.23	0.20	1.13	2.96	0.34	[-0.42,0.87]				

Note. *n* - number of effect sizes; *k* - number of independent samples; I^2 - percentage of true heterogeneity to variance across the observed effect sizes; tau^2 - between-study variance in study-average effect sizes; [a] - reference group (intercept); [b] Delay between stimulus and retrieval practice; [c] Delay between retrieval practice and misinfomation; [d] Delay between misinforation and final test; **Research Group* was run as an individual model, *Publication Status* was run with *Research Group* as a covariate, and all subsequent models were run with both as covariates.

Chapter 4: Summary and Discussion

The aim of the present meta-analysis is to provide an empirical review of retrieval-enhanced suggestibility through meta-analysis. Specifically, the objectives of this meta-analysis were to identify: (a) the overall size of the retrieval-enhanced suggestibility effect; (b) the methodological factors that moderate the size of the effect; (c) the boundary conditions of the effect; and (d) theoretical mechanisms underlying the effect. In total, 57 independent effect sizes were collected from 30 different research reports totaling 4,898 participants. The breakdown of effect sizes and independent samples per moderator can be found in Table 2. A forest plot showing the total distribution of effect sizes and overall effect can be seen in Figure 3. A summary of the moderators, hypotheses, and meta-regression results can be found in **Error! Reference source not found.**. In the next section, I discuss the overall RES effect across studies. Then I discuss the results for each moderator in relation to their respective hypotheses, potential theoretical mechanisms supporting their results, and areas for future research in relation to each moderator. I will then discuss limitations of the present meta-analysis and concluding remarks.

Overall effect

The main objective of this meta-analysis was to quantify the overall size of the RES effect. In the present meta-analysis, the mean weighted effect size for retrieval-enhanced suggestibility was $g = 0.212$, indicating a small positive RES effect (Figure 3). Much of the early RES research appeared to have overestimated the size of the RES effect, perhaps leading researchers to recruit relatively few participants in subsequent studies. For example, 36 total undergraduate students participated in Chan et al. (2009) Experiment 1A and only 18 in the retrieval-practice condition. Similarly, LaPaglia & Chan (2013) recruited 20 participants per-

condition in their experiment. The results of the overall effect analysis should lead researchers to use more conservative effect size estimates when conducting power analyses which will hopefully lead to more studies using larger sample sizes, which will lead to more-precise estimates of the RES effect.

Moderators

A summary of the moderators, hypotheses, and meta-regression results can be found in **Error! Reference source not found.**. Detailed breakdowns of moderator meta-regression results can be found in Table 3 and Table 4. When looking at the moderator-only meta-regression results collectively, seven out of the eleven moderators significantly moderated the RES effect. However, when controlling for research group (Chan and colleagues vs. other) and publication status, none of the moderators were significantly associated with the RES effect.

Report characteristics

Research group.

Consistent with my hypothesis, studies conducted by Chan and colleagues produced larger effect sizes than studies conducted by other researchers. The result is consistent with research findings in my lab and trends in the literature. Butler (2017) unsuccessfully attempted to replicate the findings of Chan et al. (2009) using the same materials and procedures. We subsequently conducted three experiments, of which only one showed an RES effect (Butler & Loftus, 2018). In addition, in the RES literature, Chan and his research group consistently produced research reports with the largest effect sizes.

Table 7. Summary of moderators, hypotheses, and meta-regression results

Moderators	Hypothesis	Moderator-only meta-regression results	Covariate meta-regression results
Report Characteristics			
Research Group	Research reports from Chan & colleagues will produce larger RES effects than reports from other researchers	**Significant moderator; Chang & colleagues research reports produced larger effect sizes**	-
Publication Status	Published research reports will have larger RES effects than unpublished reports	Not a significant moderator	Not a significant moderator
Encoding Variables			
Stimulus Type	No prediction	**Significant moderator; videos and slide shows associated with larger effect sizes**	Not a significant moderator
Stimulus Length	Longer exposure duration times will lead to smaller RES effects	**Significant moderator; longer stimuli associated with larger RES effects**	Not a significant moderator
Misinformation Type	Misinformation delivered via a narrative (either audio or written) will produce larger RES effects than misinformation delivered via misleading questions or other methods	**Significant moderator**	Not a significant moderator
Misinformation Length	Longer exposure duration times will lead to larger RES effects	Not a significant moderator	Not a significant moderator
Storage Variables			
Retention Interval 1[b]	Longer retention intervals will lead to larger RES effects.	**Significant moderator; longer retention interval associated with larger RES effect**	Not a significant moderator
Retention Interval 2[c]	Longer retention intervals will lead to larger RES effects.	Not a significant moderator	Not a significant moderator
Retention Interval 3[d]	Longer retention intervals will lead to larger RES effects.	Not a significant moderator	Not a significant moderator
Retrieval Variables			
Retrieval Practice Format	More difficult retrieval attempts (i.e., cognitive interview, free recall) will lead to smaller RES effects	**Significant moderator; cued recall associated with larger RES effect**	Not a significant moderator
Final Test Format	More difficult retrieval attempts (i.e., cognitive interview, free recall) will lead to smaller RES effects	**Significant moderator; cued recall associated with larger RES effect**	Not a significant moderator
Test Match	A match in the format of the retrieval practice and the final test will lead to increased memory performance, which will result in smaller RES effects	**Significant moderator; test format match associated with larger RES effect**	Not a significant moderator
Warning	Studies in which participants are warned about the veracity of the misinformation will produce smaller RES effects than studies without a warning	Not a significant moderator	Not a significant moderator

