Title

Factors That Influence Metacognitive Judgments: Effects at Encoding, in the Presence of Diagnostic Cues, and After Incidental Encoding

Author

Blake, Adam Bradley

Publication Date

2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Factors That Influence Metacognitive Judgments: Effects at Encoding, in the Presence of Diagnostic Cues, and After Incidental Encoding

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Psychology

by

Adam Bradley Blake

2018

ABSTRACT OF THE DISSERTATION

Factors That Influence Metacognitive Judgments:
Effects at Encoding, in the Presence of Diagnostic Cues,
and After Incidental Encoding

by

Adam Bradley Blake
Doctor of Philosophy in Psychology
University of California, Los Angeles, 2018
Professor Alan Dan Castel, Chair

People prefer methods that involve subjectively easier and faster processing fluency, and emphasize performance when making judgments about their learning. Studying cue-target pairs feels much easier than seeing a cue and laboring to retrieve an answer, and reading words in larger and clearer fonts feels easier. When people see information (like logos or flags) very often, they tend to think that they are easier to remember and consistently bias their confidence upward. The first set of studies (Chapter 2) examines a current debate in metamemory research regarding the roles of fluency (Rhodes & Castel, 2008a) and belief cues (Mueller, Dunlosky, Tauber, & Rhodes, 2014) in the construction of judgments of learning (JOLs). The results provide clear confirmatory evidence for the effects of belief on JOLs, though these data neither support a pure fluency hypothesis nor a pure belief-based hypothesis. I discuss an additive effect of perceptual fluency and belief on JOLs, and present possible mechanisms that may interact to influence and bias JOLs. In a second set of experiments (Chapter 3), I consider the generalizability of paired-associate learning for foreign-language vocabulary to the medical domain. Results show better cued-recall performance for translations compared to medications, though JOLs are somewhat insensitive to learning. Lastly, research on everyday attention suggests that frequent interaction with objects often does not benefit memory or metamemory for them. Across three experiments in Chapter 4, partici-

pants gave confidence judgments and completed eight-alternative forced-choice tests of the US, Canadian, and Mexican flags. In Experiment 1, environmental availability was correlated with confidence for the US flag, despite similar recognition performance at a saturated time point in the US (July 4th) and a neutral time point (Aug. 6th). In Experiment 2 and Experiment 3 I assess two techniques for improving both memory and metamemory for these types of materials. Via a draw–study paradigm, I introduce *disfluency* to improve performance, demonstrating a powerful metacognitive debiasing intervention and extending theories of errorful learning by highlighting the role of attention.

The dissertation of Adam Bradley Blake is approved.

Gerardo Ramirez

Daniel M. Oppenheimer

Robert A. Bjork

Alan Dan Castel, Committee Chair

University of California, Los Angeles

2018

To Calvin . . .

who—among so many other things—

made sure I didn't burn out.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This study would not have been possible without the support of advisers, friends, and family.

First, I would like to thank my committee chair, Alan Castel, for his mentor-ship and flexibility as an advisor. Alan kept me on track without a heavy hand, and was all I could ask for in an advisor. I am also indebted to my committee members, Bob Bjork, Danny Oppenheimer, and Gerardo Ramirez for their investment of time and effort in this project. The professors and mentors at UCLA have created an amazing collaborative academic ecosystem. Working within this environment has led me to learn far more about fields outside of Cognitive Psychology, and foster friendships in other labs. I cannot be more appreciative.

I wish to thank my wife, Kelly Masuda, for her love, support, and understanding during the long nights and stressful times throughout this program, and my life. I also owe a great deal of thanks for the people closest to me, that saw me through thick and thin. Jorge Pulido, Cristina Chan, Faria Sana, thank you for sharing my highs, and supporting me at my lows.

Lastly, I have to thank my parents for raising me to be someone strong enough to take on such a challenging degree. Growing up wasn't always easy, but we always made it through. Thanks, Dad, Mom, and Paul.

VITA

2016	Shepherd Ivory Franz Distinguished Teaching Assistant
2013–pres.	Teaching Fellow, Department of Psychology, UCLA
2015–2017	Lecturer, The Berkeley Review (MCAT Prep.)
2012–2013	M.A. in Cognitive Psychology, UCLA.
2008–2012	B.A. in Psychology, Minor in Instrumental Music, California State University, Fresno

PUBLICATIONS AND PRESENTATIONS

Blake, A. B., & Castel, A. D. (under revision). Memory and availability-biased metacognitive illusions for flags of varying familiarity. *Memory & Cognition*.

Blake, A. B., & Castel, A. D. (2018). On belief and fluency in the construction of judgments of learning: assessing and altering the direct effects of belief. *Acta Psychologica, 186*, 27–38.

Blake, A. B., & Castel, A. D. (2016). Metamemory. In S. K. Whitbourne (Ed.), *The Encyclopedia of Adulthood and Aging*. Hoboken, NJ: Wiley.

Murayama, K., Blake, A. B., Kerr, T., & Castel, A. D. (2015). When enough is not enough: Information overload and metacognitive decisions to stop studying information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(6), 914–924.

Blake, A. B., Nazarian, M. & Castel, A. D. (2015). The apple of the mind's eye: Everyday attention and reconstructive memory for the Apple logo. *Quarterly Journal of Experimental Psychology, 58*, 858–865.

Castel, A. D., Nazarian, M., & Blake, A. B. (2015). Attention and incidental memory in everyday settings. In J. Fawcett, E. F. Risko, & A. Kingstone (Eds.), *The Handbook of Attention* (463–483). MIT Press.

Garcia, M. A., Kerr, T. K., Blake, A. B., & Haffey, A. T. (2015) Collector (Version 2.0.0-alpha) [Software]. Available from https://github.com/gikeymarcia/Collector/releases

Oswald, K. M., Blake, A. B., & Santiago, D. T. (2014). Enhancing immediate retention with clickers through individual response identification. *Applied Cognitive Psychology, 28*(3), 438–442.

Blake, A. B., & Castel, A. D. (2017, November). *In-sight, out-of-mind: Attention and learning for familiar flags.* Poster presented at the 58th annual meeting of the Psychonomic Society, Vancouver, British Columbia, Canada.

Nguyentan, C., Blake, A. B., & Castel, A. D. (2017, May). *Memory and metamemory for medication side effects in context and out of context in younger adults.* Poster presented at the 4th annual UCLA Undergraduate Research Week, Los Angeles, CA.

Blake, A. B., Hargis, M. B., & Castel, A. D. (2016, November). *Differences in associative memory and metamemory across domains: Foreign vocabulary and medications.* Poster presented at the 57th annual meeting of the Psychonomic Society, Boston, MA.

Hansen, H. A., Middlebrooks, C. D., Blake, A. B., & Castel, A. D. (2016, May). *Eye recognize this: The effect of impression formation on memory for eye colors.* Poster presented at the 3rd annual UCLA Undergraduate Research Week, Los Angeles, CA.

Blake, A. B., Murayama, K., Kerr, T. K, & Castel, A. D. (2016, April). *Information overload and aging: When do older adults choose to stop encoding?* Poster presented at the Cognitive Aging Conference, Atlanta, GA.

Noh, S. M., Kerr, T. K., Blake, A. B., & Castel, A. D. (2015, November). *Font size and value framing effects on memory and metamemory.* Poster presented at the 56th annual meeting of the Psychonomic Society, Chicago, IL.

CHAPTER 1

Introduction

Metamemory, or knowing what you know, is a critical component of everyday life, be it in personal, professional, educational, or other social settings. A spouse heading to the grocery store must consider whether he will remember all the necessities, a programmer must consider her current level of skill and the task requirements before projecting a date of completion, and professional game testers must constantly consider how their knowledge of game mechanics influences their perceptions of usability and ease for the end-user. Each of these situations requires that a person consider their current level of knowledge, which is almost always a subjective estimate (Schwartz, Benjamin, & Bjork, 1997), and translate that into objective, concrete information that can be used to inform future behavior.

As a general framework, metacognitive processes can be categorized as either monitoring processes, which assess cognition, or control processes, which inform and regulate cognition. Monitoring processes are assessed introspectively using measures like retrospective confidence judgments or prospective appraisals of performance, whereas control processes are often associated with behavioral measures, such as the amount of time a learner devotes to a task. Conceptually, monitoring and control are strongly linked: monitoring processes assess cognitive performance and apprise control processes which lead to behavioral decisions regarding cognition (Nelson & Narens, 1990).

Effective decisions regarding control of learning are highly dependent on the accuracy of monitoring processes. A forklift operator who is studying the manual for a new type of vehicle may use confidence in how well the material has been learned as a cue to terminate study. For an extremely accurate metacognitive monitoring system, confidence is a good

cue, though in reality there is evidence that metacognitive judgments like confidence can be in direct opposition to actual performance. For example, when an answer to a question comes to mind easily it is often given with high confidence, regardless of whether it is correct (Koriat, 1997). An overly confident learner may terminate study prematurely and in the case of the forklift operator, may cause serious harm.

A common way of studying prospective judgments of one's own learning is to ask the learner to rate how well they think they have learned the information. These judgments of learning (JOLs) have gone by many different names in the literature (e.g. "feeling-that-I-will-know" in Groninger, 1979; "memorability ratings" in Mazzoni, Cornoldi, and Marchitelli, 1990) but always refer to prospective judgments of performance. JOLs are an introspective measure of metacognitive monitoring; they assess a learner's feelings and intuitions about their own memory at the time of study. In the classic paradigm (Arbuckle & Cuddy, 1969), participants make judgments about their learning performance directly following study of an item. For example, a participant studies the paired-associates apple – table and then makes a quantitative judgment of how likely that word is to be remembered later. JOLs typically correlate with ease-of-processing (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989), and when this factor is closely related to a good predictor of future recall, such as associative strength of the cue and target, JOLs are also good predictors of memory (Arbuckle & Cuddy, 1969).

In some cases, JOLs have been shown to affect memory for the items that are being judged. Specifically, the additional processing required to consider the stimulus and judge how well it has been learned is sufficient to enhance memory in relation to studying alone (Arbuckle & Cuddy, 1969), and in the case of delayed JOLs it can be likened to retrieval practice (Spellman & Bjork, 1992).

It is also important to consider the indirect impact JOLs have on control processes. Under the Discrepancy Reduction Model, it is suggested that learners allocate more study time to items that have the largest discrepancy between the current and desired learning states (Thiede & Dunlosky, 1999). Alternatively, Metcalfe and Kornell (2005) proposes that people choose study items that are in a proximal region of learning, or simply put, that

people choose items that they feel are neither too hard to learn nor too easy. The Region of Proximal Learning model further states that participants choose to terminate study of an item when they feel they are no longer learning enough from it based on judgments of rate of learning. Focusing on the higher order links between these models, both agree that there is a causal connection between JOLs and allocation of study (and see Metcalfe, 2009). This connection is generally shown in the context of study-time allocation, where items that are believed to be less well learned are given more study time (Ariel, Dunlosky, & Bailey, 2009; Metcalfe & Kornell, 2005; Nelson, Dunlosky, Graf, & Narens, 1994).

Given the option to self-pace study, however, learners are likely to allocate too much time to harder materials, leading to a "labor-in-vain effect" (Nelson & Leonesio, 1988). This has parallels in other decision making research as well. A particularly similar finding in New York City taxi drivers suggests that one reason people labor-in-vain is because their evaluations are too narrow in focus (Camerer, Babcock, Loewenstein, & Thaler, 1997; Crawford & Meng, 2011). Drivers tend to work longer hours on the days where their income rate is slower, likely because they threshold their acceptable earnings at a daily rate rather than a weekly rate. Similarly, learners tend to judge their learning at the present moment, rather than in the context of other factors, like fatigue, proactive interference, etc. But when accurate, metacognitive judgments result in more effective study and better performance (Thiede, Anderson, & Therriault, 2003, e.g.).

Metcalfe and Finn (2008) provides evidence that even inaccurate monitoring has observable effects on study-choice. In that study, JOLs were reliably influenced by the manipulation of encoding fluency at study. When given the option of which items to study, participants preferentially studied the items they had previously given lower JOLs. This finding has been extended to perceptual materials as well, where words presented at softer volumes are chosen for restudy over louder items (Rhodes & Castel, 2009). One can draw a poignant conclusion from these findings: arbitrary perceptual cues have tangible consequences on studying and learning.

1.1 Relative Versus Absolute Accuracy

JOLs are typically made as a percent likelihood that a target will be recalled at a later test. A participant's predicted performance can be compared against their actual performance to yield a measure of calibration. As such, calibration refers to the overall accuracy of a participant's JOLs; if a participant studies a list of items and gives an average JOL of 40%, a score of 40% on a subsequent test would demonstrate a perfectly calibrated metamemory. Though JOLs are often taken as subjective probabilities, it is neither clear if participants understand their memory enough to use the scale in a reliable manner, nor is it certain that participants are calculating probabilities. There is recent evidence to suggest that JOLs are two-part judgments comprised of a binary prediction of recall and a confidence rating (Hanczakowski, Zawadzka, Pasek, & Higham, 2013). Nevertheless, a vast majority of recent studies take JOLs on a percentage scale (Rhodes & Tauber, 2011), likely because it allows for a one-to-one comparison with performance.

A major strength of the JOL as a research tool is that it captures item-by-item fluctuations in metacognitive evaluation. A participant's relative accuracy is often referred to as the resolution of his or her JOLs, and it is typically represented as the computed Kruskal-Goodman gamma correlation (see Gonzalez & Nelson, 1996). The gamma correlation is a nonparametric index, and in the case of JOLs it reveals the correlation between relative magnitude of JOLs and memory, by item. Essentially it is the extent to which items given higher JOLs are remembered and items given lower JOLs are forgotten. The gamma correlation is often used as an estimate of metacognitive resolution (e.g. Koriat, 1997), but it has increased error or fails when participants only use a limited set of responses (e.g. only 0 and 100; or 0, 50, 100). In some cases, multilevel regression models are used to compute a similar statistic (e.g. Hertzog, Hines, & Touron, 2013).

Measures of calibration and resolution are key indices of metacognitive accuracy. Much of the work researching the processes underlying JOLs focuses on the relative failings of JOLs. As such it is extremely important to understand when JOLs are accurate. In many

cases a learner may be overly optimistic in his overall performance and yet be very accurate in judging the relative strengths of items in memory, and vice versa. It is thus important that research and interpretation of JOLs and related phenomena to consider both of these aspects of metacognitive accuracy.

1.2 Metacognitive Illusions

The construction of a JOL is considered to be an inferential assessment utilizing various internal and external cues (Koriat, 1997; Schwartz et al., 1997) rather than a direct assessment of a memory trace (e.g., Hart, 1967; Schwartz, 1994). Support for the inferential nature of metamemory and JOLs is rooted in the systematic deviation of judgments from performance. One of the central themes in the literature surrounding JOLs is the strength and influence of irrelevant cues on metacognitive judgments. There is no doubt that JOLs capture irrelevant cues at hand during encoding. For example, salient, non-diagnostic cues such as font-size (Rhodes & Castel, 2008a) or backward associative strength between word pairs (Koriat & Bjork, 2005) can strongly impact JOLs. Particularly strong examples of the impact of irrelevant cues include paradigms in which the influential cues are exposed separately from the stimulus and thus have no impact on encoding (e.g. Soderstrom & McCabe, 2011). These metacognitive illusions can give insight and build a foundation to examine what learners believe JOLs measure, and what JOLs actually measure. Here I present three broad categories of fluency, or ease of processing, which commonly lead to metacognitive illusions.

Encoding fluency. An early and influential study on the effects of fluency showed that participants generally expect that memory and ease of processing are positively correlated (Begg et al., 1989). Participants predicted better memory for concrete words and high frequency words than for abstract words and low-frequency words, respectively. Though concrete words and high frequency words are easier to process than their counterparts, only concrete words showed memory performance in line with participants' predictions; low-

frequency words out-performed high-frequency words on a recognition test, neatly illuminating a dissociation of memory and predictions about memory. Though Begg et al. (1989) hinged their argument upon encoding fluency, the only consideration of ease of processing was made in choosing the stimuli: concreteness and word frequency. A major critique of the study is that it did not include a measure of fluency (Hertzog, Dunlosky, Robinson, & Kidder, 2003). To address this issue, Hertzog et al. (2003) asked participants to generate interactive images of paired-associates and press a key as soon as each image was formed mentally. The latency of the key press following presentation of the stimuli was used as a measure of the relative ease of encoding. In concert with the findings in Begg et al. (1989), participants' JOLs positively correlated with encoding fluency even though encoding fluency was not an accurate predictor of recall. Other studies have found similar effects of encoding fluency on JOLs. Generally speaking, when study time or effort increases, JOLs decrease (Koriat, 2008; Koriat & Ma'ayan, 2005).

Another type of fluency relevant at encoding is the relatedness of the items. Related items are in fact better remembered at test than unrelated items, and this is tracked by JOLs reliably (Arbuckle & Cuddy, 1969; Castel, McCabe, & Roediger, 2007; Koriat, 1997). Yet, there are still some cases where this heuristic breaks down, usually when the conditions at learning do not match conditions at test. For example, during learning the question and answer often appear in conjunction. The availability of the answer can lead to a "perspective bias" where the learner has difficulty adopting the perspective that he will have at test (Koriat & Bjork, 2005). This can lead to learners rating identical pairs (apple – apple) very highly even though they essentially have no associative strength (Castel et al., 2007).

Retrieval fluency. Retrieval fluency, or the ease of retrieving an item from memory, is generally a good cue when forming JOLs as it orients learners to cues that will be relevant at test (Begg, Vinski, Frankovich, & Holgate, 1991; Koriat, 1997). In the case of delayed JOLs, retrieval is a key process involved with determining an item's relative strength in memory, and learners effectively use retrieval fluency when forming JOLs (Nelson & Dunlosky, 1991). However, there are some exceptions to the accurate use of this heuristic, namely situations

when retrieval processes at study do not match retrieval processes at test (Benjamin, Bjork, & Schwartz, 1998).

When participants are asked to retrieve information at study they tend use the fluency of retrieval as a cue for judging learning. For example, in Begg et al. (1991) learners gave higher ratings of memorability to items that had been generated rather than simply read, a process which involves retrieving a target from memory. This finding demonstrates a general understanding of the benefits of active learning over passive learning. Following from an understanding of desirable difficulties in learning, one would expect that increased difficulty retrieving an answer at study should result in better learning (E. L. Bjork & Bjork, 2011; R. A. Bjork, 1994). However, learners are often naïve to this concept, and it is reflected in their judgments.

In a clever manipulation of the relevance of retrieval fluency, Benjamin et al. (1998) asked participants to rate how well they would later recall answers to general knowledge questions. Critically, these answers were to be recalled in the absence of the original question cues. Because the cue was not present at test, increased retrieval fluency was not an accurate of better learning on a later free recall task. Instead, reaction time was positively correlated with memory performance, likely due to the benefit of desirable difficulties. JOLs were wholly dissociated with this finding and instead negatively correlated with reaction times, meaning learners gave lower JOLs to better learned material.

Perceptual fluency. Perceptual fluency refers to the ease of perceiving a stimulus. Perceptual fluency can be considered somewhat special in comparison to encoding and retrieval fluency because, unlike the other two, there is no intuitive link between ease-of-perception and semantic processing. Nonetheless, people provide larger JOLs for intact versus backward-masked words (Besken & Mulligan, 2013), large versus small words (Rhodes & Castel, 2008a), loud versus soft words (Rhodes & Castel, 2009), clear versus blurred type (Yue, Castel, & Bjork, 2013), and upright versus inverted words (Sungkhasettee, Friedman, & Castel, 2011), to name a few.

When participants are shown words in large and small font, they tend to rate the words in large font as more likely to be remembered at test than those in small font, a difference that was not reflected in recall (Rhodes & Castel, 2008a). When given multiple study-test trials participants showed an overall improvement to their calibration, but the font-sized effect remained intact with some reduction. The effect also remained when participants were overtly told that there was no correlation between font-size and memory and when a more diagnostic cue (relatedness) was introduced. It was hypothesized that participants feel that larger words are more perceptually fluent, and indeed when perceptual fluency was equated (by making all items disfluent regardless of font-size; i.e. aPpLe) the font-size effect was eliminated.

In a direct extension of the font-size effect participants were presented with words spoken at different volumes, and words spoken more loudly were given higher JOLs than those spoken very softly (Rhodes & Castel, 2009). Though the researchers make an argument for actual ease-of-perception, they do note that it is likely a subjective perception that is influencing judgments; that is, there is little reason to believe that words presented legibly in differing fonts or audibly in differing volumes would be any more or less understandable. A recent study has shown that participants rate words in large and small font similarly in a lexical decision task (Mueller et al., 2014); the lack of difference in a lexical decision task can arguably be considered objective evidence against a difference in perceptual fluency between fonts (18 pt and 42 pt). Rhodes and Castel (2008a, 2009) further note that participants may simply be integrating the most readily available perceptual cues (font-size, volume) as those are the only cues with obvious groupings and differences.

More work exploring the relationship between perceptual fluency and JOLs has shown that stronger manipulations which substantially affect perception of the stimulus have effects similar to the font-size bias. Participants shown words in blurred type and clear type rated those in clear type as better learned in a within-subjects design (Yue et al., 2013). In the same study, though JOLs were consistently higher for words in clear type, differing patterns emerged for recall depending on the relationship between context at encoding and context at

test. Specifically, for shorter presentation times and explicit tests of memory (as opposed to recognition) participants' JOLs were accurate: items in perceptually fluent type were better recalled. Importantly, across the five experiments presented, participants did not differ in their patterns of judgment even though recall fluctuated.

1.3 What Are JOLs Actually Measuring?

Considering the multiple categories of metacognitive illusions presented above, JOLs appear to be based upon a mixture of heuristics and other naïve theories about learning (Alter & Oppenheimer, 2009; R. A. Bjork, 1999; Koriat, 1997; Rhodes, 2016). It is very likely that the reason why these illusions are so powerful is because they exploit heuristics which are generally predictive in the real world and they are being incorporated at study in a more-or-less automatic fashion.

Research shows that learners incorporate various contextual cues when making JOLs (see Rhodes, 2016; Schwartz & Efklides, 2012). Under the "cue-utilization framework" (Koriat, 1997), learners integrate three different classes of cues—intrinsic, extrinsic, and mnemonic—when judging their learning. The three types of cues are not given equal weight by learners, nor are they equally predictive of future performance. Intrinsic here refers to the qualities and characteristics of a stimulus. Examples of these cues include perceptual salience or concreteness, but can include anything about a stimulus that is indicative of learning or that is believed to be indicative of learning. Intrinsic cues are contrasted with extrinsic cues, which refer to the contextual conditions surrounding encoding or testing, such as the amount of time allotted for a task or the number of preceding items in a list. There are also mnemonic cues which are concerned with the way stimuli are experienced by the learner. Most often these cues are described in terms of ease-of-processing (e.g. Begg et al., 1989) but also include other indices such as familiarity.

This framework provides an accessible method of considering the different factors involved in metacognitive judgments and how undue attention to non-diagnostic cues such as spoken

volume (Rhodes & Castel, 2009) might interact with beliefs about memory (Mueller et al., 2014) or important contextual shifts between study and test (e.g. Benjamin et al., 1998; Koriat & Bjork, 2005). The cue-utilization framework has strong explanatory power, especially when considering the discrepancies between predictions and performance. At study learners rely on intrinsic cues like font-size (Rhodes & Castel, 2008a) and mnemonic cues like memory for a past test (Finn & Metcalfe, 2008), and they use these cues at the exclusion of relevant extrinsic cues like serial position (Castel, 2008). Further, the more temporally distal a judgment is made after seeing a stimulus, the more accurate that judgment (Koriat & Ma'ayan, 2005; Nelson & Dunlosky, 1991). Under this framework, this is likely because cues intrinsic to the stimuli have dissipated and the learner is captured by much more diagnostic cues such as difficulty in recalling a target.

Though the cue-utilization framework does well to explain many of the relationships among cues and judgments, Rhodes (2016) notes that as a framework it lacks a clear mechanism of influence. The framework contends that cues have direct effects on judgments, but it is unclear whether the cues driving these illusions have direct effects on metacognition (e.g. Rhodes & Castel, 2008a) or if instead it is the effect of beliefs and naïve theories about how the cues should work (e.g. Alter & Oppenheimer, 2009). Matvey, Dunlosky, and Guttentag (2001) provides empirical evidence for theory-driven, analytic construction of JOLs: in an experiment where participants observed and rated others' learning, the watchers produced a pattern of judgments that matches self-judgments given by performers. Even in a removed state where the effects of retrieval fluency could not be felt, participants made identical judgments.

Another method of assessing participants' beliefs about learning is via the pre-JOL paradigm (Castel, 2008). Under this paradigm, participants are asked about how well they would expect to learn some information, and they are asked this in the absence of the information. Using such a paradigm, multiple studies have shown a bias toward large-font words being easier to learn (Kornell, Rhodes, Castel, & Tauber, 2011; Mueller et al., 2014). In these studies, participants were asked prior to studying any words the relative percentages

of small-font and large-font words that they would be able to remember at test. Participants consistently predict better performance for large-font words. Mueller et al. (2014) further suggests that these beliefs may account for nearly all of the variance in JOLs. The major argument of the study is that items in 48 pt font are not disfluent compared to 16 pt font (because participants do not show differences in lexical decision times across font size) but they do show a preference for large-font words in their pre-JOLs. Recent work presented in this dissertation suggests that these effects of belief play a large component but cannot completely account for all of the effects of perceptual fluency on JOLs. It is clear though, that beliefs do play a strong role when judging learning.

Given that non-diagnostic cues and beliefs can have a strong impact on predictive judgments of memory, it should be questioned whether JOLs are not truly judgments of learning but rather judgments of performance. Though learning and performance are highly dissociable, learners may not be sensitive to this distinction (see Koriat & Bjork, 2005). That is, when participants make their judgments, they may be focusing on cues indicative of their current performance instead of cues—or lack of cues—regarding their learning. Such performance cues include ease-of-processing at encoding (Hertzog et al., 2003; Koriat & Ma'ayan, 2005), ease of perception (Besken & Mulligan, 2013; Rhodes & Castel, 2008a, 2009), and even the ability to retrieve an item from memory (Benjamin et al., 1998; Koriat & Ma'ayan, 2005; Matvey et al., 2001; Metcalfe & Finn, 2008).

Postal workers learning a type using a spaced schedule (once a day for 1 hr) rather than a massed schedule (twice a day for 2 hr) show a faster rate of acquisition, faster speed of typing, and fewer errors after completing 60 hr of training (Baddeley & Longman, 1978). Yet the same participants subjectively rated massed schedules as more satisfactory and were more likely to choose massed schedules if they were trained again or given the opportunity to train further. In a similar context, when participants are tasked with learning painting styles where paintings are grouped by artist (i.e. blocked study), the styles are less-well learned at a later test than when the artists are interleaved in presentation (Kornell & Bjork, 2008). Though performance is better in the interleaved condition, participants rate the blocked

schedule as more preferable.

Participants systematically show a preference for methods that involve a higher processing fluency and an emphasis on performance. Studying cue-target pairs together is much easier than laboring to retrieve an answer, typing feels much more fluent after four hours of practice in one day rather than reloading that process once a day for a short time, and it feels easier to compare the similarities between a group of paintings by once artist than it does to search for differences among artists and compare to past paintings by the same artist. Most importantly, these preferences exist despite running counter to more effective learning methods. Even in these cases when participants have experiential evidence contrary to their beliefs, it is not always enough to combat the influence that current level performance and ease-of-processing fluency can have on their judgments.

Finally, in an even more striking example of the preferential position of performance-based judgments, there exists a strong bias toward stability in judgments (Koriat, Bjork, Sheffer, & Bar, 2004; Kornell & Bjork, 2009; Kornell et al., 2011). Participants neither take into account the probable benefit of future study (Kornell & Bjork, 2009) nor do they adjust judgments of retention to account for extremely long delays (Koriat et al., 2004). This erroneous discounting of relevant information occurs despite the fact that participants have beliefs to the contrary and understand the relative effects that retention and restudy have on memory. Their personal beliefs are only integrated when the researchers manipulate cues to reorient participants' thoughts toward consequences, e.g. changing from judgments to learning to judgments of forgetting (Koriat et al., 2004).

It is difficult to reconcile these findings from the stability bias work into a belief-based model for fluency and JOLs: learners have overarching beliefs about retention and restudy that are not affecting their judgments. Instead these findings, and the strong effects of belief about fluency, fit well within the three-stage framework of cognition offered by Alter and Oppenheimer (2009), and can be considered an extension or elaboration of the dual-basis view discussed in Koriat et al. (2004). In this framework, a primary array of input largely influenced by processing fluency is integrated with domain-specific naïve theories

before a final judgment is made. These domain-specific theories include beliefs about how fluency should function in various domains, and the model as a whole does a fantastic job of incorporating the effects of both fluency and belief on judgments in a wide-variety of domains.

CHAPTER 2

Assessing the Contributions of Fluency and Belief to Immediate JOLs

As discussed in the introduction, there is a varied body of literature that has examined the ways fluency impacts metacognitive judgment. Often these studies focus on the negative influence of fluency and how it can create a metacognitive illusion of confidence, and, arguably, these illusions share a common theme: easily processed information should lead to easily recalled information (Koriat, 2008; Miele, Finn, & Molden, 2011). Research on JOLs has shown that peculiar but predictable metacognitive errors can arise due to influence from peripheral and irrelevant factors such as perceptual fluency (e.g. Rhodes & Castel, 2008a) or beliefs regarding the saliency of specific features (e.g. Mueller et al., 2014). Often, these factors do not positively correlate with memory performance yet exhibit an effect on metamemory nonetheless. A current debate in the metacognitive literature is whether belief or fluency is the primary instigator for these errors in metacognitive judgment. The following line of studies examines the font-size bias in memory as a method of assessing how fluency directly and indirectly affects JOLs.

2.1 Study 1[1]

There are many ways to conceptualize fluency and its effects on judgments, though generally fluency is considered the subjective ease-of-processing of information (Alter & Oppenheimer,

[1] Study 1 is now published: Blake, A. B., & Castel, A. D. (2018). On belief and fluency in the construction of judgments of learning: Assessing and altering the direct effects of belief. *Acta Psychologica*, *186*(May 2018), 27–38. doi:10.1016/j.actpsy.2018.04.004

2009). There is a varied body of literature that has examined the ways fluency impacts metacognitive judgment. Often these studies focus on the negative influence of different types of fluency and how they can create metacognitive illusions of confidence. For example, in the case of perceptual fluency, learners rate items as better remembered when they are presented in a larger font (Rhodes & Castel, 2008a) or louder volume (Rhodes & Castel, 2009). Similarly, learners are susceptible to retrieval fluency where items which come to mind easily are deemed better learned (Begg et al., 1989; Benjamin et al., 1998).

It is a strong view in the metacognitive literature that these processing fluencies nearly exclusively influence JOLs (Begg et al., 1989; Koriat et al., 2004; Koriat, Ma'ayan, Sheffer, & Bjork, 2006; Kornell et al., 2011). In a particular demonstration of the effects of perceptual fluency on metacognitive judgments, participants have a strong bias toward words presented in larger versus smaller fonts (Rhodes & Castel, 2008a). Presumably participants experience a higher degree of perceptual fluency when reading large-font words as opposed to small font words which may be perceived as harder to read. The strong argument in that study was that when the perceptual fluency of larger font words is equated with smaller font words, the bias is eliminated: words presented in alternating capital letters (e.g. "hElLo" "WoRlD") are disfluent in any font size and the font size bias is removed (Rhodes & Castel, 2008a). Similarly, when the perception of a word has been reduced or altered via backward masking, people are likely to give lower JOLs though subsequent memory is not impaired (Besken & Mulligan, 2013). Interestingly, the perceptual characteristics only appear to have a strong effect immediately after presentation, and a delay is sufficient to remove their bias in participants (Luna, Martín-Luengo, & Albuquerque, 2017), suggesting that perceptual fluency is a factor that is weighted less when compared against more diagnostic factors like retrieval fluency. Nevertheless, perceptual characteristics do have an effect on participants' choices to restudy information, and thus merit further study (Luna et al., 2017).

A criticism of Rhodes and Castel (2008a) suggests that the font-size manipulation is not enough to affect processing fluency, and thus the differences in JOLs across font size must be due to another factor. Specifically, when items on a lexical decision task are manipulated with

the same font-sizes there are no differences in classification times across font-size (Mueller et al., 2014). However, similar effects of fluency occur for more direct manipulations of perceptual fluency as well: participants rate blurry words as less memorable than clear words, though there is little, if any, effect on memory (Yue et al., 2013). Regardless, it is a valid point that for the materials commonly used to show the font-size effect there may be no true differences in processing fluency.

The competing explanation for the variability in JOLs posits that JOLs are driven instead by participants' beliefs and expectations about the memorability of words in larger fonts (Mueller et al., 2014; Mueller, Tauber, & Dunlosky, 2013). Under this view, it is not that fluency has direct effects on JOLs, but rather that it takes the form of a heuristic: subjectively easily processed information must result in easily recalled information (Koriat, 2008; Miele et al., 2011). A fairly direct method of measuring one's belief about a stimulus is to use a pre-JOL paradigm where the judgment must be made before seeing the word (see for example Castel, 2008). In the particular case of assessing belief in the face of fluency, judgments made prior to the stimulus are necessarily unaffected by the fluency of the following stimulus. With only their beliefs to guide them, participants still show a bias toward larger fonts, providing strong evidence for the influence of beliefs when making metacognitive judgments (Mueller et al., 2014).

Logically, when utilizing a belief-based system for JOLs one would expect that dispelling any erroneous belief about font-size should in turn dispel the effect. When participants are explicitly told to ignore any variances in font-size when making JOLs because they are not predictive of later recall, there is in fact a dramatic reduction, though not complete elimination, of the font-size effect (Rhodes & Castel, 2008a). A very similar finding was shown in a study examining the effects of perceptual fluency on JOLs in auditory stimuli, where participants failed to discount perceptual fluency as a predictor for learning even when the experimenters gave overt warnings that it was non-diagnostic of memory (Besken & Mulligan, 2014). It would appear that the competing belief introduced by the experimenter had a direct effect on judgments. Alternatively, this belief can be introduced experientially:

after studying and testing their knowledge of a word list, participants have the opportunity to assess their own metacognitive decisions. When given a second study-test cycle participants show effects of debiasing, generally exhibited through lower confidence and JOLs, for large-font words (Rhodes & Castel, 2008a), though this interaction was not statistically significant.

This refinement of judgment through experience has been shown in multiple cases, providing some further evidence suggesting that there may be some effect of belief behind any differences in JOLs (Benjamin, 2003; Koriat et al., 2006; Rhodes & Castel, 2008a). An explanation for this refinement is that participants become more sensitive to the manipulated mnemonic cues, for example, associative direction and strength of word pairs in Koriat et al. (2006) or word frequency in Benjamin (2003). However, in the case of the font-size effect, participants tend to believe that font-size influences memory (Kornell et al., 2011; Mueller et al., 2014) and if they were to rely on this "mnemonic cue" one would expect an increase in the difference between large- and small-font JOLs across lists. Instead, there a numerical reduction (not statistically significant) in the effect (e.g. Rhodes & Castel, 2008a, Experiment 2), which possibly indicates some type of discounting of the effects of perceptual fluency. In all of these studies, there is no change in the type of items presented across list, yet average predicted performance is more closely aligned with average memory performance in an absolute sense. Since there is no perceptual change introduced during encoding across lists, one must infer that participants developed different beliefs about the materials, beliefs that affected their judgments.

Importantly though, the above description of experiential debiasing is not necessarily affecting all beliefs about list-learning. That is, there are various beliefs that people have when approaching a study-test cycle, one of which is likely an overly optimistic idea of their memory capacity as a whole. In a study-test cycle using items in different font-sizes, this optimism exists alongside their beliefs regarding font-size and perceptual fluency. When a learner recalls fewer items than expected, she may refine her beliefs about her memory capacity, but not those about perceptual fluency. This refinement is reflected in an experiment showing that the basic font-size effect persists across multiple lists despite reductions in JOLs

as a whole (Rhodes & Castel, 2008a). It is possible that multiple biases (font-size, memory capacity, clarity, accessibility) are adjusted together through a common experience, though it is not clear in the data whether this occurs or if the basic JOL paradigm allows for it.

Though Mueller et al. (2014) very importantly showed that participants have beliefs about font-size that may result in a font-size bias, these beliefs were assessed in the absence of the stimuli (i.e., before encoding the information). Frank and Kuhlmann (2017) show that even participants that do not report beliefs about fluency (in this case volume of a spoken word) still exhibit a pattern of responding consistent with a bias toward more perceptually fluent items. Further, there is still the unanswered question of whether the effects of belief persist in the presence of perceived differences in fluency, that is, whether beliefs are considered during a post-stimulus paradigm and affect on-line JOLs rather than fluency taking the reins, so to speak. The approach in the current study is to examine the font-size effect (Rhodes & Castel, 2008a) using the basic post-stimulus JOL paradigm.