When looking at the moderator-only meta-regressions as a whole, seven out of the eleven moderators significantly moderated the RES effect. However, when controlling for research group and publication status, none of the moderators significantly moderated the RES effect. Although I hypothesized that research group and publication status moderators would account for some of the variation in effect sizes, it was surprising to find them to have such a strong impact. It is not clear why Chan and colleagues typically produce reports with larger RES effects than most other researchers. The results of the present meta-analysis cannot answer this question, but I can offer some speculation.

There may be certain methodological procedures that their labs use that they are not communicating in their reports. As mentioned above, I performed a direct replication ostensibly using the same materials that Chan used — he sent them to me personally — and failed to find an RES effect. I considered regional differences in participant populations but have ruled that reason out due to the Chan & colleagues research groups being spread across the United States. It is possible that there is some form of confirmation bias occurring which has led the group to seek or interpret evidence of RES in a biased fashion (Nickerson, 1998). Chan was the first researcher to find RES, and he and his colleagues have produced the majority of the RES publications to date (36 for Chan and colleagues, 21 for other researchers). To be clear, I have absolutely no reason to suspect the Chan research group of nefarious practices, but rather highlight a potential explanation to a puzzling finding. This bias can affect decisions made in the research process, such as which variables to analyze and when to stop data collection. These "researcher degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011) could contribute to the stark contrast in findings between Chan's research group and the other researchers. Research has

shown that blinded interpretation of study results can diminish interpretation bias (Javinen et al., 2014).

Publication status.

I hypothesized that published research reports will have larger RES effects than unpublished reports. The results of the meta-regressions did not support this hypothesis. As noted earlier in the publication bias section, RES research reports do not appear to be significantly more likely to be published if they are confirmatory.

Encoding variables

Stimulus type.

I did not have a prediction for this moderator. In the individual model, stimulus type significantly moderated the RES effect. Specifically, when the stimulus was a video or slideshow, the RES effect was larger. This finding makes sense in light of previous research showing that memory for dynamic stimuli can be better than memory for static stimuli (Goldstein et al., 1982).

Stimulus length.

I hypothesized that longer exposure duration times would lead to smaller RES effects. Contrary to this hypothesis, longer stimuli led to a statistically significant increase in the RES effect in the individual model, although it was a small increase. In hindsight, this aligns with previous research on memory decay (e.g. Deffenbacher et al., 2008). As the length of the stimulus increases, more time has elapsed between the encoding of each key detail in the stimulus and the subsequent misinformation and final test.

Misinformation type.

I hypothesized that misinformation delivered via a narrative (either audio or written) would produce larger RES effects than misinformation delivered via misleading questions or other methods because when misinformation is delivered via a narrative, it is possible that participants view the information as corrective feedback and thus update their memories. Misinformation type overall was a significant moderator of the RES effect, but my specific hypothesis was not supported. In the classroom context, research has shown that feedback given verbally or written can have differing effects. For example, Merry and Orsmond (2008) suggested that students may perceive audio feedback in a more meaningful way because speech seems more genuine and is received in a more personal way. Applying this idea to the RES context, subjects receiving misinformation via an audio narrative might be more likely to accept false information than subjects receiving misinformation via written narratives. To my knowledge, there are no studies that directly manipulate the method of misinformation delivery; a direct manipulation of misinformation type would provide a more-direct comparison.

Misinformation length.

I hypothesized that longer exposure duration times would lead to larger RES effects. This hypothesis was not supported by the results. A potential explanation for this is that there simply wasn't enough variability in misinformation length ($M = 6.41$, $SD = 1.54$) between studies.

Storage variables

Retention intervals.

There are three important retention intervals in RES research: the delay between the stimulus and the initial test; the delay between the initial test and the misinformation; and the delay between the misinformation and the final test. For all three retention intervals, I

hypothesized that longer retention intervals would be associated with larger RES effects due to the tendency for memories to fade and become weaker over time (Ebbinghaus,1964). The only retention interval that was significantly associated with the RES effect was the first, namely the delay between the initial test and the misinformation. This result seems to imply that as time passes and memory for the individual event fades, participants use the misinformation as corrective feedback or a second chance to encode the stimulus and update their memories accordingly.

Retrieval variables

Retrieval practice and final test format.

For both of these moderators, I hypothesized that more difficult retrieval attempts (i.e., cognitive interview, free recall) would lead to smaller RES effects while cued recall and recognition would lead to larger RES effects. According to the effortful retrieval hypothesis, more difficult retrieval attempts lead to better retention of the tested material (Jacoby, 1978). The moderator-only meta-regressions support this hypothesis, where cued-recall retrieval practice and final tests were both associated with larger RES effects compared to other formats.

Test match.

According to the theory of transfer-appropriate processing, the strength of the testing effect is positively associated with the similarity between retrieval practice and final test (Rowland, 2014). Thus, I hypothesized that a match in the format of the retrieval practice and the final test would lead to increased memory performance, which would thus result in smaller RES effects. Contrary to this hypothesis, matching formats was associated with *larger* RES effects.

Despite research showing the memorial benefits of matched testing, some research suggests there is no benefit at all (e.g., Carpenter & DeLosh, 2006).

Warning.

I hypothesized that studies in which participants are warned about the veracity of the misinformation would produce smaller RES effects than studies without a warning. Although the hypothesis was not supported by a statistically significant result, the data were in the right direction: studies with warnings showed smaller RES effects in both the moderator-only model (g = -.23) and the covariate model (g = -0.11).

Limitations

One limitation of the present meta-analysis is the small number of independent effect sizes in the literature for subgroups of certain moderators, particularly *stimulus type* and *retrieval practice/final test format*. More effect sizes in these categories would allow for more fine-tuned analyses form which more specific conclusions could be drawn. As more RES studies continue to be conducted, a follow-up meta-analysis sometime would provide more precise effect size estimates.

Another limitation of the current meta-analysis is the small number of unpublished research reports included in the analysis. Although I discussed three different ways to assess publication bias (Fail-safe *N*, visual funnel plot inspection, and Egger's regression for funnel plot asymmetry), the best way account for the file drawer problem is to have an accurate number of unpublished reports in the analysis. Several researchers I contacted did have unpublished data, but they could not share the findings as they were being prepared for journal submission.

Future directions

The present meta-analysis is an important first step in trying to make sense of retrieval-enhanced suggestibility. Until now, research was being conducted based on the results of individual studies with varying methodologies, results, and sample sizes. Moving forward, I see two clear ways for RES researchers to improve our understanding of the RES effect and improve the quality of research on it.

The first area of growth revolves around researchers designing experiments that can answer research questions more precisely, instead of trying to draw conclusions across experiments. For example, Butler and Loftus (2018) used two different discrepancy detection measures in Experiments 2 and 3. Instead of comparing the results across those two experiments, a beneficial alternative could have been to have presented the two different measures to different intervention groups in the same experiment and directly measured the difference. This applies to most of the moderators included in the present meta-analysis. Researchers can manipulate multiple moderators within experiments conducted at the same time in order to test the effects of multiple manipulations on the RES effect. Of course, this requires more resources (e.g., time, funding for more participants, research assistants), but it is necessary for the field to draw more precise conclusions.

The second suggestion for future research is in regard to the *Open Science* movement, which aims to improve science by testing the replicability and reproducibility of research findings (Crüwell et al., 2018). Researchers are encouraged to pre-register their studies before conducting them. Generally, this entails explicitly stating your research questions, hypotheses, data source(s), methodology, and analysis plan, along with time-stamping them in an online database. Additionally, as the name implies, a key tenant of open science is open access to all parts of the study including its materials, data, code used to analyze the data, and the actual

research report itself. The hope is that with increased transparency, researcher degrees of freedom in scientific research are reduced and science benefits.

Conclusion

After conducting the present meta-analysis, I still find retrieval-enhanced suggestibility to be an intriguing finding with several areas that need further exploration. With the meta-analytic findings, my hope is that researchers approach these areas of exploration with transparency and precision such that the conclusions drawn are generalizable and reproducible.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research*, 87(3), 659–701. https://doi.org/10.3102/0034654316689306

Belli, R. F. (1989). Influences of misleading postevent information: Misinformation interference and acceptance. *Journal of Experimental Psychology: General*, 118(1), 72-85.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The Mismeasure of Memory : When Retrieval Fluency Is Misleading as a Metamnemonic Index. *Journal of Experimental Psychology: General*, 127(1), 55–68.

Besnard, A., Caboche, J., & Laroche, S. (2012). Reconsolidation of memory: a decade of debate. *Progress in Neurobiology*, 99(1), 61-80.

Blank, H., & Launay, C. (2014). How to protect eyewitness memory against the misinformation effect: A meta-analysis of post-warning studies. *Journal of Applied Research in Memory and Cognition*, 3(2), 77-88.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). Introduction to meta-analysis. John Wiley & Sons.