Finally, the following studies take a direct approach in manipulating participants' beliefs about font-size, a distinction that separates the current work from the current corpus of work on the effects of belief and fluency on font-size. To complement and extend the finding that participants have prior beliefs about font-size that affect their JOLs, the current study employs methods to manipulate participant beliefs about font-size to assess the direct effects of belief. These methods follow both a belief-strengthening paradigm (Experiments 1A and 1B), where participants are given information that is intended to increase the intensity of their beliefs and thus affect their subsequent JOLs, as well as a counter-belief paradigm (Experiments 1B and 2), where participants are introduced to research that runs counter to their current beliefs about their own memory.

2.1.1 Experiment 1A

If metamemory follows a belief-driven model, it should be possible to manipulate intensity of belief and observe an effect on subsequent judgments. That is, the hypothesis implies

that stronger beliefs about font-size can lead to greater differences between JOLs for smaller and larger font words. In this experiment participants were informed of some suggestive findings that showed that larger-font words, relative to smaller-font words, may be easier to recall for college students (something that could be true in certain settings, but was not anticipated by the authors in the current design). Many people have pre-existing beliefs that words in larger fonts are easier to learn and remember (Kornell et al., 2011; Mueller et al., 2014; Rhodes & Castel, 2008a), and the introduction of the suggestive research supported this notion. This was intended to ensure that all participants have this belief, have recently considered it, and to lay a foundation for any later confirmatory evidence.

According to the knowledge-updating literature, participants should be updating their judgments of learning following a first study-test cycle (see Mueller, Dunlosky, & Tauber, 2015). Before starting a second study list, it is expected that participants will both experience a deficit in their recall performance relative to their JOLs on the first list, and if given no specific feedback, will likely not have enough information available to evaluate their performance for each font size. At this point, any information given regarding their specific item performance is expected to be utilized when participants update their understanding of their own memories.

To assess whether stronger beliefs alter JOLs, differing levels of confirmatory feedback were administered between study-test cycles. Confirmatory feedback has been shown to strengthen beliefs and can increase false memories (Zaragoza, Payment, Ackil, Drivdahl, & Beck, 2001), and it is expected that participants given confirmatory feedback suggesting they recalled more of one font than another will use this information to update their JOLs on the next study list, thereby increasing the font-size effect. If JOLs are belief-driven, more strongly held beliefs should increase the magnitude of the difference between JOLs for large- and small-font words.

2.1.1.1 Method

Participants and design. The participants were 88 introductory psychology students from the University of California, Los Angeles, who participated for course credit. Participants completed two study-test sets where the font-size of the studied items was manipulated within-subjects along two levels (large, 48pt; small, 18pt). Instructions were manipulated between subjects prior to the second set across three levels: none, where no additional instruction was given; repeated, where the study's initial instructions were shown on-screen again; and feedback, where pre-scripted feedback on performance was given. Two dependent variables, JOL and recall, were measured for each participant.

Materials. Two study lists of 42 nouns each were taken from the Kučera and Francis (1967) norms. For each participant, each list was randomly divided into two sets of 18 items that were presented equally often in 18 pt or 48 pt Arial font. The remaining six items served either as primacy or recency buffers, presented equally often in 18 pt or 48 pt font, and were excluded from all analyses reported. The study lists were equated for frequency ($M = 45.35$), number of syllables ($M = 1.76$), and number of letters ($M = 5.74$), again using the Kučera and Francis (1967) norms. The presentation of the lists was counterbalanced such that half of the participants saw the one list first, and the other half saw the other list first.

Procedure. After providing informed consent, participants were told that they would study words presented in different font sizes. Additionally, they were given the information that "research has shown that, for college-age participants, words in larger fonts are easier to recall than words in smaller fonts," but that they should attempt to remember as many words as possible. After receiving these instructions participants both had to acknowledge to the experimenter that they read and understood the instructions, and then answer a question about the prompt. The answer to the question reiterated the information about font-size and memory. Participants then began the presentation of the first study list. All study stimuli were presented for 4 s each in white, lowercase letters in the center of a black background on a computer screen. Immediately following the presentation of each word participants were

given 4 s to rate their confidence that they would be able to recall that item on a scale from 0% (no chance of recall) to 100% (certain of recall). Participants were encouraged to use the entire range of the scale. Immediately following the study lists participants engaged in a 4-min distractor task where they recalled either as many Presidents of the United States as they could or as many states of the United States as they could. After this distractor, they were given 4 min to recall as many of the items as they could from the study list on a blank sheet of paper.

After recall of the first list and before presentation of the second the instructions were manipulated to vary the intensity of the belief for each group. In one group, no additional instructions were given between lists (low-intensity), a second group received the exact same instructions that preceded the first list (medium-intensity), and the final group was given the instructions again along with general feedback that they had performed better on words in larger font sizes (high-intensity). Following this, participants received a new study list, distractor task, and recall phase in the same fashion as the first. After completing the study, participants were debriefed and informed of the specific instances when font-size is likely to impact memory, e.g. the Von Restorff effect (Von Restorff, 1933).

2.1.1.2 Results

To test these any differences in JOLs and performance across font size and list, as well as the effects of instruction, a 3 (instruction) x 2 (list) x 2 (font) mixed ANOVA, with instruction as the between-subjects factor, was performed. The alpha level was set to .05 and all effect sizes are reported in terms of η_p^2 for ANOVAs and Cohen's d for t tests.

JOLs. We first examined the data for any effects of the central instruction manipulation on JOLs, and there was neither a significant three-way interaction between instruction, font, and list [$F(2, 85) = 1.08, \eta_p^2 = .025,\ MSE = 46.56, p = .34$]; nor a significant two-way interaction of instruction and font [$F(2, 85) < 1, \eta_p^2 = .008,\ MSE = 89.25$]; nor a significant two-way interaction of instruction and list [$F(2, 85) < 1, \eta_p^2 = .013,\ MSE = 127.74$]. Finally,

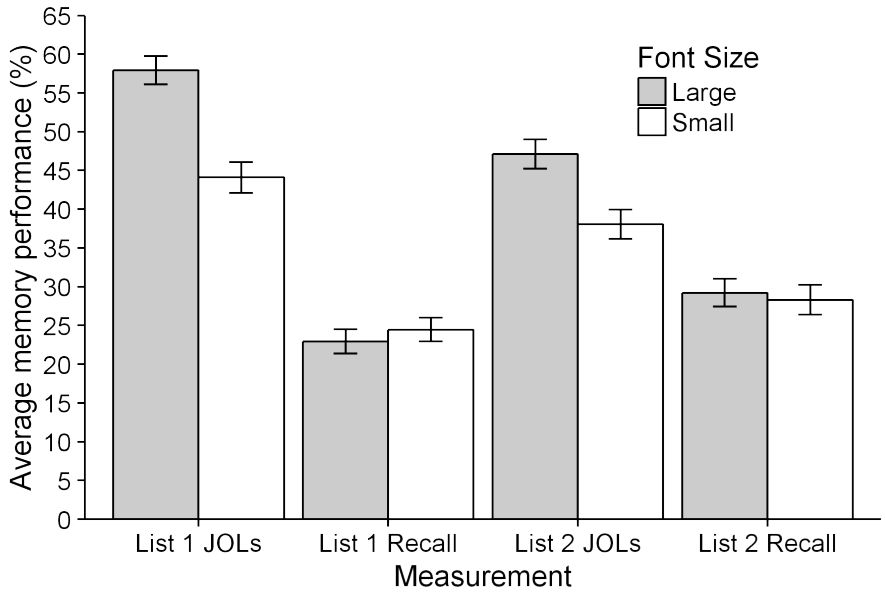


Figure 2.1. Mean predicted recall (JOLs) and free recall for words in Experiment 1A, divided by list and by font-size. Error bars represent standard errors of the mean.

instruction had a similar effect on JOLs regardless of whether participants received no instructions between lists ($M = 45.53$, $SD = 13.89$), repeated instructions between lists ($M = 48.25$, $SD = 18.49$), or feedback with instructions between lists ($M = 46.63$, $SD = 17.54$), $F(2, 85) < 1, \eta_p^2 = .006$, $MSE = 857.40$. The effects regarding the instruction manipulation being null, the descriptive statistics for JOLs have been collapsed across the three instructional conditions in the analyses below (see Table 2.1 for the full set of means and standard deviations).

Figure 2.1 shows the effect of font-size in both lists collapsed across the three instruction types. On average, small-font words appeared to be given lower JOLs than words in large font. This trend held across both lists, with what appears to be a reduction in the magnitude of the effect as well as an increase in the calibration of the JOLs and performance across lists. As suggested by Figure 2.1, the font-size effect for JOLs was reliably replicated in this experiment as well as the hypothesized reduction in the font-size effect across lists, yielding a significant interaction of font-size and list, $F(1, 85) = 10.49, \eta_p^2 = .110$, $MSE = 46.56, p = .002F(1, 85) = 10.49, \eta_p^2 = .110$, $MSE = 46.56, p = .002$. As shown in Figure 2.1, in

the first study list participants gave higher JOLs for words in 48pt font (M = 57.95, SD = 16.17) than those in 18pt font (M = 44.10, SD = 17.38), $t(87) = 9.60, d = 1.02, p < .001$. This pattern persisted to the second study list where participants again rated words in 48pt (M = 47.13, SD = 16.43) with higher JOLs than 18pt words (M = 38.06, SD = 16.45), $t(87) = 6.66, d = .71, p < .001$. Note both the large overall reduction in JOLs across lists ($M_{list1} = 51.03, M_{list2} = 42.60$) and the decrease in the effect size of the font-size effect ($d_{list1} = 1.02, d_{list2} = .71$) which reflect the relative debiasing and calibration of JOLs across lists with respect to the recall scores reported below. The reduction in JOL magnitudes across lists was significant in both the judgments for large font words [$t(87) = 9.27, d = .99, p < .001$] and small font words [$t(87) = 4.64, d = .49, p < .001$].

Gamma correlations. Metacognitive accuracy can also be operationalized as the degree to which the magnitude of a JOL for a word was associated with the probability that that word was recalled. This measure is generally referred to as metacognitive resolution and can be computed as the Goodman-Kruskal γ correlation between JOLs and recall for each participant (Nelson, 1984). In the following analyses, note that the degrees of freedom may differ from those in the JOL or recall analyses because the correlation cannot be computed when there is not enough variance in participants' responses.

Considering first the effects of the instructional manipulation on metacognitive resolution, participants were no more accurate when given confirmatory feedback (γ = .30, SE = .08), than when they were given repeated instructions (γ = .32, SE = .08), or no between list instructions (γ = .26, SE = .08), $F(2, 72) < 1, \eta_p^2 = .011$, $MSE = 0.07$. This lack of a main effect for the instructional condition was neither qualified by a significant two-way interaction with font-size [$F(2, 72) = 2.12, \eta_p^2 = .056$, $MSE = 0.12, p = .13$], nor list [$F(2, 72) < 1, \eta_p^2 = .007$, $MSE = 0.12$; nor was there a significant three-way interaction [$F(2, 72) = 1.01, \eta_p^2 = .027$, $MSE = 0.16, p = .37$].

Next, γ was not significantly different for items in large font (γ = .30, SE = .05) versus small font (γ = .29, SE = .04), $F(2, 72) < 1, \eta_p^2 = .001$, $MSE = 0.12$. Similarly, there was no

Table 2.1

Means and standard deviations for Experiment 1A

Condition	Measure	List	Font	*M*	*SD*
Low-intensity (n = 30)	JOL	1	Small	42.46	14.22
			Large	55.93	14.08
		2	Small	38.69	14.38
			Large	45.05	12.83
	Recall	1	Small	26.03	13.34
			Large	23.81	13.87
		2	Small	27.62	16.56
			Large	28.41	15.84
Medium-intensity (n = 30)	JOL	1	Small	45.52	18.78
			Large	59.61	19.16
		2	Small	39.47	17.35
			Large	48.40	18.63
	Recall	1	Small	23.33	16.56
			Large	22.86	13.95
		2	Small	29.37	18.59
			Large	27.30	14.63
High-intensity (n = 28)	JOL	1	Small	44.34	19.28
			Large	58.33	15.07
		2	Small	35.87	17.85
			Large	47.98	17.68
	Recall	1	Small	23.98	10.16
			Large	22.11	13.42
		2	Small	27.89	16.17
			Large	32.14	16.68

significant difference between the first list ($\gamma = .30$, $SE = .05$) and the second list ($\gamma = .29$, $SE = .04$), $F(2, 72) < 1, \eta_p^2 = .001$, $MSE = 0.12$. Further, there was no significant two-way interaction of font-size and list, $F(2, 72) < 1, \eta_p^2 = .01$, $MSE = 0.16$.

Finally, we drew from Koriat (1997) and examined the correlation between font-size and JOLs, as opposed to recall and JOLs. In Koriat (1997), the materials were manipulated to be more or less difficult, however, here we use perceptual fluency as a proxy for item difficulty. To interpret these values, values father from 0 (bounded at -1, 1) will indicate more of a reliance on font-size, where positive values indicate that participants find larger items more memorable and negative values indicate participants find smaller items more memorable. A two-way ANOVA comparing the Goodman-Kruskal γ correlations across list and condition was computed. There were no significant differences in correlations between conditions [$F(2, 85) < 1, \eta_p^2 = .01$, $MSE = 0.15$], nor was there a significant interaction [$F(2, 85) < 1, \eta_p^2 = .01$, $MSE = 0.07$]. However, there was a decreased correlation between font-size and JOLs from the first list ($\gamma = .38$, $SE = .03$), to the second list ($\gamma = .29$, $SE = .04$), $F(1, 85) = 6.02, \eta_p^2 = .07$, $MSE = 0.07, p = .02$.

Recall. The results for the effects of instruction on memory performance were similar to those for the JOL measure: there was neither a neither a significant three-way interaction between instruction, font, and list [F(2,85) = 1.16, $\eta_p^2 = .026$, MSE = 94.79, p = .32]; nor a significant two-way interaction of instruction and font [$F(2, 85) < 1, \eta_p^2 = .016$, $MSE = 161.07$]; nor a significant two-way interaction of instruction and list [$F(2, 85) < 1, \eta_p^2 = .009$, $MSE = 129.96$]. Lastly, main effect of instructions was not significant such that participants remembered similar percentages of the word lists when no instructions were given between lists ($M = 26.47$, $SD = 14.96$), instructions were repeated between lists ($M = 25.72$, $SD = 16.03$), and feedback was given with the instructions ($M = 26.53$, $SD = 14.34$), $F(2, 85) < 1, \eta_p^2 = .001$, $MSE = 532.07$. The effects regarding the instruction manipulation being null, the descriptive statistics for recall have been collapsed across the three instructional conditions in the analyses below (see Table 2.1 for the full set of means and standard deviations).

Like past research, the recall scores did not differ across font-size in this experiment. A similar percentage of words in large font ($M = 26.08$, $SD = 14.68$) were recalled as words in small font ($M = 26.38$, $SD = 15.37$), $F(1, 85) < 1, \eta_p^2 = .001$, $MSE = 129.96$. This pattern was consistent across lists and there was no significant interaction of font-size and list, $F(1, 85) = 1.47, \eta_p^2 = .017$, $MSE = 94.79, p = .23$. Finally, there was a main effect of list on recall such that a higher percentage of words were recalled on the second list ($M = 28.76$, $SD = 16.34$) than the first list ($M = 23.70$, $SD = 13.59$), $F(1, 85) = 14.21, \eta_p^2 = .143$, $MSE = 161.07, p < .001$.

Post-task items. To ensure that the instructions and research were believed by participants, a manipulation check was included in the post-task questionnaire. When asked if they had believed the research presented in the instructions at the start of the experiment, 62.50% indicated that they definitely believed the research.

2.1.1.3 Discussion

In this experiment, we attempted to bias people's beliefs regarding font size so that they were consistent with, and possibly stronger than, the standard belief that larger font words are easier to remember than smaller font words. Particularly, we supplied participants with information on font-size and memory that coincides with the very common belief that words in larger font are easier to recall. We expected that when participants' beliefs were reinforced participants would give JOLs consistent with their strengthened beliefs: increased JOLs for large words and decreased JOLs for words in small font with larger differences at each level of the instruction manipulation.

It is unclear why the instructional manipulation did not cause any effects on the font-size effect. A possible explanation is that participants did not believe the information being presented to them via the instructions, a point which is supported by the fact that only about 63% reported accepting the "research" they were presented with. However, this check was performed after the participants had undergone the experiment, and additional analyses

performed on that subset yielded the same patterns presented above. Instead we speculate that participants already had a strong belief about how font-size would affect memory. This prior belief is discussed in the introduction of this paper, and has been shown multiple times in past research (e.g. Kornell et al., 2011; Mueller et al., 2014). Additionally, the initial instructions may have already amplified the font-size bias which is suggested by the slightly larger effect size shown in list one of this experiment ($d = 1.02$) compared to Rhodes and Castel (2008a) Experiment 1 ($d = 0.83$, computed from their reported statistics). It is plausible that participants were already at or near ceiling for this belief and thus unable to strengthen it.

Though the expected finding was not shown, it is important to note that participants did show a strong font-size bias toward larger font items. The expected overconfidence in JOLs on list one was found, and the overall decrease in JOLs across lists likely reflects experiential debiasing regarding memory capacity. That is, that participants experienced the limits of their memory capacity in the context of the task, and on the second list, generally felt that items were less likely remembered. That this improved accuracy was not reflected in participants' metacognitive resolution (γ) suggests that these judgments were not updated at the item level but as a global reduction in expected performance. Recent work suggests this may be because they began using number of items as a metacognitive cue, or at least that it became a more salient cue (Murayama, Blake, Kerr, & Castel, 2016).

At a glance, the effect size of the font-size effect appears to be diminished across lists, a finding which if significant would suggest that across lists participants reduced their font-size bias. However, this is a misleading line of thought because the higher-order analysis of the interaction between font-size and list was non-significant, meaning that the difference in effect size was negligible. However, there is evidence to suggest that participants were less reliant on font-size as a predictor of learning: there was a significant reduction in the correlation between font-size and JOLs from the first list to the second. If we interpret this reduction in correlation in the same light as Koriat (1997) did with difficulty and JOLs, we might conclude that participants partially shifted their focus from font-size to some other

factor from the first list to the second.