Brewer, G. A., Marsh, R. L., Meeks, J. T., Clark-Foos, A., & Hicks, J. L. (2010). The effects of free recall testing on subsequent source memory. *Memory (Hove, England)*, *18*(4), 385–393.

Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, 13(4), 273-281.

Butler, B. J., & Loftus, E. F. (2018). Discrepancy detection in the retrieval-enhanced suggestibility paradigm. *Memory*, 26(4), 483-492.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal ofExperimental Psychology: Learning, Memory, and Cognition*, 35, 1563–1569. doi:10.1037/ a0017021

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, 34, 268–276. doi:10.3758/BF03193405

Chan, J. C. K., & Langley, M. M. (2011). Paradoxical effects of testing: Retrieval enhances both accurate recall and suggestibility in eyewit- nesses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 248–255.

Chan, J. C., Manley, K. D., & Lang, K. (2017). Retrieval-Enhanced Suggestibility: A Retrospective and a New Investigation. *Journal of Applied Research in Memory and Cognition*, 6(3), 213-229.

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: a dual process account. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(2), 431–437.

Chan, J. C., McDermott, K. B., & Roediger III, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, 135(4), 553-571.

Chan, J. C. K., Thomas, A. K., & Bulevich, J. B. (2009). Recalling a Witnessed Event Increases Eyewitness Suggestibility. *Psychological Science*, 1–8.

Chan, J. C. K., Wilford, M. M., & Hughes, K. L. (2012). Retrieval can increase or decrease suggestibility depending on how memory is tested: The importance of source complexity. *Journal of Memory and Language*, 67, 78–85.

Cochran, K. J., Greenspan, R. L., Bogart, D. F., & Loftus, E. F. (2016). Memory blindness: Altered memory reports lead to distortion in eyewitness memory. *Memory & Cognition*, 717–726.

Coe, R. (2002). It's the effect size, stupid: What effect size is and why it is important.

Cuddy, L. J., & Jacoby, L. L. (1982). When forgetting helps memory: An analysis of repetition effects. *Journal of Verbal Learning and Verbal Behavior*, 21(4), 451-467.

Crüwell, S., Doorn, J. van, Etz, A., Makel, M., Moshontz, H., Niebaum, J., … Schulte-Mecklenbeck, M. (2018). 8 Easy Steps to Open Science: An Annotated Reading List. PsyArXiv Preprints. http://doi.org/10.31234/OSF.IO/CFZYX

Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: a meta-analysis. *Educational Psychology Review*, 28(3), 425-474.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. *Publication bias in meta-analysis: Prevention, assessment and adjustments*, 11-33.

Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. *Contemporary Educational Psychology*, 6(3), 217-226.

Dudai, Y. (2004). The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology*, 55(1), 51-86.

Ferguson, C.J., & Brannick, M.T. (2012). Publication bias in psychological science: Prevalence, method for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*(1), 120-128.

Fisher, Tipton and Zhipeng (2017). robumeta: Robust Variance Meta-Regression. R package version 2.0. https://CRAN.R-project.org/package=robumeta

Flavell, J. H. (1992). Cognitive development: Past, present, and future. *Developmental Psychology*, 28(6), 998.

Gabbert, F., Hope, L., Fisher, R. P., & Jamieson, K. (2012). Protecting against misleading post-event information with a self-administered interview. *Applied Cognitive Psychology*, 26, 568–575.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, 6(40).

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition*, 7(2), 95-112.

Goldstein, A. G., Chance, J. E., Hoisington, M., & Buescher, K. (1982). Recognition memory for pictures: Dynamic vs. static stimuli. *Bulletin of the Psychonomic Society*, 20(1), 37-40.

Gordon, L. T., & Thomas, A. K. (2014). Testing potentiates new learning in the misinformation paradigm. *Memory & Cognition*, 42, 186–197. http://dx.doi.org/10.3758/s13421-013-0361-2

Gordon, L. T., & Thomas, A. K. (2017). The forward effects of testing on eyewitness memory: The tension between suggestibility and learning. *Journal of Memory and Language*, 95, 190–199. https://doi.org/10.1016/j.jml.2017.04.004

Gordon, L. T., Thomas, A. K., & Bulevich, J. B. (2015). Looking for answers in all the wrong places: How testing facilitates learning of misinformation. *Journal of Memory and Language*, *83*, 140–151. http://doi.org/10.1016/j.jml.2015.03.007

Greene, E., Flynn, M. S., & Loftus, E. F. (1982). Inducing resistance to misleading information. *Journal of Verbal Learning and Verbal Behavior*, *21*(2), 207-219.

Glover, J. A., (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, 10(5), 562-567.

Holliday, B. E. (2003). The effect of a prior cognitive interview on children's acceptance of misinformation. *Applied Cognitive Psychology*, 17, 443–457.

Huff, M. J., Weinsheimer, C. C., & Bodner, G. E. (2016). Reducing the misinformation effect through initial testing: Take two tests and recall me in the morning? *Applied cognitive Psychology*, 30(1), 61-69.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal ofVerbal Learning and Verbal Behavior*, 17, 649–667. doi:10.1016/S0022-5371(78)90393-6

Jang, Y., & Huber, D. E. (2008). Context retrieval and context change in free recall: recalling from long-term memory drives list isolation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(1), 112–127.

Järvinen, T. L., Sihvonen, R., Bhandari, M., Sprague, S., Malmivaara, A., Paavola, M., ... & Guyatt, G. H. (2014). Blinded interpretation of study results can feasibly and effectively diminish interpretation bias. *Journal of Clinical Epidemiology*, 67(7), 769-772. Chicago

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, 101, 621–629. doi: 10.1037/a0015183

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, 114(1), 3-28.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, 319(5865), 966-968.

Kulik, J. A., & Kulik, C. L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, 58(1), 79-97.

LaPaglia, J. A., & Chan, J. C. (2012). Retrieval does not always enhance suggestibility: Testing can improve witness identification performance. *Law and human behavior*, 36(6), 478-487.

LaPaglia, J. A., & Chan, J. C. K. (2013). Testing increases suggestibility for narrative-based misinformation but reduces suggestibility for question-based misinformation. *Behavioral Sciences and the Law*, 31(5), 593-60.

LaPaglia, J. A., Wilford, M. M., Rivard, J. R., Chan, J. C. K., & Fisher, R. P. (2014). Misleading suggestions can alter later memory reports even following a Cognitive Interview. *Applied Cognitive Psychology*, 28, 1–9.

Lindsay, D. S., Johnson, M. K., & Kwon, P. (1991). Developmental changes in memory source monitoring. *Journal of Experimental Child Psychology*, 52(3), 297-318.

Loftus E. F. (1979). Reactions to blatantly contradictory information. *Memory & Cognition*, 7, 368–374.

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, 12(4), 361-366.

Mafinejad, M. K., Aghili, R., Emami, Z., Malek, M., Baradaran, H., Taghavinia, M., & Khamseh, M. E. (2014). Study guides: effective tools to improve self-directed learning skills of medical students. *Acta Medica Iranica*, 781-785.

McCloskey, M., & Zaragoza, M. (1985). Misleading postevent information and memory for events: Arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General*, 114(1), 1.

McGaugh, J. L. (1966). Time-dependent processes in memory storage. *Science*, 153(3742), 1351-1358.

Memon, A., Zaragoza, M., Clifford, B. R., & Kidd, L. (2010). Inoculation or antidote? The effects of cognitive interview timing on false memory for forcibly fabricated events. *Law and Human Behavior*, 34, 105–117.

Stephen Merry & Paul Orsmond (2008) Students' Attitudes to and Usage of Academic Feedback Provided Via Audio Files, *Bioscience Education*, 11(1), 1-11, DOI: 10.3108/ beej.11.3

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, 6, e1000097. doi:10.1371/journal.pmed.1000097.

Mondadori, C., & Ducret, T. (1991). How long does 'memory consolidation take? New compounds can improve retention performance, even if administered up to 24 hours after the learning experience. *Brain research*, 555(1), 107-111.

Montori, V. M., Smieja, M., & Guyatt, G. H. (2000, December). Publication bias: a brief review for clinicians. In Mayo Clinic Proceedings (Vol. 75, No. 12, pp. 1284-1288). *Elsevier*. Chicago

Murphy, G., & Greene, C. M. (2016). Perceptual load induces inattentional blindness in drivers. *Applied Cognitive Psychology*, 30(3), 479-483.

Nader, K., & Hardt, O. (2009). A single standard for memory: the case form reconsolidation. *Nature Reviews Neuroscience*, 10(3), 224.

Nader, K., Schafe, G. E., & Le Doux, J. E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, 406(6797), 722.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review*

of General Psychology, 2(2), 175-220.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology*, 74(1), 18.

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature News*, 506(7487), 150.

Okado, Y., & Stark, C. E. L. (2005). Neural activity during encoding predicts false memories created by misinformation. *Learning & Memory*, 12(1), 3–11.

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, 19(1), 55.

Pansky, A., & Tenenboim, E. (2011). Inoculating against eyewitness suggestibility via interpolated verbatim vs. gist testing. *Memory & Cognition*, 39, 155–170.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(1), 3.

Pastötter, B., & Bäuml, K. H. T. (2014). Retrieval practice enhances new learning: the forward effect of testing. *Frontiers in Psychology*, 5, 286.

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K.-H. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(2), 287–297.

Pena, M. M., Klemfuss, J. Z., Loftus, E. F., & Mindthoff, A. (2017). The effects of exposure to differing amounts of misinformation and source credibility perception on source

monitoring and memory accuracy. *Psychology of Consciousness: Theory, Research, and Practice*, 4(4), 337.