On the whole, it appears that the instructional manipulation was not strong enough to alter participants' JOLs, that their bias toward font-size was already very high, and that their judgments appear to be less related font-size differences across lists.

2.1.2 Experiment 1B

Experiment 1A was unable to show an effect of instruction in terms of strengthening belief in the font-size effect. The interaction of font-size and list suggested a possible experiential debiasing of their beliefs about font-size but this is a tenuous claim. The current experiment took a more direct approach to manipulating participants' beliefs by introducing a competing belief. Informing participants of the true, null correlation between font-size and memory has been shown to reduce bias dramatically, however this does not eliminate the font-size effect (Rhodes & Castel, 2008a). The lack of elimination there could mean that participants either did not trust the information they were being given, or that the fluency of the text is truly driving the results. The current experiment avoids subtlety and provides a belief that is in direct opposition of the common belief about font size: smaller words are easier to remember (something that could be true in certain settings, but not in the current design). A belief-based account for JOLs should hypothesize that when participants believe that smaller words are more memorable than large words there will be a reversal in the font-size effect such that JOLs are larger for smaller words. This would be completely counter to a fluency account which would hypothesize that 18pt font words, which are presumably less fluent than 48pt words (Rhodes & Castel, 2008a), should always be given lower JOLs. Additionally, as this belief runs counter to normal intuitions about font-size and recall, it is expected to have a larger, more noticeable effect than the previous attempt in Experiment 1A to strengthen what was possibly an already strong belief.

Because this manipulation was expected to reduce, eliminate, or reverse the font-size effect, a power analysis was performed using the effect size from Rhodes and Castel (2008b,

Experiment 4) ($\eta_p^2 = .13$) and a modest correlation among the measures (.3). This particular effect size was chosen because (1) in that experiment the authors warned the participants that font-size is not diagnostic of memory performance, and (2) it was the lowest effect size of the full study and thus the most conservative estimate (in comparison the largest effect size was .45 and the average in the study was .36). The results of the power analysis suggested that in any given group judging words in small and large font a sample of at least 21 should show the font-size effect in JOLs.

2.1.2.1 Method

Participants were 90 workers recruited through Amazon Mechanical Turk and were paid $4.50 to participate in the study. The experiment used the same design, procedure, and materials as in Experiment 1A with the exception of the research that was presented prior to study. Instead of informing participants of research showing that large words are easier to remember, the opposite information was given: smaller words are easier to remember. As in Experiment 1A, the high-intensity condition participants were told that their recall on the first study list was consistent with the stated research, the moderate-intensity condition participants were only reminded of the stated research after the first recall session, and the low-intensity condition participants were given the manipulation at the beginning of the study and received no further instruction regarding it.

Following the experiment, participants were told about the goals of the experiment, and thanked for their participation. After eliminating 12 participants for failure to complete the full experiment, the final group sizes for the between-subjects instruction manipulation were 30 for low-intensity, 21 for moderate-intensity, and 27 for high-intensity.

As a note, because these workers used their own computer and browser settings, the exact size of the font could not be controlled. However, the ratio of the font sizes was preserved (48:18).

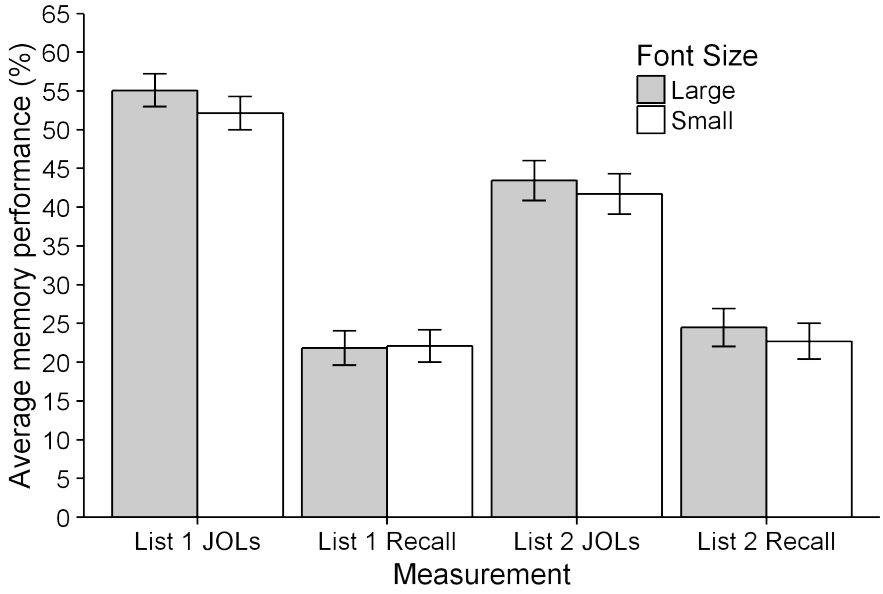


Figure 2.2. Mean predicted recall (JOLs) and free recall for words in Experiment 1B, divided by list and by font-size. Error bars represent standard errors of the mean.

2.1.2.2 Results

A simplified representation of the data is presented in Figure 2.2, which shows the effect of font-size in both lists collapsed across the three instruction types. There appear to be no strong differences across font-size in either the JOL or recall measures, regardless of list. A 3 (instruction) x 2 (list) x 2 (font) mixed ANOVA, with instruction as the between-subjects factor, was performed to test the effects of the manipulations on JOLs and recall scores. The alpha level was set to .05 for all inferential statistics, and all effect sizes are reported in terms of η_p^2 for ANOVAs.

JOLs. Examining the data for the effects of instruction, there was neither a significant three-way interaction between instruction, font, and list for JOLs [$F(2, 75) < 1, \eta_p^2 = .0067$, $MSE = 40.40$]; nor a significant two-way interaction of instruction and font [$F(2, 75) < 1, \eta_p^2 = .017$, $MSE = 80.51$]; nor a significant two-way interaction of instruction and list [$F(2, 75) < 1, \eta_p^2 = .017$, $MSE = 163.99$]. Lastly, participants gave similar JOLs regardless of whether they received no instructions between lists ($M = 46.68$, $SD = 20.39$), repeated

instructions ($M = 43.69$, $SD = 17.94$), or feedback with repeated instructions ($M = 53.09$, $SD = 19.23$), $F(2, 75) = 1.88, \eta_p^2 = .048$, $MSE = 1215.00, p = .16$. As no significant effects regarding the instruction manipulation were found, the descriptive statistics for JOLs have again been collapsed across the three conditions in the analyses below (see Table 2.2 for the full set of means and standard deviations).

Interestingly, JOLs for words in large font ($M = 49.19$, $SD = 19.36$) were always higher than JOLs for words in small font ($M = 46.92$, $SD = 19.60$). The interaction between font size and list was indeed non-significant [$F(2, 75) < 1, \eta_p^2 = .01$, $MSE = 40.40$] and the main effect of font size was significant, $F(1, 75) = 5.35, \eta_p^2 = .067$, $MSE = 80.51, p = .02$. Thus, the observed interaction of font and list on JOLs found in Experiment 1A was not replicated in Experiment 1B, where the instruction manipulation was altered. A reduction of JOLs across list was evident [$F(1, 75) = 54.74, \eta_p^2 = .422$, $MSE = 163.99, p < .001$], but this reduction was consistent across font size.

Gamma correlations. As in Experiment 1A we considered the effects of the independent variables on metacognitive resolution. Twenty-one participants were excluded from these analyses because their correlations could not be computed due to invariance in either their JOLs or recall in one or more conditions.

First, we analyzed the effects of the instructional manipulation on metacognitive resolution and found that participants were no more accurate when given confirmatory feedback ($\gamma = .28$, $SE = .08$), than when they were given repeated instructions ($\gamma = .31$, $SE = .10$), or no between list instructions ($\gamma = .29$, $SE = .09$), $F(2, 54) < 1, \eta_p^2 = .010$, $MSE = 0.22$. There were neither a significant two-way interaction with font-size [$F(2, 54) < 1, \eta_p^2 = .001$, $MSE = 0.11$], nor with list [$F(2, 54) = 1.74, \eta_p^2 = .060$, $MSE = 0.15, p = .39$]; nor was there a significant three-way interaction [$F(2, 54) = 1.51, \eta_p^2 = .053$, $MSE = 0.11, p = .23$].

Looking at font-size, γ was not significantly different for items in large font ($\gamma = .30$, SD = .05) versus small font ($\gamma = .26$, $SE = .05$), $F(2, 54) < 1, \eta_p^2 = .010$, $MSE = 0.11$. Similarly, there was no significant difference between the first list ($\gamma = .28$, $SE = .08$) and

the second list ($\gamma = .28$, $SE = .05$), $F(2, 54) < 1, \eta_p^2 = .001$, $MSE = 0.15$. Further, there was no significant two-way interaction of font-size and list, $F(2, 54) < 1, \eta_p^2 = .010$, $MSE = 0.11$.

Lastly, we again computed a two-way ANOVA comparing the Goodman-Kruskal γ correlations across list and condition to assess changes in reliance on font-size as a predictor of learning. There were again no significant differences in correlations between conditions [$F(2, 74) = 1.33, \eta_p^2 = .04$, $MSE = 0.16, p = .27$], nor was there a significant interaction [$F(2, 74) < 1, \eta_p^2 = .02$, $MSE = 0.08$]. Unlike Experiment 1A, there was no change in the correlation between font-size and JOLs from the first list ($\gamma = .11$, $SE = .04$), to the second list ($\gamma = .04$, $SE = .04$), $F(1, 74) = 1.59, \eta_p^2 = .02$, $MSE = 0.08, p = .21$.

Table 2.2

Means and standard deviations for Experiment 1B

Condition	Measure	List	Font	*M*	*SD*
Low-intensity (n = 30)	JOL	1	Small	51.34	20.08
			Large	55.14	19.21
		2	Small	38.69	20.75
			Large	41.56	21.47
	Recall	1	Small	19.84	16.83
			Large	23.81	15.62
		2	Small	21.27	19.24
			Large	23.65	19.35
Medium-intensity (n = 21)	JOL	1	Small	46.81	14.04
			Large	50.37	14.09
		2	Small	37.55	21.66
			Large	40.02	20.55
	Recall	1	Small	17.91	11.94
			Large	14.06	14.24
		2	Small	20.41	16.23
			Large	23.58	20.56
High-intensity (n = 27)	JOL	1	Small	57.07	16.89
			Large	58.72	17.30
		2	Small	48.33	21.03
			Large	48.23	21.28
	Recall	1	Small	27.87	20.28
			Large	25.75	22.42
		2	Small	26.10	21.19
			Large	26.10	20.98

Recall. Similar to the findings from the JOL analyses regarding the instruction condition, recall showed neither a significant three-way interaction between instruction, font, and list [F(2, 75) = 1.30, η_p^2 = .034, p = .28]; nor a significant two-way interaction of instruction and font [$F(2, 75) = 2.17, \eta_p^2 = .055$, $MSE = 66.66, p = .12$]; nor a significant two-way interaction of instruction and list [$F(2, 75) = 1.31, \eta_p^2 = .034$, $MSE = 221.54, p = .28$]. Finally, participants recalled the same percentage of words whether they were in the no repeated instructions condition ($M = 22.14$, $SD = 17.83$), the repeated instructions condition ($M = 18.99$, $SD = 16.06$), or the feedback with instructions condition ($M = 26.46$, $SD = 21.23$), $F(2, 75) = 1.34, \eta_p^2 = .034$, $MSE = 1016.00, p = .27$. See Table 2.2 for the full set of means and standard deviations.

In regards to the font-size effects, the percentage of words recalled was consistent across large ($M = 23.03$, $SD = 19.24$) and small fonts ($M = 22.27$, $SD = 18.29$) [$F(1, 75) < 1, \eta_p^2 = .005$, $MSE = 66.66$] and did not interact with list, $F(1, 75) = 1.37, \eta_p^2 = .034$, $MSE = 88.18, p = .28$.

Post-task items. To ensure that the instructions and research were believed by participants, a manipulation check was included in the post-task questionnaire. When asked if they had believed the research presented in the instructions at the start of the experiment, 84.62% indicated that they definitely believed the research.

2.1.2.3 Discussion

Akin to Experiment 1A, the procedure did not result in a graded, strengthening effect across instruction condition. The lack of any effects across this manipulation is not particularly surprising given the similar results in Experiment 1A. However, the belief presented in this experiment is not a commonly pre-held belief about font-size and memory: for example, in Mueller et al. (2014) only 10 participants of 48 reported small-font words as more memorable. We expected that the novelty of the belief would result in varying levels of acceptance, as well as gradation across the intensity of the manipulation. Instead, participants showed a

uniform font-size bias in their JOLs, indicating a similar change in behavior toward font-size across all instructional conditions. The lack of effect here may reflect some strong tendency of participants to accept counterintuitive research about their own memory, or alternatively may show that any effects are too variable to show fine-grained differences in the studied sample.

Performance across lists again showed a reduction in JOLs which brought them closer to actual recall performance. However, as in Experiment 1A the improved global calibration of JOLs was not qualified by a significant interaction effect between font-size and list, nor were there any improvements in resolution across lists in this experiment. As such, the difference between Experiments 1A and 1B appears to be solely due to the change in the initial instructions and any effects they may have had on participants' beliefs. Here, participants were under the impression that smaller words would be easier to remember than larger words, and though there was not a complete reversal of the font size effect, there is a definite effect of the opposing belief crippling the effect.

The fluency argument and the belief argument for JOLs make competing predictions for this experiment. For fluency, smaller words are less perceptually fluent than large words and thus no matter what information the experimenter gives the participant, JOLs should be larger for the more perceptually fluent words and should be correlated with font-size to some degree. Alternatively, belief that smaller words are more easily remembered should produce larger JOLs for words in small font and should be negatively correlated with font-size. The results show that presenting a belief that small font words are easily remembered nearly eliminates the font-size effect and a near-null correlation between font-size and JOLs ($\gamma = .04$), which suggest a very low reliance on font-size as a predictor of learning. Rhodes and Castel (2008b) showed a similar reduction in the effect after instructing participants to discount perceptual fluency as a diagnostic cue for memory. Critically, here we show a dramatic reduction in the font-size effect even when participants are aware of and even primed to attend to font-size specifically. The fact that the effect is not eliminated possibly suggests a lingering effect of perceptual fluency on JOLs that is not easily overcome by

changing one's beliefs.

2.1.3 Experiment 2

In Experiments 1A and 1B we attempted to manipulate participants' beliefs about font-size and memory and the strength of those beliefs and it was shown that participants' responses are not particularly affected by the attempts made to strengthen those beliefs. Unfortunately, these experiments lacked an appropriate control group that was not given any instruction regarding the effects of font-size on memory. It is thus unclear if the instructions themselves, regardless of the intensity manipulation, had any effect at all. Further, participants were all given reason to believe in font-size effects at the beginning of the study. If participants had given substantial effort to applying those beliefs over the course of a study list, it is not surprising that the instructional manipulations did not have a strong (or any) effect.

Importantly though, the patterns of data presented in 1A and 1B suggest that the content of the beliefs and their effects on behavior may be relatively malleable via instructional manipulations at the start of the study. That is, though people tend to show strong beliefs about font-size prior to starting studies like these (Kornell et al., 2011; Mueller et al., 2014), participants in Experiment 1B appeared to accept that words in small font might be easier to remember that words in large font. After being presented with this information, there was no effect of font-size on their JOLs though a sample size that large should have shown the standard effect under the fluency hypothesis.

In the previous experiments the instruction manipulation was performed prior to any baseline measurements regarding the font-size effect in the participant pool. That is, participants likely had *a priori* beliefs about the effects of font-size that could potentially confound the data (e.g. if participants in Experiment 1B had a strong belief that font-size does not affect learning at all), and further, participants may not have truly believed the information given to them by the researchers. To confirm that this nullification of the font-size effect was reliable, this experiment analyzes the belief-instruction manipulation against participants'

non-biased (i.e. not influenced by the researchers) beliefs about font-size.

We expected that participants would all show the standard font-size effect on a first list, such that JOLs for large font words are higher than those given for small font words. We further expected that participants' JOLs on a second list would be similar across font-size if given the belief instructions from Experiment 1B or show only a slightly reduced effect of font-size if given the instructions from Experiment 1A. Even though processing fluency should remain constant across lists, it is hypothesized that the introduction of a counter-belief will have an additive effect reducing the differences across font-size caused by fluency and prior beliefs.

2.1.3.1 Method

The participants were 62 workers recruited through Amazon Mechanical Turk, who were paid $4.50 to participate. Participants completed two study-test sets where the font-size of the studied items was manipulated within-subjects along two levels (large, 48pt; small, 18pt). These were the same word lists that were constructed for Experiment 1. Again, it should be noted that because these workers used their own computer and browser settings, the exact size of the font could not be controlled, but the ratio of the font sizes was preserved (48:18). Information about the effects of font-size was manipulated between subjects prior to the second list across two levels: participants were either informed that large or small words are easier to learn.

Procedure. After providing informed consent, participants were told that they would study words presented in different font sizes and that they should attempt to remember as many words as possible. Participants indicated their understanding of the instructions by answering a multiple choice reading check question before moving on. All study stimuli were presented in black, lowercase letters in the center of a white background on a computer screen. Immediately following the presentation of each word participants were instructed to rate their confidence that they would be able to recall that item on a scale from 0% (no

chance of recall) to 100% (certain of recall). Participants were encouraged to use the entire range of the scale. Immediately following the study lists participants engaged in a 2-min distractor task where they played a game like Tetris. After this distractor, they were given 4 min to recall as many of the items as they could from the study list by typing them into a text box.

After recall of the first list and before presentation of the second, participants were given the information regarding the effects of font-size on memory. The wording here was identical to Experiments 1A and 1B: "research has shown that, for college-age participants, words in [larger/smaller] fonts are easier to recall than words in [smaller/larger] fonts." Again, participants answered a multiple choice reading check question before moving on, and in this case the correct answer reiterated the font-size information briefly. The participants then studied the second list, played another Tetris distractor for 2 min, and finally recalled the second list for 4 min.

After completing the study, participants were debriefed and informed of the specific instances when font-size is likely to impact memory, e.g. the Von Restorff effect (Von Restorff, 1933), and also completed other questionnaire regarding how they study information and their beliefs about font-size and memory.

2.1.3.2 Results

JOLs. As shown in Figures 2.3 and 2.4, which are the graphs for the instructed-large condition and the instructed-small conditions, it appears that the instruction manipulation produces different results depending on the instruction. The baselines in each condition (list one) have a similar pattern, as expected, but after the belief is introduced the JOLs move in different directions, such that in Figure 2.3 we see a clear font size effect in the JOLs for list two and in Figure 2.4 there are no differences between the JOLs in list two. This direction of change is made more apparent in Figure 2.5, which shows the average difference in JOLs across list for each font-size. In the condition where participants were instructed that large

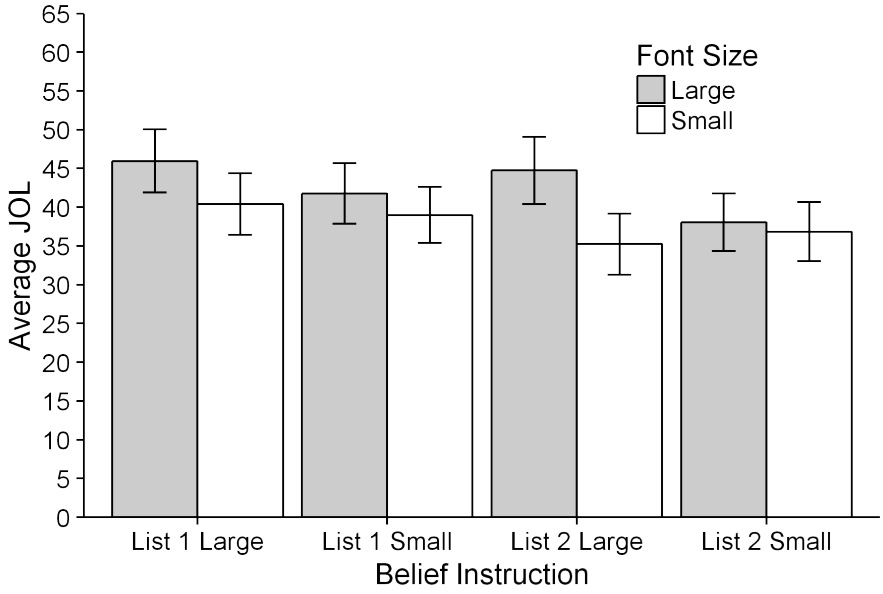


Figure 2.3. Mean JOLs for large and small font words in Experiment 2, divided by list and the belief instruction condition ("Large" indicates that participants were instructed that larger words are easier to recall). Error bars represent standard errors of the mean.

words are easier to learn the font-size effect appears to grow: JOLs for large words do not change across lists but JOLs for small words decrease across lists. In the condition where participants were instructed that small words are sometimes easier to learn the opposite is found: JOLs for words in large font become smaller and JOLs for words in small font do not change. Thus, for the instruct-large condition, it appears that the JOLs for small words become even smaller, whereas for the instruct-small condition the JOLs for the large words are reduced due to the manipulation. To test these apparent effects, a 2 (instruction) x 2 (list) x 2 (font size) mixed ANOVA, where the belief instruction was between subjects, was performed. The three-way interaction between font size, list, and belief instruction was significant, $F(1, 60) = 4.73, \eta_p^2 = .073$, $MSE = 92.54, p = .03$.

The three-way interaction seems to be driven by the simple two-way interaction between font size and the critical instruction manipulation, $F(1, 60) = 4.76, \eta_p^2 = .073$, $MSE = 74.15, p = .03$. This interaction was such that in the condition where participants were given reason to believe words in large font are easier to learn, words in large font ($M =$

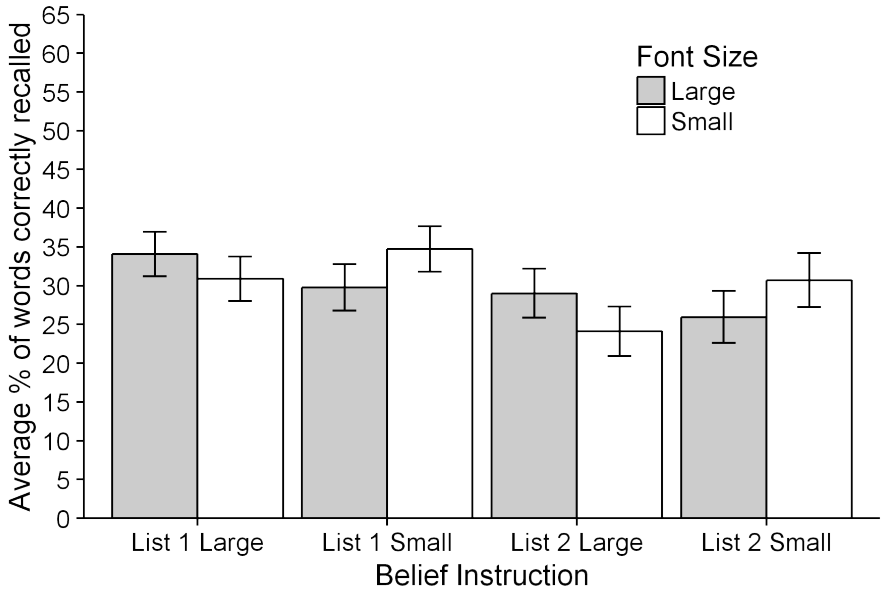


Figure 2.4. Mean percentage of words correctly recalled for large and small font words in Experiment 2, divided by list and the belief instruction condition ("Large" indicates that participants were instructed that larger words are easier to recall). Error bars represent standard errors of the mean.

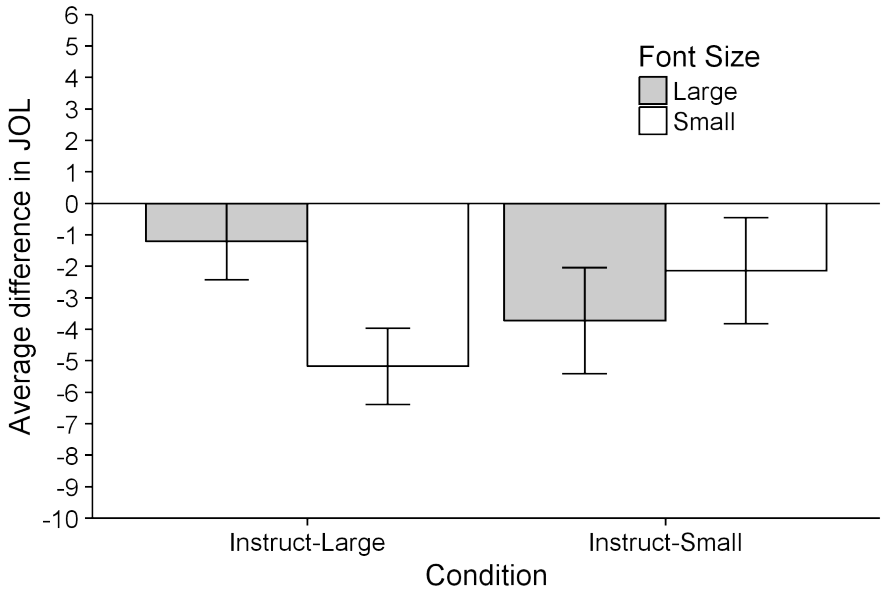


Figure 2.5. The average difference of JOLs from list one to list two for words in large and small font in Experiment 2, divided by the instruction manipulation condition. Error bars represent standard errors of the mean.

46.75, $SD = 21.55$) were given higher JOLs than words in small font ($M = 39.16$, $SD = 21.59$) [$t(61) = 4.43, d = 0.35, p < .001$], but this effect was reduced in the opposite belief condition where words in large font ($M = 41.30$, $SD = 20.99$) were given only slightly higher JOLs than those in small font ($M = 38.47$, $SD = 18.92$), $t(61) = 2.08, d = 0.14, p = .046$. The magnitude of JOLs was significantly different across the instruction manipulation such that words in large font were given larger JOLs in the condition where participants had reason to believe words in larger fonts are easier to learn [$t(61) = 2.99, d = 0.26, p = .004$] but there were no differences in the JOLs for small font across the instruction conditions [$t(61) = 0.37, d = 0.03, p = .72$].

The simple two-way interaction between the list and instruction manipulation was not significant [$F(1, 60) = .25, \eta_p^2 = .004$, $MSE = 96.26, p = .62$]. On average, participants gave similar magnitudes of JOLs regardless of whether they were in the believe-large condition ($M = 42.95$, $SD = 22.57$) or the believe-small condition ($M = 39.88$, $SD = 21.19$), $F(1, 60) = .35, \eta_p^2 = .006$, $MSE = 1654.5, p = .56$. On average participants gave lower JOLs to items on the second list ($M = 40.06$, $SD = 21.86$) than the first list ($M = 42.78$, $SD = 21.84$), $F(1, 60) = 4.80, \eta_p^2 = .074$, $MSE = 96.26, p = .03$. Finally, the general font size effect was shown such that words in large font ($M = 44.02$, $SD = 22.50$) were given larger JOLs than words in small font (M = 38.81 SD =21.27), $F(1, 60) = 22.70, \eta_p^2 = .275$, $MSE = 74.15, p < .001$.

Gamma correlations. We computed Goodman-Kruskal γ correlations to analyze the effects of the independent variables on metacognitive resolution. Eight participants were excluded from these analyses because their correlations could not be computed due to invariance in either their JOLs or recall in one or more conditions.

Considering the effects of the instructional manipulation on metacognitive resolution, we found that participants were no more accurate with their judgments when given information that suggested words in large font were more memorable ($\gamma = .37$, $SE = .09$), than when given the same information about small-font words ($\gamma = .32$, $SE = .08$), $F(1, 51) < 1, \eta_p^2 =$

.001, $MSE = 0.28$. Similarly, there were no significant effects of font-size on metacognitive resolution: large-font words ($\gamma = .36$, $SE = .06$) were judged just as accurately as small-font words ($\gamma = .34$, $SE = .06$), $F(1, 51) < 1, \eta_p^2 = .001$, $MSE = 0.17$. However, a marginally significant main effect of list was found such that judgments in the first list ($\gamma = .40$, $SE = .05$) were more accurate than judgments in the second list ($\gamma = .29$, $SE = .07$), $F(1, 51) = 3.54, \eta_p^2 = .07$, $MSE = 0.14, p = .07$.

Looking at the more complex effects of the list variable, there was neither a significant interaction with font-size [$F(1, 51) < 1, \eta_p^2 = .008$, $MSE = 0.15$], nor with condition [$F(1, 51) = 2.33, \eta_p^2 = .044$, $MSE = 0.15, p = .39$]; however, there was a marginally significant three-way interaction [$F(1, 51) = 3.24, \eta_p^2 = .060$, $MSE = 0.15, p = .08$].

Though the three-way interaction was only marginally significant, a brief exploratory suggests the interaction may be driven by differences in judgments about small-font words. Judgments for words in large font do not appear to differ across condition regardless of whether they are made in the first list [$t(26) = 1.02, d = .20, p = .32$] or the second list [$t(25) = 0.54, d = .11, p = .60$]. However, though accuracy for words in small-font in list one is better in the bigger-is-better condition ($\gamma = .46$, $SE = .07$) than the opposite ($\gamma = .31$, $SE = .07$) [$t(28) = 1.56, d = .29, p = .13$], this appears to be reversed on the second list, such that the bigger-is-better condition is less accurate ($\gamma = .20$, $SE = .11$) than the other ($\gamma = .36$, $SE = .08$), [$t(23) = -1.08, d = .22, p = .29$]. Further, this may be due to an increase in accuracy for small font words in the smaller-is-better condition from the first list ($\gamma = .31$, $SE = .07$) to the second ($\gamma = .36$, $SE = .08$) [$t(27) = -0.90, d = .17, p = .38$] coupled with a decrease in the same words in the bigger-is-better condition from the first list ($\gamma = .46$, $SE = .07$) to the second ($\gamma = .31$, $SE = .07$) [$t(25) = 2.059, d = .40, p = .05$].

Lastly, we again computed a two-way ANOVA comparing the Goodman-Kruskal γ correlations across list and condition to assess changes in reliance on font-size as a predictor of learning. In this experiment, there were no changes in correlations across lists, $F(1, 56) < 1, \eta_p^2 = .003$, $MSE = 0.11$. However, there was a significant effect of condition, such that participants in the bigger-is-better condition ($\gamma = .36$, $SE = .04$) showed higher

correlations than those in the smaller-is-better condition ($\gamma = .14$, $SE = .04$), $F(1, 56) = 10.97, \eta_p^2 = .16$, $MSE = .12, p = .002$. This main effect was driven by the two-way interaction, $F(1, 56) = 7.80, \eta_p^2 = .12$, $MSE = .11, p = .007$. Two dependent samples t tests were run to determine the nature of the interaction. In the bigger-is-better condition, there was no difference from the first list ($\gamma = .26$, $SE = .06$) to the second list ($\gamma = .21$, $SE = .06$), t(28) = 1.60, d = 0.15, p = 12. However, in the smaller-is-better condition, there was a decrease from the first list ($\gamma = .36$, $SE = .04$) to the second list ($\gamma = .36$, $SE = .04$), $t(28) = 2.56, d = 1.05, p = .02$.

Recall. Again, referring to Figures 2.3 and 2.4, the recall scores appear to have a slight decline across lists. Additionally, there appears to be a difference in the pattern of which font size was better recalled depending on the condition that the participant was in. Specifically, in the condition where participants were instructed that large words would be easier to recall, the small words were recalled less often on both lists. This pattern was reversed in the opposite belief condition. To test these apparent effects, a 2 (instruction) x 2 (list) x 2 (font size) mixed ANOVA, where the belief instruction was between subjects, was performed. The three-way interaction between font size, list, and belief instruction was not significant, $F(1, 60) < 1, \eta_p^2 = .002$, $MSE = 84.14$.

The simple two-way interactions were non-significant for both the font size and list [$F(1, 60) < 1, \eta_p^2 = .002$, $MSE = 84.14$] and the list and instruction pairings [$F(1, 60) < 1, \eta_p^2 = .007$, $MSE = 146.66$], though the effects of font size do vary across the instruction variable, [$F(1, 60) = 12.62, \eta_p^2 = .174$, $MSE = 97.70, p < .001$]. In the condition where participants were instructed that large fonts are sometimes easier to learn, words in large font ($M = 31.57$, $SD = 15.75$) were recalled more often than words in small font ($M = 27.49$, $SD = 15.82$), $t(30) = 2.57, d = .26, p = .02$. In the opposite instruction condition, the recall scores were reversed and participants recalled a higher percentage of words in small font ($M = 32.72$, $SD = 16.84$) than words in large font ($M = 27.88$, $SD = 16.84$), $t(30) = 2.49, d = .29, p = .02$.

Participants recalled similar amounts of items whether they were in the condition where they believed larger words were easier to learn ($M = 29.53$, $SD = 15.78$) or the condition where they believed smaller words were easier to learn ($M = 30.30$, $SD = 16.69$), $F(1, 60) = .05, \eta_p^2 = .001$, $MSE = 726.53, p = .82$. Lastly, participants recalled a higher percentage of items on average on the first list ($M = 32.38$, $SD = 15.25$) than the second list ($M = 27.46$, $SD = 17.24$), $F(1, 60) = 10.24, \eta_p^2 = .146$, $MSE = 146.66, p = .002$.

Post-task items. Again, to ensure that the instructions and research were believed by participants, a manipulation check was included in the post-task questionnaire. When asked if they had believed the research presented in the instructions at the start of the experiment, 70.97% indicated that they definitely believed the research. Considering each instructional manipulation separately, 83.87% of participants that received instructions about small-fonts providing memory benefits reported believing the research compared to 58.06% of participants in the opposite condition. Also, when asked about their *a priori* beliefs about font-size 62.90% indicated an *a priori* belief that large font is more memorable, 8.06% small font, and 29.03% no bias.

2.1.3.3 Discussion

In this experiment, participants viewed words in large and small font and made JOLs about these words before attempting to recall them. As expected, when participants were given no instruction other than to judge the likelihood that they would recall the information later, JOLs were larger for words in large font than words in small font, either because the fonts caused a difference in processing fluency (Rhodes & Castel, 2008a) or prior beliefs about font-size (Mueller et al., 2014).

We do note that there were slight differences in recall scores in this experiment that differ from prior work with the font-size effect. These differences are puzzling and may be due to a slight demand characteristic of the instruction manipulation. That is, it is possible that participants gave differential levels of effort to memorizing words printed in the "easier" and

"harder" font-sizes. This explanation falls short when considering that these differences were not found in Experiment 1A or 1B, which had a similar manipulation. Further, the difference is present in the recall for the first list, at which point both groups had gone through the same exact procedure. We conclude that these differences must be due to some random noise in the data, and as they are small differences on the order of 1-2 words, we believe they do not present any complications when considering the more central JOL effects.

Though the processing fluency did not differ between lists one and two of this experiment, there was a significant effect of the instructions given before the second list. For participants that were given information suggesting that words in large font would be easier to recall than those in small font, the font-size bias in JOLs was shows for words in large font across both lists. In contrast, for participants who were given information suggesting that words in small font would be easier to recall, there was only an effect of font-size on JOLs on the first list but no difference in JOLs on the second list. It is hypothesized that this counter-belief, counter in that it is the opposite of what participants have indicated they believe prior to these types of experiments (Kornell et al., 2011; Mueller et al., 2014), is competing with the effects of processing fluency. We suggest that when participants see a large font word and give a JOL, they are giving the sum of the effects of fluency (positive) and belief (negative). This is supported by the results in the condition where the information given was in concert with prior beliefs that large font words are easier to learn and participants showed a strong font-size effect on their JOLs.

To reiterate a point from the discussion of Experiment 1B, a pure-belief driven account of JOLs would predict that if participants believe the instructional manipulation and update their JOLs accordingly, JOLs and font-size would have a strong positive correlation in the bigger-is-better condition and a strong negative correlation in the smaller-is-better condition. Here, participants all showed a small-to-moderate correlation between JOLs and font-size on list one which is concordant with the font-size bias. Participants in the bigger-is-better retained this bias at list two, but participants in the smaller-is-better condition showed a negligible correlation. This shift in correlations suggests that participants have changed their

beliefs about the effects of font-size on learning. There is no reversal to a negative correlation which leaves three possible interpretations: participants discarded their beliefs about font-size entirely, participants changed their beliefs and failed to update their judgments fully (similar to Mueller et al., 2015), or they changed their beliefs and those beliefs are now in competition with perceptual fluency.

One important consideration regarding the critical manipulation of the instructions is that there was no independent measure of the impact of instructions. Conclusions can only be made regarding the differences in the instruction sets, and readers should be cautious in making assertions about the impact of instructions alone. Future studies concerning paradigms of this sort should consider a control group that does not receive any instructions. This control group would allow for a cleaner analysis of metacognitive changes as a result of both task experience and instruction set content.