Pustejovsky (2020). clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections. R package version 0.4.2. https://CRAN.R-project.org/package=clubSandwich

Putnam, A. L., Sungkhasettee, V. W., & Roediger, H. L. (2017). When Misinformation Improves Memory: The Effects of Recollecting Change. *Psychological Science, 28*(1), 36-46.

Przybyslawski, J., & Sara, S. J. (1997). Reconsolidation of memory after its reactivation. *Behavioural brain research*, 84(1-2), 241-246.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, 60, 437– 447. doi:10.1016/j.jml.2009.01.004

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rindal, E. J., DeFranco, R. M., Rich, P. R., & Zaragoza, M. S. (2016). Does reactivating a witnessed memory increase its susceptibility to impairment by subsequent misinformation? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42, 1544–1558.

Roediger III, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5), 1155.

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Imrpoves Long-Term Retention. *Association for Psycological Science*, *17*, 249–255. BB Typo

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences*, 15(1), 20-27.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, 140(6), 1432.

Rubio-Aparicio, M., López-López, J. A., Viechtbauer, W., Marín-Martínez, F., Botella, J., & Sánchez-Meca, J. (2020). Testing categorical moderators in mixed-effects meta-analysis in the presence of heteroscedasticity. *The Journal of Experimental Education*, 88(2), 288-310.

Schwarzer, G. (2007), meta: An R package for meta-analysis, R News, 7(3), 40-45.

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). Meta-analysis with R (Vol. 4784). Cham: springer.

Sederberg, P. B., Schulze-Bonhage, A., Madsen, J. R., Bromfield, E. B., McCarthy, D. C., Brandt, A., ... & Kahana, M. J. (2006). Hippocampal and neocortical gamma oscillations predict memory formation in humans. *Cerebral Cortex*, 17(5), 1190-1196.

Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science*, 22(11), 1359–1366. http://doi.org/10.1177/0956797611417632.

Slamecka, N. J., & Katsaiti, L. T. (1988). Normal forgetting of verbal lists as a function of prior testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(4), 716.

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.) *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 99--110). Chichester, England: Wiley.

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education*, 30(4), 325-341.

Sundar, S. S. (2000). Multimedia Effects on Processing and Perception of Online News: A Study of Picture, Audio, and Video Downloads. *Journalism & Mass Communication Quarterly*, 77(3), 480–499. https://doi.org/10.1177/107769900007700302

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2009). "Testing during study insulates against the buildup of proactive interference": Correction. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(6), 156.

Thomas, A. K., Bulevich, J. B., & Chan, J. C. K. (2010). Testing promotes eyewitness accuracy with a warning: Implications for retrieval enhanced suggestibility. *Journal of Memory and Language*, *63*, 149–157. http://doi.org/10.1016/j.jml.2010.04.004

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory*, 4(3), 210.

Tousignant, J. P., Hall, D., & Loftus, E. F. (1986). Discrepancy detection and vulnerability to misleading postevent information. *Memory & Cognition*, *14*(4), 329–338. http://doi.org/10.3758/BF03202511

Tulving, E., & Watkins, M. J. (1974). On negative transfer: Effects of testing one list on the recall of another. *Journal of Verbal Learning and Verbal Behavior*, *13*, 181–193.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48. URL: http://www.jstatsoft.org/v36/i03/

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. Psychonomic bulletin & review, 14(5), 779-804.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... & Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. Psychonomic bulletin & review, 25(1), 35-57.

Wahlheim, C. N., & Jacoby, L. L. (2013). Remembering change: The critical role of recursive remindings in proactive effects of memory. *Memory and Cognition*, *41*, 1–15. http://doi.org/10.3758/s13421-012-0246-9

Weber, K. (2012, September 6). The reversed testing effect: Unraveling the benefits of practiced recall. State University of New York Albany

Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face-name learning. *Psychonomic Bulletin & Review*, *18*, 518–523.

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, *3*(4).

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory*, 11(6), 571-580.

Whitten, W. B., II, & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior*, 16, 465–478. doi:10.1016/S0022-5371(77)80040-6

Wilford, M. M., Chan, J. C. K., & Tuhn, S. J. (2014). Retrieval enhances eyewitness suggestibility to misinformation in free and cued recall. *Journal of Experimental Psychology: Applied*, 20, 81–93.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review*, 18(6), 1140-1147.

APPENDIX A

Table 3 List of information coded from research reports

Report characteristics
Author name(s)
Year
Publication status

Participant characteristics
Age

Encoding variables
Stimulus type (e.g., mock-crime video, slideshow)
Stimulu exposure duration
Misinformation type
Misinformation exposure duration

Storage variables
Delay between stimulus and retrieval practice
Delay between retrieval practice and misinformation
Delay between misinformation and final test

Retrieval variables
Format of initial retrieval practice
Format of final test
Initial-final retrieval type match
Warning

Effect size
Effect size index (as labeled by authors)
Inferential information
Sample size
Misinformation production/endorsement rate
Calculated effect size for inconsistent items
Calculated effect size for consistent items
Calculation method

RES Coding Protocol

Report Characteristics

R1) What is the report ID number? (ID)

R2) What is the first author's last name? (NAME)

R3) What was the year of appearance of the report or publication? (YEAR)

R4) What type of report? (PUBSTATUS)

published

unpublished (e.g., thesis, dissertation, conference presentation)

Encoding Variables

E1) Stimulus (STIM)

mock-crime video

mock-crime slideshow

word list

word pairs

pictures

other _____

E2) Stimulus duration (STIM_LEN)

E3) Misinformation type (MISINFO_TYPE)

written narrative

audio narrative

misleading questions

other _____

E4) Misinformation duration (MISINFO_LEN)

Storage Variables

S1) Delay between stimulus and retrieval practice (DELAY1)

immediate

24hrs

1 week

other _____

S2) Delay between retrieval practice and misinformation (DELAY2)

S3) Delay between misinformation and final test (DELAY3)

Retrieval Variables

R1) Format of retrieval practice (RP)

free recall

cued recall

recognition / AFC

cognitive interview

MMFR

other _____

R2) Format of final test (TEST)

free recall

cued recall

recognition / AFC

cognitive interview

MMFR

other _____

R3) Retrieval practice — final test match (MATCH)

R4) Warning (WARN)

before misinformation

after misinformation

other _____

Participant Characteristics

P1) What is this sample ID number? (SAMPLE ID)

P2) Provide any "defining" characteristics of the sample. (DEFINING)

gender

age range

college student?

P3) What is the proportion of males in the sample? (PROPMALES)

Effect size

E1) Effect size index (as labeled by authors) (EFFECT)

E1b) Page found

E2) Sample size

E2b) Retrieval-practice condition (RP_N)

E2c) Control condition (C_N)

E2d) Page found

** If the sample size presented in Results and Method section differ, use the sample size from the Results section (e.g., table or matrix Notes).

E3) Retrieval-practice condition misinformation production (RP_PROD)

E3b) standard error, standard deviation, or confidence interval

E4) Control condition misinformation production (C_PROD)

E4b) standard error, standard deviation, or confidence interval

E5) Calculated effect size for inconsistent items (MISINFO_EFFECT)

E6) Calculation Method (CALC METHOD)

E5) Calculated effect size for consistent items (CONSISTENT_EFFECT)

E6) Calculation Method (CALC METHOD)

E7) Initials of person calculating the effect size

Coder and Coding Characteristics

C1) What are your initials? (INITIALS)

C2) In minutes, approximately how long did it take you to code this study? (MINUTES)

C3) Provide any notes about the reports or concerns regarding your coding of it. (NOTES)

** Make sure to include all notes in a single cell and “drag them down” to all rows.

Appendix B – Email sent to authors

Hello XX,

My name is Brendon Jerome Butler, and I'm working on a retrieval-enhanced suggestibility meta-analysis. While we've attempted to ensure all published studies on this research question have been collected, we also hope to include all relevant unpublished empirical work in our meta-analysis. To that end, we would greatly appreciate if you'd be willing to share any unpublished studies on this topic or could recommend other sources of unpublished empirical work. In particular, I am looking for any research where: (a) All subjects were presented some type of stimulus material; (b) a randomly-assigned group of subjects engaged in retrieval practice that pertained to the details from the stimulus; (c) a different randomly-assigned group of subjects did not engage in retrieval practice but performed some alternate task(s) (control condition); (d) all subjects were exposed to misinformation; and (e) all subjects were tested on their memory for details in the originally-encoded stimulus.

I would be very grateful if you could send or refer me to an electronic copy of relevant research reports. A citation of empirical work that I could track down would also be great, as any information is helpful in our search for studies! Of course, we would be happy to share a copy of our results after completion of the search and analysis upon request.

Many thanks for your time and help!

Sincerely,

Brendon Jerome Butler, M.A.

Graduate Student

Psychology and Social Behavior

University of California, Irvine

Appendix C — Abstract Screening Guide

For all questions below, answer “yes”, “no”, or “maybe/unsure”.

Any question answered “no” is excluded.

Do not answer any further questions after the first “no”.

1. **Primary Research**:

Is the abstract *not* a review of research (e.g., systematic review, meta-analysis)?

2. **Language**:

Is the abstract written in English?