In Experiment 1A and Experiment 1B participants did not show any changes in metacognitive resolution scores across any variable. An increase in resolution would reflect either changes in JOL production leading to more accurate judgments, similar JOL production along with changes in recall that better reflect judgments, or some combination of both. Overall in this experiment there were no significant changes in JOLs across any of the variables, though it may be of interest to consider the marginal trends in the data.

There was a trend toward decreased metacognitive resolution across lists, which is probably best explained by the apparent reduction in resolution from list one to list two for participants in the smaller-is-better condition. These participants gave less accurate JOLs for words in small font on list two than they did on list one. This is in contrast with their performance on large-font words, and with participants in the opposing condition, both of which showed no trends. The reduced resolution was accompanied by decreased JOLs for words in small font from list one to list two, along with decreased recall performance for words in small font on list two. Together, these trends may suggest a tendency for participants to give less effort to memorizing words in small-font because they believe they will be easier to recall. If so, this would not only support the notion that beliefs about font-size

influence JOLs but that they also influence the memory task performance.

Concerning the relevant manipulation checks from the post-task questionnaire, it is somewhat surprising to see that only about 58% of participants reported believing the instructional manipulation when the research presented indicated that large-fonts are more memorable. These participants produced a font-size effect in their data, nonetheless, suggesting that they were unaware of how the font-size was affecting their judgments (consistent with findings from Frank & Kuhlmann, 2017). This contrasts with the high acceptance of the instructional manipulation in the opposing condition (about 84%). It is possible that this large difference in acceptance rates is due to the difference in the relationship between the presented information and prior beliefs about font-size and memory. That is, we speculate that participants are more likely to accept research that directly contrasts with their own beliefs.

2.1.4 General Discussion

The literature strongly supports an effect of font-size on JOLs which is not reflected in recall, and this effect is replicated here in several experiments. The mechanism driving this effect is in debate currently and there are two competing hypotheses. The fluency hypothesis argues that increased processing fluency, which is inherent in more perceptually fluent items, leads to a more familiar, more accessible representation in working memory. A more accessible representation can lead participants to judge more perceptually fluent items as better remembered, though their judgments are often wrong because perceptual fluency is not necessarily a useful cue. The belief hypothesis argues that participants make their judgments based almost solely on belief, and that it is belief about large font words, and larger font in general, which is driving the font-size effect. To test these hypotheses against one another, the experiments presented here attempted to manipulate belief while holding fluency constant. In this way, any differences in JOLs would be driven by belief.

In Experiment 1A, participants were presented with evidence confirming the font-size

effect before study began and between the first and second study list. It was expected the instruction manipulation would strengthen participants' pre-existing beliefs in the font-size effect and that their JOLs would reflect this strengthened belief. Though the basic finding was replicated and large-font words were rated as more memorable than small font words, there were no differences in the effect across instruction. These findings suggest that either manipulation did not strengthen the belief, the belief was not strengthened appreciably, or the font-size effect is not mediated by belief. Taken alone, the results of Experiment 1A provide no direct confirmation or refutation of either the fluency or belief hypotheses but do show that the effect persists across conditions.

When presented with the information that small-font words are more easily remembered than large-font words (1B), participants showed a markedly reduced difference between their JOLs for words in large-font and words in small-font. Importantly, it was only the introduction of an incompatible belief that led participants to give similar JOLs to items in different font-size, a behavior which has not been observed in past studies. However, the effects found in Experiment 1B do not necessarily refute the fluency hypothesis.

If JOLs were driven entirely by belief, participants who believed that words in smaller font should have given those items larger JOLs. In the raw data for Experiment 1B, a negative difference between large and small font words was only found in 3 of the 79 participants; it was much more common to see only slight one-to-two-point differences in JOLs between the font-sizes. As these results further showed, participants did not systematically alter JOLs according to font-sizes, suggesting that perceptual fluency may still be significantly influencing JOLs. JOL construction might be represented in summation sense where base memory, (e.g. if an item has repeated presentations or longer presentation times), item effects (e.g. familiarity, word frequency, or personal relevance), fluency, and beliefs are related:

$$\text{JOL} = [\text{BaseMemory} + \text{ItemEffects}] + [\text{Fluency}] + [\text{Beliefs}]$$

, where beliefs would encompass beliefs about memory capacity, the effects of font-size,

etc. In the present study, base memory and item effects were not manipulated, but prior studies have shown direct effects of these factors (cf. Rhodes, 2015). In this model, any negative belief would counter the positive effects of fluency, and any positive effects of fluency would mitigate negative beliefs. This equation is, of course, likely an over-simplification, but represents an attempt to better articulate the multiple bases for JOLs, how they interact, and can be altered. Additionally, it could be argued that the current set of studies does not unambiguously show a direct effect of belief on JOLs; though this study shows some evidence that belief has a tangible effect on metamemory and previous studies have shown that there must be some effects of belief, they may only be mediated effects or otherwise indirect. What is clear though is that belief is in some way affecting JOLs, and when beliefs are changed or challenged JOLs similarly change. This is consistent with recent research showing that participants' JOLs are affected by novel beliefs that were not held prior to the experiment (Mueller & Dunlosky, 2017).

A counterargument to this hypothesis is that participants are simply exhibiting demand characteristics. That is, because the instructions include information about the relative sizes of the words in the subsequent lists, participants are catering to the experimenters' expectations when making their judgments. It is difficult to disentangle any demand characteristics in these data, however there is some evidence to show that participants are responding in earnest. Namely, there was no reversal of the font-size effect—that is, if participants were responding entirely based on demand characteristics we should expect a pattern in Experiment 1B and Experiment 2 such that smaller words are being rated as more memorable. Instead the effect is null. Further, if participants are guiding their beliefs using the information given in the instruction sets, this is simply further evidence that belief is a main component in the formation of JOLs. This new information should be used to inform JOLs as it is presumably more diagnostic (as a piece of researcher coming from a scientist) than other naïve heuristics like perceptual fluency. Again, this pattern of responding is not fully endorsed as we do not see a reversal of the font-size bias.

It appears as though the font-size effect is not being driven solely by belief or solely

by fluency, but rather that there are additive effects of both factors. Past research has strongly suggested that fluency must be playing a role in metacognitive monitoring (Baddeley & Longman, 1978; Begg et al., 1989; Benjamin, 2003; Kelley & Rhodes, 2002; Koriat & Bjork, 2005; Koriat & Ma'ayan, 2005; Rhodes & Castel, 2008a, 2009). Mueller et al. (2014) challenges whether it is in fact fluency that is mediating those effects or merely beliefs about fluency. The current study seems to suggest that though beliefs play a large role, there is still a residual effect of fluency. However, it remains unclear whether fluency is a direct factor or if it is instead beliefs about fluency driving these effects. That is, the influence of perceptual fluency may be due to a direct impact of fluency, or alternatively, it may be mediated by participants' beliefs about fluency with little to no direct effect of fluency. It is plausible that that instead of fluency, there are competing beliefs about how font-size may affect memory: prior beliefs about positive effects of large fonts on memory (Kornell et al., 2011) and the information about font-size and memory given by the instruction sets in these experiments.

Additionally, participants may simply not be fully implementing any new beliefs. They may fully believe that the smaller font words are easier to remember but not fully adapt their heuristics in such a short time. This line of reasoning resonates with recent research showing that participants do not fully-update their strategies for the formation of JOLs after learning from their experiences (Mueller et al., 2015; Tullis, Finley, & Benjamin, 2013). The aforementioned research (Mueller et al., 2015) also highlights the effects of scaling artifacts over time, which are likely present to some extent in this experiment as well—as JOLs drop on list 2 the effects of belief may be hidden. A lack of complete knowledge updating coupled with the possible presence of these kinds of scaling artifacts certainly casts doubt on any direct effects of perceptual fluency reported here.

If it were entirely the effects of belief, the effects of the currently study might simply be explained by altered, and not replaced beliefs: participants may not have switched over to the new counter-belief presented to them, rather they partially adjusted current beliefs. The current set of experiments cannot speak to this alternate explanation directly, though we speculate that the strong acceptance of the belief-instruction in Experiment 1B (about 84%)

coupled with the lack of reversal of the font-size effect is compelling evidence. Despite our speculations, to understand whether this is a direct effect of perceptual fluency, it would be useful to have clear measures of perceptual and processing fluencies which would allow for the evaluation of their contributions. Without an independent measure of fluency, it is difficult to make strong claims as to whether these effects regarding fluency are distinctly different from other research suggesting they are instead effects of beliefs about fluency (Mueller et al., 2014).

Another approach to clarify this ambiguity is to focus on some key individual differences which are known to affect beliefs. For example, Miele et al. (2011) showed that differences in theory of intelligence (Dweck, 1999) predict differences in whether or not the "easily learned; easily remembered" (ELER) heuristic (Koriat, 2008) is used. Dweck (1999) has shown that there are different views on how intelligence can be increased or decreased. Incremental theorists, or growth theorists, have the view that a person can work hard and improve one's own intelligence. Fixed theorists on the other hand feel that intelligence is something you are born with and it cannot be changed. Fixed theorists tend to be more susceptible to using heuristics, and Miele et al. (2011) has specifically shown that fixed theorists are more likely to use the ELER heuristic.

Heuristics, though more or less automatic processes, reflect participants' beliefs about certain situations. In theory, individual differences on theory of intelligence should yield differential effects of belief such that fixed theorists show much greater effects of belief than incremental theorists. If the effects of fluency are in truth effects due to beliefs about fluency, people who are closest to the incremental end of the theory of intelligence scale would likely display minimal differences between any factors, be they belief or fluency.

In conclusion, the perceptual characteristics of words certainly have an effect on how people judge their learning which is evident across many studies (see Luna et al., 2017, for a meta-analysis of these effects). In these experiments, we believe that the results show both belief and fluency to be strong cues utilized by people when monitoring their learning. Whether it is perceptual fluency or beliefs about fluency that is a prime mover in the devel-

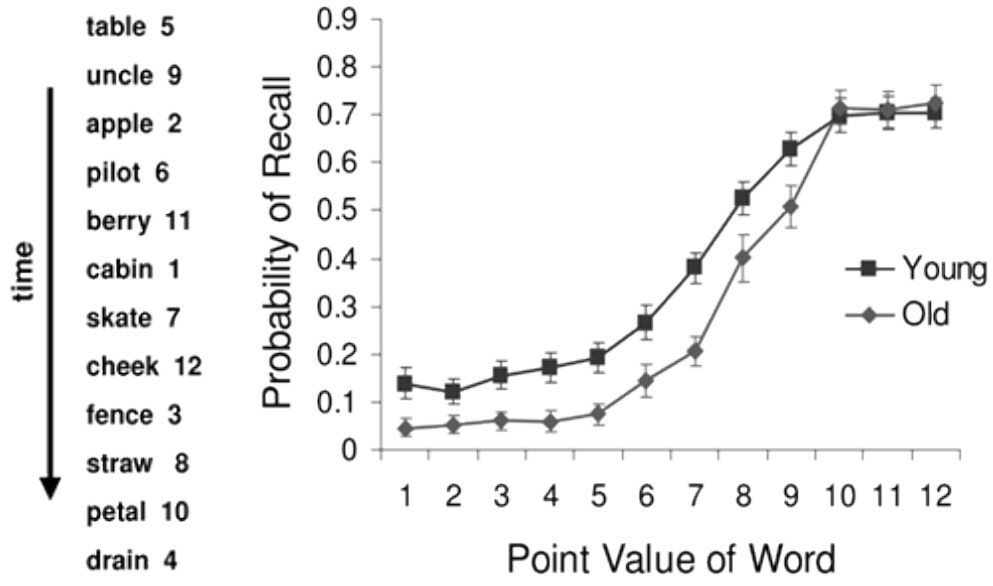
opment of JOLs is uncertain and may be hard to de-couple. Yet this study does show that belief and competition of belief about something as simple as font-size can have a direct and observable effect on an individual's perceived level of learning.

2.2 Study 2

Continuing with this examination of the effects of fluency and belief on JOLs, it is important to understand how different strata of beliefs interact with one another. A common criticism of the font-size effect is that in the absence of any other cues, font-size is as good as any to make a JOL. Words chosen in memory studies are specifically controlled to be similarly recallable, so participants do not have any good diagnostic cues available to them.

A number of studies have shown that participants are sensitive to value effects when learning new information (see Castel, McGillivray, & Friedman, 2012, for a review). The general procedure is depicted in Figure 2.6, panel A. Words are presented to participants one-at-a-time with values attached to remembering the items. Participants are instructed to attempt to maximize the number of points they earn during the study. Panel B shows that younger adults generally perform better than older adults, but both groups are drawn to remembering the high value words, and older adults do just as well as younger adults on memory for the highest value words. Value is thus directing memory in the participants and is a diagnostic cue for memory.

Given these effects on memory combined with research showing that participants are metacognitively sensitive to value effects on memory (McGillivray & Castel, 2011), the current study asks whether prior beliefs regarding font size influence subsequent memory and JOLs in the face of a more diagnostic cue.



(a) Paradigm *(b)* Results

Figure 2.6. The selectivity procedure (a) and results (b) from the selectivity paradigm [Figure adapted from Castel, McGillivray, and Friedman (2012)].

2.2.1 Method

Participants and design. Amazon Mechanical Turk (MTurk) was used to recruit 88 participants (41 female, $M_{age} = 36.07$) for this study. Participants were paid approximately \$6 / hr. as compensation for their participation. This study utilized a 3 (value-framing) x 3 (font-size: large, medium, small) mixed subjects design, where value was manipulated between-subjects and font-size was manipulated within-subjects. The participants were approximately equally split across the between-subjects groups.

Four other participants completed the experiment but their metacognitive data was corrupted and unreadable, so they have been excluded from all counts and analyses.

Materials. Value framing was manipulated between-subjects across three levels: congruent, incongruent, and control. In the congruent framing condition, participants were told that the importance the words in the study was positively correlated with the font-size they were presented in. This positive correlation is congruent with the general belief that partici-

pants hold about the relationship between font-size and memory: larger fonts are thought to be easier to remember (Kornell et al., 2011). In the incongruent condition, participants were instead instructed that the words in smaller fonts were more important to remember. The control condition did not include any instructions about the value of remembering any category over another. In all conditions, participants were told that they should try to remember as many words as they could for a later test.

Font-size was manipulated across three levels: small, medium, and large. Small words were assigned to be presented in 16 pt font, medium words in 32 pt font, and large words in 48 pt font. These font-sizes are the same sizes used in Rhodes and Castel (2008a).

The word list used was a set of 90 words pulled from the English Lexicon Project (Balota et al., 2007) with a mean log hyperspace analog frequency of 8.68 and an average word length of 5.21 characters. For each participant, these words were randomly assigned to three separate lists of 30 words each. Within these subsets, the words were grouped into 10 groups of three words. Within these groups of three, the words were randomly assigned to one of the three font-size conditions, such that each font-size was represented once within each group. This ensured that the font-sizes were equally distributed throughout the list in a pseudo-random pattern.

Procedure. Participants logged into the study with their MTurk worker IDs from a link provided to them through the MTurk interface. After logging in, participants were randomly assigned to one of the value-framing conditions, with the constraint that group membership sizes were roughly equal. The participants were shown the instructions regarding the value-framing and were reminded that they should do their best to remember as many of the words as they could for a later test. Then, the study phase began and the study words were presented one-at-a-time in the center of a white screen, in black font, in the assigned font-size, for 3 s. Following the presentation of each word, participants made a JOL on a scale from 0 (not likely to later recall) to 10 (sure to recall). After studying and judging 30 words, participants played Tetris for 2 min before the test phase began. During the test

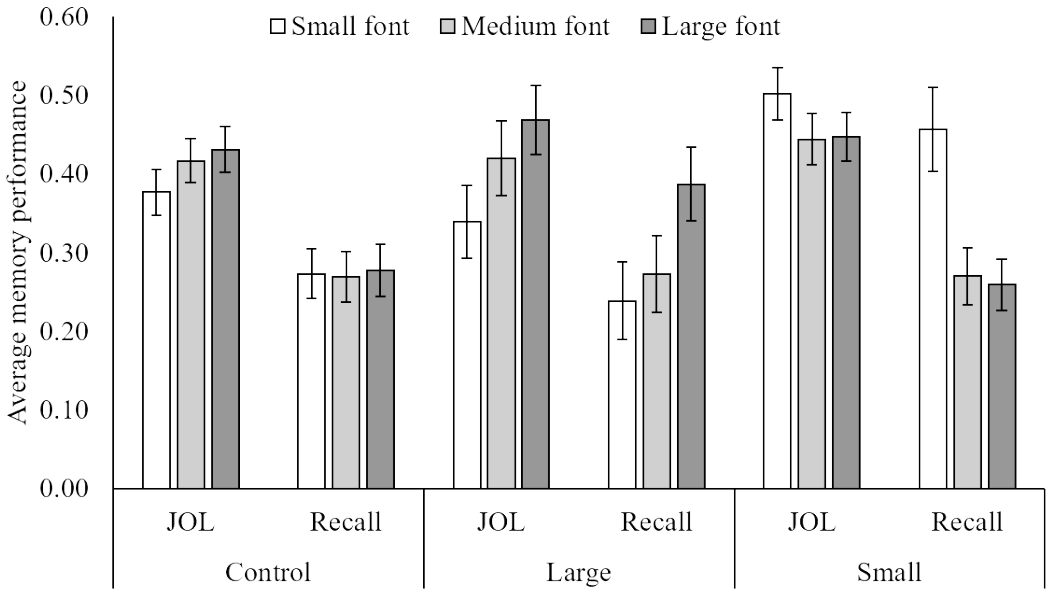


Figure 2.7. The average predicted memory performance (JOL) and actual performance (recall) for words in each font-size condition (small font, medium font, large font) split by the value-framing condition (control, large, small). Error bars represent standard error of the mean.

phase, participants were given 2 min to free recall words from the list by typing them in to a text box. This study-test sequence repeated twice more.

2.2.2 Results

The data were analyzed with a 3 (list) x 3 (value-framing) x 3 (font-size) x 2 (measure) repeated measures ANOVA. First, we looked for any changes in JOLs or recall across lists and found no significant effects of list (all F's < 1). As such, we can consider the data collapsed across list, as displayed in Figure 2.7. Reading the figure, there appear to be dramatic changes in the patterns of judgments and recall across both the value-framing conditions and the font-size conditions, and indeed this interaction was significant, $F(4, 85) = 3.95, \eta_p^2 = .09, p < .01$. To further explore the three-way interaction, we analyzed the patterns of differences within each value-framing condition via paired samples t tests. These tests are given below, grouped by measure.

JOLs. In the control condition, participants gave significantly different JOLs for words in small-font ($M = 0.38$, $SD = 0.16$) compared to medium-font [$M = 0.42$, $SD = 0.15$; $t(29) = 2.29, p = .03$] and large-font [$M = .43$, $SD = .16$; $t(29) = 2.37, p = .03$], but there was no difference in JOLs between medium-font and large-font, $t(29) = 0.94, p = .35$. In the condition where participants were told that the large-font words were more important to remember, there were significant differences in JOLs for small-font ($M = .34$, $SD = .24$) compared to medium-font [$M = .42$, $SD = .25$; $t(26) = 3.44, p = .002$], small-font compared to large-font [$M = .47$, $SD = .23$; $t(26) = 5.01, p < .001$], and medium-font compared to large-font, $t(26) = 3.44, p = .002$. Lastly, in the condition where participants were told that the small-font words were more important to remember, there were no significant differences between judgments for small-font ($M = .50$, $SD = .19$) compared to medium-font [$M = .44$, $SD = .18$; $t(30) = 1.67, p = .10$], small-font compared to large-font [$M = .45$, $SD = .17$; $t(30) = 1.66, p = .10$], or medium-font compared to large-font, $t(30) = 0.16, p = .87$.

Recall. Unlike JOLs for the control condition, recall was not different between small-font ($M = .27$, $SD = .17$) and medium-font [$M = .27$, $SD = .18$; $t(29) = 0.20, p = .84$], small-font and large font [$M = .28$, $SD = .18$; $t(29) = 0.22, p = .83$], nor medium-font and large-font, $t(29) = 0.42, p = .68$. However, in the condition where large-font words were more important, large-font words ($M = .39$, $SD = .24$) were remembered more often than both small-font words [$M = .24$, $SD = .26$; $t(26) = 5.01, p < .001$], and medium-font words [$M = .27$, $SD = .25$; $t(26) = 4.26, p < .001$], though small-font words were remembered as often as medium-font words, $t(26) = 1.75, p = .09$. Similarly, when small-font words were more important to remember, small-font words ($M = .46$, $SD = .29$) were recalled more often than medium-font words [$M = .27$, $SD = .20$; $t(30) = 4.04, p < .001$], and large-font words [$M = .26$, $SD = .18$; $t(30) = 3.97, p < .001$], though there was no difference in recall between medium-font words and large-font words, $t(30) = 0.58, p = .57$.

2.2.3 Discussion

In this study, font-size and value were both manipulated to observe the effects they have on memory and metamemory. Over three between-subjects conditions, participants were instructed that either large-font words were more important, small-font words were more important, or were given no instructions on importance of words. As we might expect, in the control condition were participants were not given any instructions regarding the importance of one category of words or another, participants' responses displayed the same pattern of results as found in other studies on font-size bias (Blake & Castel, 2018; Mueller et al., 2015; Rhodes & Castel, 2008a). That is, words in larger fonts were rated as better remembered, but no differences were found in recall. The more interesting findings are in the value-framing conditions.

In the condition where large-fonts were framed as valuable, Figure 2.7 shows an exaggeration of the font-size bias in JOLs: the separation between the means becomes more apparent. Further, as value-directed remembering research would predict, there is a font-size bias in the pattern of recall such that words in large-font are recalled more often than others. This pattern in JOLs and recall is exactly what one would expect given the task instructions; when participants are told certain items are more important, they give more effort to learning those items at the exclusion of others, and their JOLs reflect that.

A belief-based hypothesis of JOL-generation would predict similar effects in the pattern of data for the condition where small-fonts were framed as more valuable. However, this was not the case. The recall data matches well with what we would expect from the value-directed remembering literature, where the more valuable small-font items are better remembered than the less valuable larger-fonts. Yet there were no differences among the JOLs. Looking again at Figure 2.7, it appears that there is a hint of a pattern in the JOLs, however, the amount of variability in the responses was very high compared to the other conditions.

So, what is happening in the small-font value-framing condition? It appears as though we again see competing effects of belief and fluency, as in Blake and Castel (2018). Participants'

JOLs for larger font words remain high, likely due to some effect of the perceptual qualities of the stimuli, and the JOLs for the small-font words are increased due to the effects of the value-framing.

Lastly, it appears that the unimportant items are being treated differently across groups. In the large-font value-framing condition, the JOLs for small-font items appear to be slightly suppressed. However, in the small-font value-framing condition there is an increase, if any change, in the JOLs for the unimportant large-font items. Again, this appears to point to an underlying effect of the perceptual fluency of the items.

An alternative account of these data would argue that the fluency effects I have been describing are instead the effects of competing beliefs. It is likely safe to say that participants believe that if they give more effort to learning an item, it will be learned better. In the small-font value framing condition, participants are told that the small-font items are more important and likely give those items more effort and consequently higher JOLs as well. Additionally, the vast majority of participants have prior notions about the relationship between font-size and memory, namely that larger fonts are easier to remember (Kornell et al., 2011; Mueller et al., 2014). Thus, participants may be using the more diagnostic cue of "effort given" for the small font items, but are still being duped by the effects of their faulty beliefs for the larger-font items.

This study presents a similar view as Blake and Castel (2018) in regards to the competing belief-based and fluency hypotheses. This view is supported by research with ease-of-learning judgments showing that participants are sensitive to processing fluency in the case of words in alternating case (hElLo) but that this fluency is moderated by beliefs about how it will affect learning (Jemstedt, Schwartz, & Jönsson, 2017). It seems that perceptual fluency plays some role, though it is yet unclear whether it is a direct effect of fluency that is contributing to differences in JOLs or whether it is mediated by beliefs about fluency.

In Mueller et al. (2014) lexical decision time is used as a measure of processing fluency. The study found no differences between how fast words in small-font and words in large-font

were processed, though using lexical decision times may not be the best measure of fluency. One issue with using the lexical decision task is that it only measures how fast the word is read and judged as a word. It does not measure how long it takes a participant to try to learn a word, nor does it provide an approximation of the perceived difficulty of learning the word; the task only measures how fast participants can recognize that the item is a word or not.

To account for the fact that lexical decision is not a perfect measure of fluency, Mueller et al. (2014) utilized a second measure of fluency used in a separate experiment: study time. Presumably participants will devote more effort to items that are harder to learn. This idea is in-line with research suggesting that participants' learning follows a discrepancy reduction model wherein more difficult items are allocated more study time (Thiede & Dunlosky, 1999). It would follow that if larger-font words are perceived as easier to remember, participants would subsequently spend less time studying those words. The study found no evidence of study time differences and concluded that participants must not be making their JOLs based on any difficulties inherent in learning words in one size font over another.

One thing that Mueller et al. (2014) fails to consider is that participants may be approaching "difficult" items in different ways. Another model of study time suggests participants seek items that are in their region of proximal learning (Metcalfe & Kornell, 2005). Though participants following a discrepancy reduction model might be expected to spend more time on items that are perceived to be harder, others may simply give less attention to these items. As the font-size effect generally describes aggregate effects, individual differences in study approaches may be averaged out and hidden in the data.

One instance where individual differences have been shown regarding processing fluency is with the easily-learned easily-remembered heuristic and theories of learning (Miele et al., 2011). Miele et al. (2011) showed that differences in theory of intelligence (Dweck, 1999) predict differences in whether or not the "easily learned; easily remembered" (ELER) heuristic (Koriat, 2008) is used. Dweck (1999) has shown that there are different views on how intelligence can be increased or decreased. Incremental theorists, or growth theorists, have

the view that a person can work hard and improve one's own intelligence. Fixed theorists on the other hand feel that intelligence is something you are born with and it cannot be changed. Fixed theorists tend to be more susceptible to using heuristics, and Miele et al. (2011) has specifically shown that fixed theorists are more likely to use the ELER heuristic.

Heuristics, though more or less automatic processes, reflect participants' beliefs about certain situations. In theory, individual differences on theory of intelligence should yield differential effects of belief such that fixed theorists show much greater effects of belief than incremental theorists. If the effects of fluency are in truth effects due to beliefs about fluency, people who are closest to the incremental end of the theory of intelligence scale would likely display minimal differences between any factors, be they belief or fluency.

Clearly, there are many different factors at play when considering the driving forces in the font-size effect. When these factors are taken in isolation it is easy to focus on the items that seem to address one issue or another, but it would appear that they should be taken together. By modeling the interplay between these factors, we can begin to see a clearer picture. This experiment attempts to do so by measuring participants' memory and metamemory for words in different-sized fonts, assessing their theories of intelligence, taking multiple measures of fluency and belief, and modeling how these factors influence one another to generate JOLs.

CHAPTER 3

Differences in Memory and Metamemory Across Domains

Learning occurs in many different domains, yet to this point I have focused on learning arbitrary word lists. It has been argued that the font-size effect may only be occurring in these types of materials because there is a dearth of diagnostic cues. The words used are chosen to be relatively infrequent yet common enough that they are not "difficult" words, they are presented at the same rate, and the legibility of the words is controlled.

In the absence of better cues than font-size, participants use the only source of variation they perceive. This is not a complete explanation, as participants have responded that they believe font-size can and does impact memory (Blake & Castel, 2018; Kornell et al., 2011; Mueller et al., 2014). However, in these studies most other factors *other than* font-size are controlled away, which may explain the particular attention to font-size. Given more diagnostic cues, participants may marginalize the effects of task fluency and instead make decisions based on better heuristics. This idea was touched on in Experiment 3 of Rhodes and Castel (2008a) where perceptual fluency was manipulated along with associative strength—a cue which can be used to predict memory quite well (Koriat & Bjork, 2005; Koriat et al., 2004). When studying related and unrelated pairs in small and large font, participants were much less sensitive to the effects of perceptual fluency (Rhodes & Castel, 2008a), suggesting that they gave more credence to associative strength, but were still influenced by fluency. Even in these cases however, it is clear that the word pairs and perhaps the pairings themselves are arbitrary, and there is no true motivation to remember them.

Instead, it is important to generalize these findings to materials that have clearer, more

salient motivations for truly learning the pairs involved. Two such domains are foreign language learning and medical information learning. In the following studies, I consider how learning medications and side-effects is a non-trivial task akin to learning a new language, and that it in fact may be more difficult than foreign-language learning.

3.1 Study 3[1]

In this study I will be focusing on the similarities and differences between learning Lithuanian – English translations and medication – side-effect pairings. First, it is important to consider why medication learning is important, and how it compares superficially to foreign language learning.

Medication usage in the US is widespread and has grown substantially over the last few years (Catlin et al., 2008). Though a lot of this growth and use of prescription drugs has been centered in elderly populations (e.g. Qato et al., 2008), medication usage is common across the lifespan. Additionally we are in the midst of an "opioid epidemic" which has developed from prescription drug misuse by people of all ages (Van Zee, 2009).

Understanding how people learn information about the drugs they take is certainly a core aspect to understanding drug use. Ley's model (Ley, 1988) is a prevalent model for communicating with patients about medication usage. Under the model, memory is a core component, but even so, people forget a very large portion of what is learned in the doctor's office almost immediately (Kessels, 2003). This is a substantial problem, and medical information has been shown to be surprisingly resistant to learning, even under good learning conditions like proper organization (McGuire, 1996).

Paired-associate learning (PAL) is a common tool for assessing cognitive and metacognitive strategies. In the basic PAL paradigm, participants study pairs of words one-at-a-time and a test they are only shown one half of the pair and asked to recall the other half. When

[1] Portions of Experiment 1 and some pilot studies for Study 3 were developed in collaboration with Mary B. Hargis.

using a PAL paradigm with foreign language materials, a foreign word is generally paired with its English translation. The foreign word is generally presented on the left, with the familiar, English translation on the right. Considering that many drugs have uncommon characters, bigrams, and pronunciation, there are some clear parallels to learning a foreign language. If one is learning what symptom a drug treats and comparing this to foreign language learning, the unfamiliar drug name is analogous to the foreign word and the more familiar symptom is analogous to the English associate.

Using this type of paradigm (PAL) to study drug information allows us to draw comparisons and contrasts to the greater body of literature on learning in general. This set of studies will assess the factors that make medical information somewhat harder to learn than other materials, and the effects of those factors on metacognitive judgments.

3.1.1 Experiment 1

Learning for Lithuanian translations and medication – side-effect pairings was assessed using a PAL paradigm. As the construction of the paradigm was parallel for each stimulus set (unfamiliar word on the left, familiar word on the right), it was expected that performance would be relatively similar across sets. However, if participants approach the task with different motivations regarding the purpose and utility of learning one set or another, or if the stimuli sets are intrinsically differential in their ease of learning, there may be differences in both memory and metamemory.

3.1.1.1 Method

Participants. For this experiment, 118 students (70 female, $M_{age} = 20.89$) from the University of California, Los Angeles were recruited to participate in this study. Students were recruited from the Psychology Department's participant pool and were compensated with class credit for their participation.

Materials and design. This experiment examined memory and metamemory for two different types of materials: foreign language word pairs and medication – side-effect pairs. Two lists of 26 pairs each were created for this experiment. Each list was comprised of only one type of pairs (translations or medications).

For the foreign word list, 26 pairs of words were pseudo-randomly selected from a Lithuanian – English database (Grimaldi, Pyc, & Rawson, 2010) with the constraint that the ease-of learning Z-scores were close to the mean ($Mz = -.05$).

For the medication – side-effects list, 13 real drug names and 13 fake drug names with 26 moderate side effects taken from an Internet database (Aubuchon, 2015). We attempted to equate the fake drugs and real drugs by asking Amazon Mechanical Turk workers (paid \$0.75 each) to rate 94 different drug (half real, half fake) in terms of how real they seemed and how familiar they seemed. The items used in this study were lowest in familiarity but highest in realness.

All of the word pairs are listed in Appendix A Table A.1.

Procedure. Participants were randomly assigned to either complete the translations or the side-effects study-test cycle first. Participants were instructed that in the study they would be studying word pairs with the purpose of recalling them on a later test. They were informed that in the task there would be two study-test phases where they would be learning either Lithuanian – English translations or medication – side-effect pairs. Further, they were told that each test would only cover the material from the study phase directly preceding it, and that each study list would only contain one type of materials (translations or medications).

After indicating that they understood the instructions, a randomized study list was generated for each participant and participants studied each word pair for 8 s. During study, each pair was presented in black text, on the center of a white background, separated by a colon (e.g. cue : target). The Lithuanian words and the medications were always presented on the left, and the English words and the side-effects were always presented on the right.

After studying each pair, participants were asked to rate how well they thought they would be able to recall the pair at a later test on a scale from 0 (not at all) to 100 (sure to recall).

Immediately after studying all 26 pairs, participants began the test-phase where they were shown the left side of the studied pair (i.e. either a Lithuanian word or a medication) and asked to recall the right side of the pair (i.e. either the English word or the side-effect). Participants were encouraged to guess if they were unsure, and had as long as they wanted to type the words in.

After the first study-test phase, participants were reminded that the materials would be different on the next list, but the task was the same. After completion of the second study-test cycle, participants judged each list on how easy it was to learn (global ease-of-learning; gEOL). Finally, participants answered a series of questions regarding any strategies they may have used and whether there were any problems during the experiment.

Recall was scored by computing the Damerau-Levenshtein edit distance (Brill & Moore, 2000). This distance was subtracted from the total word length, and then divided by the total word length to yield an approximate percent deviation from the correct spelling of the word. The threshold for correctness was set at 70% deviation to allow for minor misspellings.

3.1.1.2 Results

First, the fictitious drugs and real drugs were compared in regards to JOLs and recall performance as a percentage of the number of words recalled. Participants were not sensitive to the distinction between real ($M_{jol} = 34.46, SD_{jol} = 26.67; M_{recall} = .19, SD_{recall} = 0.40$) and fake drugs ($M_{jol} = 34.31, SD_{jol} = 27.06; M_{recall} = 0.18, SD_{recall} = 0.39$). The distinction between fictitious and real drugs will no longer be considered in these results.

For the remaining results, linear mixed effects models (LMEM) were used to analyze the data. There are multiple advantages to using LMEM over traditional analyses like analysis of variance (Baayen, 2008; Baayen, Davidson, & Bates, 2008; Quené & Van Den Bergh, 2004; Quené & van den Bergh, 2008). In particular, many cognitive psychology studies strive to

"control away" the differences in word type, frequency, length, etc., but here the medications and translations are derived as a more random sample from the English lexicon. It is plausible that participants will have differential prior experience with each of the cues and targets, and that this experience will manifest as differences in JOLs and recall scores. Also, LMEM does not require that data be aggregated, which hides important variance among pairs and participants. This is an important consideration for JOLs where each participant may be using the scale in a different fashion. For recall data, which is binary, traditional methods are considered inadequate (Jaeger, 2008; Quené & van den Bergh, 2008).

JOLs. As recommended in Barr, Levy, Scheepers, and Tily (2013), the model was specified as maximally as possible in regard to random effects. JOLs were analyzed using a LMEM with random intercepts for the cues and targets, and random intercepts and slopes for participants within stimulus-type (translation, medication) conditions. The maximal model (random slopes of participants within conditions, and random intercepts for participants, cues, and targets) failed to converge and the random slope factor was dropped.

The model was fit using the 'lme4' package (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2018), using restricted maximum likelihood estimation. The model fit was first analyzed by partitioning the variance explained in to the part explained by the fixed effects, or model ($R^2_m < .001$), and the part explained by the random effects ($R^2_c = 0.39$ Nakagawa & Schielzeth, 2013). Very little of the variance in JOLs was explained by the stimulus-type, however, the significance of the fixed effect was still calculated with the 'lmerTest' package (Kuznetsova, Brockhoff, & Christensen, 2017), using Satterthwaite approximations to degrees of freedom. There was no significant effect of stimulus type on JOLs ($F(1, 147.4) = 0.04, p = .85$), indicating that the groups were no different from the intercept, or grand mean, which was significantly greater than zero ($\beta = 34.14$, $SE = 1.87$), $t(145.6) = 18.22, p < .0001$).

Recall. The same maximal model was fit for recall scores, albeit using logistic regression because recall correctness is a binary measure. The model the was fit included the fixed

effect of stimulus type and the random intercepts for participant, cue, and target.

The model was again fit using the 'lme4' package (Bates et al., 2015), using maximum likelihood estimation via Laplace approximation. Partitioning the variance explained into the amount explained by the fixed effects ($R^2 = 0.03$) compared to the random effects ($R^2 = 0.24$) shows not only that the stimulus type again has lower explanatory power than individual differences in recall, but that there is quite a bit of variance unexplained by the model in general.