3. **Retrieval practice and final test**:

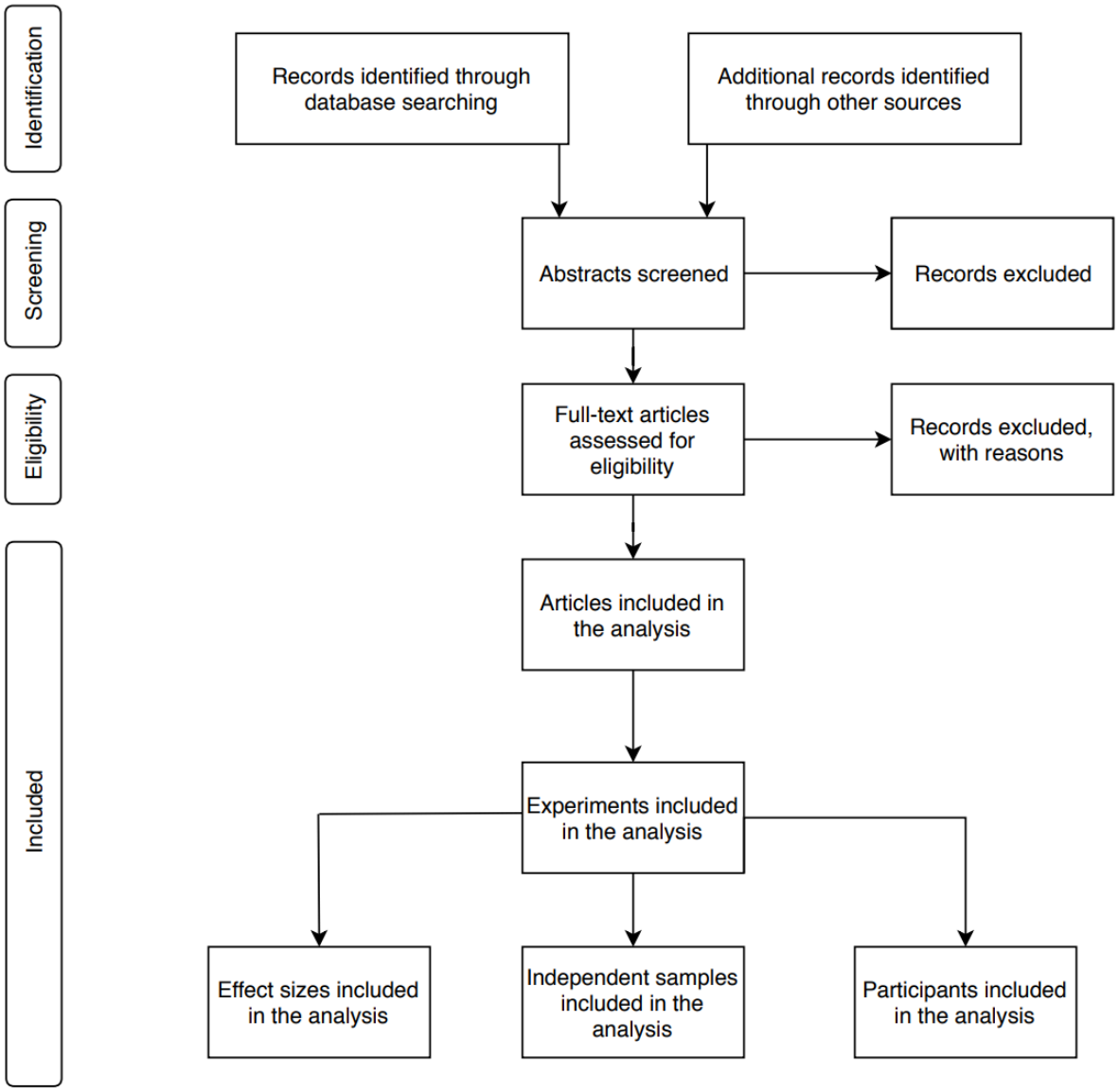
Do participants engage in some form of retrieval practice and later take a final test?

4. **Misleading post-event information or questions**:

Are participants exposed to misleading information? Either via a narrative, questioning, or some other method?

Decision: **Keep** (all “yes” or “maybe/unsure” answers) or **Drop** (at least one “no” answer)

Appendix D — PRISMA FLOWCHART



APPENDIX E — Full-text Screening Tool

For all questions below, answer “yes”, “no”, or “maybe/unsure”.

Any question answered “no” is excluded.

Do not answer any further questions after the first “no”.

1. **Primary Research**:

Is the research report *not* a review of research (e.g., systematic review, meta-analysis)?

2. **Language**:

Is the abstract written in English?

3. **Encoded stimuli**:

Do all participants encode some initial stimuli?

2. **True Experiment**:

Are participants randomly assigned to either a retrieval-practice or control group?

3. **Retrieval practice and final test**:

Do participants engage in some form of retrieval practice and later take a final test?

4. **Misleading post-event information or questions**:

Are participants exposed to misleading information following retrieval-practice? Either via a narrative, questioning, or some other method?

Decision: **Keep** (all “yes” or “maybe/unsure” answers) or **Drop** (at least one “no” answer)

Appendix F – Coding Database

https://docs.google.com/spreadsheets/d/1-