Considering the regression coefficients, the intercept ($\beta = -1.76$, $SE = 0.14$), indicates that the odds of recalling a pair in the medications stimulus group was 0.17:1 (found by exponentiating β, which is the log-odds-ratio). A Wald χ^2 analysis of deviance test was performed using the 'car' package (Fox & Weisberg, 2011) to determine if the translations ($\beta_{med.\rightarrow trans.} = 0.74, SE = 0.17$) were significantly different from the medications. The main effect of stimulus type was significant, indicating that the when a pair was in the translations group, participants had 2.10 times greater odds (0.35:1) of recalling it than in the medications group, $\chi^2(1) = 20.20, p < .0001$.

3.1.1.3 Discussion

In this experiment, though both tasks involve binding an unfamiliar word with a familiar associate, participants do not perform similarly on each set of materials. Contrary to predictions that learning would be similar due to task similarity, learning for the Lithuanian translations was better than for medications and their side-effects. JOLs were largely insensitive to learning, though metacognition for the materials was not entirely impaired as global EOLs were very accurate.

There are multiple reasons why the medications and side-effects were learned less well than the translations. One reason may have been motivation; participants may not have been motivated to learn the medications. This could be due to differences in perceived relevance between the translations and medications, namely that learning a new language

seems more important and fun to participants. Alternatively, it could be that participants find the medications particularly irrelevant: "I don't need to learn any of these drugs — I won't ever use them." Further, it is less appealing to learn about side-effects, which are by definition secondary to the nature of the drug which is primarily used to treat an ailment. This question of motivation is considered in Experiment 4.

Another distinction between the stimuli types is the composition of the targets. First, participants may be suffering from the fact that the Lithuanian word pairs include many concrete common nouns and the side effects are all abstract sensations. It is a well-established finding that concrete words are easier to remember than abstract concepts, likely due to the image-ability of the words (Ellis & Beaton, 1993; Paivio, 1986). Second, side-effects are always in the same category: side-effects. As such they are more susceptible to build-up of proactive interference (e.g. Wickens, Born, & Allen, 1963) than are the translation pairs which span multiple categories. Experiment 2 will address this issue by comparing mixed and uniform category translation pairs to the medication pairs.

Global metacognitive judgments nearly perfectly matched memory performance, consistent with other work regarding delayed and global judgments (Besken & Mulligan, 2013; Nelson & Dunlosky, 1991). Generally speaking, participants usually make fairly good metacognitive judgments for paired-associates when they are not being deliberately fooled by experimenters. However, that accuracy is usually fairly well correlated with associative strength, suggesting that associative strength is the salient cue that participants focus on (Koriat & Bjork, 2005). In this case though, the associative strength between all pairs in the current experiment is effectively zero as half of the pair is foreign. Instead, it is likely that the gEOLs are so accurate here because they are made following the test and participants are making their judgments based on their impressions of their average retrieval fluency on each test. This is consistent with previous literature on delayed JOLs (Nelson & Dunlosky, 1991), though generally, delayed JOLs are made on an item-by-item basis. It is impressive that participants in this study were so accurate in judging their performance considering that the judgments were made absent of any explicit feedback, and were made well after the items

were recalled (in the case of the first list, an entire study-test cycle intervened before the judgment was made).

Participants were overconfident overall in their JOLs, a very common finding for prospective judgments, but they appear to be more overconfident for the medications and side-effects. That is, if we consider the mathematical differences between JOLs and recall, participants had a greater discrepancy between recall and JOLs for medications (JOL $\beta_0 = 34.14$ compared to 0.17:1 odds of correct recall) than they did for translations (0.35:1 odds of correct recall). This might suggest that during the task participants are experiencing similar levels of perceived difficulty across materials. Though, as discussed at length in Assessing the Contributions of Fluency and Belief to Immediate JOLs, immediate JOLs are very susceptible to influence from processing fluency and beliefs about the materials. It may be that the JOLs are inaccurate due to a mis-perception of the processing fluency, e.g. the translations were interpreted as harder to process than they were. Alternatively, both sets may have similar levels of processing fluency, and beliefs about the materials are shifting the JOLs to be in line across materials.

A possible solution for teasing apart these effects would be to ensure that the processing fluency for each set of materials is equivalent. Grimaldi et al. (2010) have provided a fantastic set of norms for the Lithuanian word pairs that can be used to match the items in terms of their ease-of-learning and other measures, however the specific medication – side-effect pairings used in these experiments have not been normed along the same measures. It is beyond the scope of this work to develop these types of ratings and materials, however, it is possible to equate the lists experimentally by recombining the pairings, i.e. pairing Lithuanian words with side-effects and medications with the English words (examined in Experiment 3).

3.1.2 Experiment 2

Recall was very poor for medication – side-effect pairs in Experiment 1. A possible explanation for this is that the side-effects are all from a single category, and the English targets in the translation pairs are in multiple categories. A cursory examination of the translations in Grimaldi et al. (2010) yielded seven broad natural categories: nature-related, human-related, activity-related, edible items, clothing, urban-life-related, and household items. The medication targets were all from the same category: common, moderate side-effects.

The semantic relatedness of the targets may pose a problem for recall. Learning items from the same category can be more difficult because learners are more susceptible to build-up of proactive interference (Wickens, 1970, for a review). These types of findings are shown in paired-associate learning as well, and it is suggested that output interference is the cause of lower rates of recall for categorized lists (Roediger & Schmidt, 1980). Categorization and output interference may be a factor in the lower rates of recall with the medication materials.

To test whether the difficulty in learning medication information in Experiment 1 was due to category composition of the targets, the current experiment compares three types of stimuli sets. The same medication pairings are used as a reference set to compare against translations from Lithuanian to either a uniform target category or a mixed target category consisting of three categories. If the target composition has a strong effect on recall, participants should recall more words from the mixed category than the others, and the medications and uniform category should yield similar results.

Additionally, under the cue-utilization framework, it is suggested that JOLs are comparative in nature (Koriat, 1997). It should be very noticeable that one of the Lithuanian sets is all from the same category, and it will be informative to examine their judgments in relation to their JOLs in the other categories and to recall performance.

3.1.2.1 Method

Participants. For this experiment, 64 participants (27 female, M_{age} = 28.6) were recruited to participate in this study via Amazon Mechanical Turk. Participants were told that the study would last approximately 25 min and that they would be paid $2.30 for their participation.

Materials and design. This experiment was conducted using a within-subjects design across one variable, category type, with three levels: medications, uniform, and mixed. For each participant, three separate lists of 18 word pairs each were constructed to satisfy these labels. The order of these lists was counterbalanced.

In the medications list, 18 fictitious medication names were randomly drawn (without replacement) from a list of 30 fictitious medications created for the experiment (see Table A.3 in Appendix A). For each of these medication names, a side effect was randomly selected (without replacement) to pair with the medication.

For the uniform and mixed groups, 36 Lithuanian words were selected randomly, without replacement, from Grimaldi et al. (2010). Letters with inflections or similar markings were replaced with their base counterparts, e.g. "ò" would be stripped to "o". These 36 words were then divided into two 18 word lists that were used as cues in the mixed and uniform category conditions.

The targets in the translation conditions were drawn from four categories: animals, fabrics, fruit, and occupations. These categories were taken from Van Overschelde, Rawson, and Dunlosky (2004). For each participant, a random category was selected for the uniform condition, and 18 words were randomly drawn and paired with the pre-selected Lithuanian words. For the mixed condition, six words were drawn from each of the remaining three categories to yield an 18 word list, which was paired with the remaining Lithuanian cues.

Procedure. Participants were randomly assigned to one of the counterbalanced sets of stimuli. Then, participants were instructed that in the study they would be studying word

pairs with the purpose of recalling them on a later test. They were informed that in the task there would be study-test phases where they would be learning either Lithuanian – English translations or medication – side-effect pairs. Further, they were told that each test would only cover the material from the study phase directly preceding it, and that each study list would only contain one type of materials (translations or medications).

After indicating that they understood the instructions, participants studied each word pair in the first list for 8 s each. During study, each pair was presented in black text, on the center of a white background, separated by a colon (e.g. cue : target). The Lithuanian words and the medications were always presented on the left, and the English words and the side-effects were always presented on the right. After studying each pair, participants were asked to rate how well they thought they would be able to recall the pair at a later test on a scale from 0 (not at all) to 100 (sure to recall).

Immediately after studying all 18 pairs, participants began the test-phase where they were shown the left side of the studied pair (i.e. either a Lithuanian word or a medication) and asked to recall the right side of the pair (i.e. either the English word or the side-effect). Participants given a maximum of 20 s to make each response, but could move on sooner if they typed their answer in.

After the first study-test phase, participants were told that they would be starting list "2 of 3" and that their task was the same. This cycle was repeated again for the third list as well. Finally, participants answered a series of questions regarding any strategies they may have used and whether there were any problems during the experiment.

Recall was scored by computing the Damerau-Levenshtein edit distance and converting it to a percent deviation, as in Pairings used in Chp. 3 Experiment 1.

3.1.2.2 Results

JOLs. As in 3.1.1 the data were analyzed using a LMEM with a maximal random effects structure. The critical fixed effects of category type was regressed on JOL along with random

intercepts of participants, cues, and targets, as well as random participant-condition slopes.

The model was fit using the 'lme4' package (Bates et al., 2015), using restricted maximum likelihood estimation. The model fit was first analyzed by partitioning the variance explained in to the part explained by the fixed effects ($R^2_m = .006$), and the part explained by the random effects ($R^2_c = 0.67$ Nakagawa & Schielzeth, 2013).The significance of the fixed effects was analyzed calculated with the 'lmerTest' package (Kuznetsova et al., 2017), using Satterthwaite approximations to degrees of freedom. There was no significant effect of category type on JOLs ($F(2, 86.19) = 2.77, p = .07$), indicating that the groups were no different from the intercept ($\beta_0 = 34.50$, $SE = 2.93$).

Recall. Like the JOLs, the recall scores were analyzed using a LMEM with the fixed effect of category type, random intercepts for participants and cues and targets, and random slopes for participants along category type. The model was fit using the 'lme4' package (Bates et al., 2015), using maximum likelihood estimation via Laplace approximation. The full model with fixed and random effects yielded a significantly better model fit than the null model with no fixed effects $\chi^2(2) = 28.10, p < .001$). The following analyses thus pertain to the full model.

Tables 3.1 and 3.2 display the the estimates of the random and fixed effects. Note that unlike the JOLs, these estimates are in terms of logits, and represent the log odds ratio. For a more understandable interpretation consider the "Odds" column, which is the exponentiated estimate and can be interpreted as relative change in the odds of correctly recalling a word. For category type, the reference group was the medications, so the intercept represents the coefficient for medications, where the odds of recalling a pair is 0.23:1. Participants had 2.79 times greater odds of recall for the mixed category translations than the medications (0.64:1), and 3.05 greater odds of recall for the uniform category than medications (0.70:1).

To determine which of the fixed effects significantly predicted recall, a Wald χ^2 analysis of deviance test was performed using the 'car' package. The main effect of category type was significant, $\chi^2(2) = 30.00, p < .0001$. To determine which conditions were different from the

Table 3.1

Random effects regression coefficients for Experiment 2

Group	Effect	Variance	Std. Deviation
Cue	Intercept	0.25	0.40
Target	Intercept	0.16	0.50
Participant	Intercept	3.13	1.77
	Mixed	0.86	0.93
	Uniform	0.64	0.80

Table 3.2

Fixed effects regression coefficients for Experiment 2

Fixed Effect	Estimate	Std. Error.	Odds	Z
Intercept (Medications)	-1.47	0.27	0.23	-5.45
Category: Mixed	1.024	0.22	2.79	4.61
Category: Uniform	1.114	0.21	3.05	5.29

others, multiple post-hoc pairwise tests were performed using the 'emmeans' package (Lenth, 2018), with the Tukey method for p-value adjustment. The contrast of the medication category and the mixed category was significant ($z_{ratio} = -4.61, p < .0001$), as was the medication category against the uniform category ($z_{ratio} = -5.29, p < .0001$). The mixed and uniform categories did not significantly differ, $z_{ratio} = -0.51, p = .87$.

3.1.2.3 Discussion

It was expected in this experiment that the category composition would affect recall performance. The medications group replicated the findings from the medication pairs in Experiment 1, and participants showed very similar odds of recalling a pair correctly for those pairs in both Experiments (odds: 0.20:1; probability: 0.16). Despite predictions to the contrary, the uniform category did not yield performance similar to the medications, and participants in fact performed twice as well (odds: 0.64:1, probability: 0.39). Further, the mixed category pairs were remembered just as often as the uniform category pairs (odds: 0.70; probability: 0.41), indicating that the category composition of the targets is categor-

ically *not* affecting recall appreciably. Interestingly participants in this sample performed somewhat *better* for the translations than in Experiment 1, though this is could be due to the difference in population (UCLA vs. Amazon Mechanical Turk).

Despite the strong differences in learning for each of the category types, JOLs did not discriminate amongst the lists. This suggests that, in the moment,s each of the lists felt as hard to learn as the others did. This is consistent with the findings from Experiment 1 where participants did not discriminate between the medications and the translations. Additionally, considering that Roediger and Schmidt (1980) shows that category effects on recall are likely due to output interference, it is unlikely that participants would show sensitivity in their JOLs.

3.1.3 Experiment 3

Considerations of categories notwithstanding, there are other differences in both the cues and targets that may be affecting recall. First, there may be differential difficulty in reading and learning the cues — both are difficult and foreign, yet this may be variable across sets. Second, the translation targets to this point have almost been exclusively concrete common nouns, and the side-effects are much less image-able. Words that are higher in imageability are more easily recalled than abstract words (Paivio, 1986). To assess these issues, the current experiment switched the Lithuanian cues with the medication cues. Using the same set of materials used in the previous studies, each of the Lithuanian cues was randomly paired with the side-effects, and each of the medications was randomly be paired with one of the English words from the Lithuanian set. Each study cycle will be framed to participants as a second language acquisition task.

By framing the task in this manner, there should be relatively few differences in how participants approach the materials, and judgments will be more dependent on the actual materials themselves rather than prior beliefs about the materials. Any persisting differences in recall and JOLs will be more likely to have originated in the differences between the

materials. By equating perceptual fluency it might be possible to more directly compare the remaining effects of beliefs about the materials that may be influencing the JOLs (if any exist that do such). In the case of these materials, the relative amount of uncommon characters (e.g. x, z) and bi-grams (e.g. tv, yp) is likely different across the materials. This likely difference may be causing differences in ease of reading and ease of pronouncing the cues, which may in turn be influencing JOLs. Controlling for these factors would enable a clearer view of what is driving these metacognitive errors.

3.1.3.1 Method

Participants. For this experiment, 34 participants (18 female, M_{age} = 34.6) were recruited to participate in this study via Amazon Mechanical Turk. Participants were told that the study would last approximately 25 min and that they would be paid $2.30 for their participation.

Materials and design. This experiment was conducted using a within-subjects design across two variables, cue type (Lithuanian, medication) and target type (English, side-effect). For each participant, two lists of 20 pairs each were constructed. Each list had five each of the four possible combinations of the experimental levels.

To construct the list, 10 words were drawn randomly without replacement from each of four possible sources: Grimaldi et al. (2010) (Lithuanian), Van Overschelde et al. (2004) (English; either only from animals or only from fruit, category randomly chosen for each participant), Appendix A Table A.2 (medications), and Appendix A Table A.3 (side-effects). All of the words in each of the list were stripped of special characters and converted to lowercase. The pairs were composed into the two above-mentioned lists and then shuffled.

Procedure. Participants were instructed that in the study they would be studying Lithuanian – English translations with the purpose of recalling them on a later test. After indicating that they understood the instructions, participants studied each word pair in the first list for 8 s each. During study, each pair was presented in black text, on the center of a white

background, separated by a colon (e.g. cue : target). After studying each pair, participants were asked to rate how well they thought they would be able to recall the pair at a later test on a scale from 0 (not at all) to 100 (sure to recall).

Immediately after studying all 20 pairs in the first list, participants began the test-phase where they were shown the left side of the studied pair (cue) and asked to recall the right side of the pair (target). Participants given a maximum of 20 s to make each response, but could move on sooner if they typed their answer in.

After the first study-test phase, participants were told that they would be starting list "2 of 2" and that their task was the same. Finally, participants answered a series of questions regarding any strategies they may have used and whether there were any problems during the experiment.

Recall was scored by computing the Damerau-Levenshtein edit distance and converting it to a percent deviation, as in Pairings used in Chp. 3 Experiment 1.

3.1.3.2 Results

JOLs. The data were analyzed using a LMEM with a maximal random effects structure. The critical fixed effects of cue-type and target-type were regressed on JOL along with random intercepts of participants, cues, and targets, as well as random participant-condition slopes.

The model was fit using the 'lme4' package (Bates et al., 2015), using restricted maximum likelihood estimation. This model was compared to a random-effects-only model, where the fixed effects of cue-type, target-type, and their interaction were removed. The full model with fixed effects yields a significantly better model fit than the restricted model $\chi^2(3) = 14.4, p = .003$. The following analyses thus pertain to the full model. See Tables 3.3 and 3.4 for the estimates of the random and fixed effects, respectively.

The model fit was first analyzed by partitioning the variance explained in to the part explained by the fixed effects ($R^2_m = .006$), and the part explained by the random effects

Table 3.3

Random effects regression coefficients for Experiment 3

Group	Effect	Variance	Std. Deviation
Cue	Intercept	8.39	2.90
Target	Intercept	0.00	0.00
Participant	Intercept	394.37	19.86
	Cue Type	1.15	1.07
	Target Type	14.89	3.86
	Cue * Target Type	0.37	0.61

Table 3.4

Fixed effects regression coefficients for Experiment 3

Fixed Effect	Estimate	Std. Error.	df	t
Intercept	23.47	3.50	33.39	6.72
Cue Type (Medication - Lithuanian)	-0.77	1.21	129.65	-0.64
Target Type (English - Side-effect)	4.10	1.24	41.57	3.31
Cue Type * Target Type	-2.62	1.48	911.82	-1.77

($R_c^2 = 0.70$ Nakagawa & Schielzeth, 2013).The significance of the fixed effects was then calculated with the 'lmerTest' package (Kuznetsova et al., 2017), using Satterthwaite approximations to degrees of freedom. There was no significant interaction of cue and target types, $F(1, 911.82) = 3.12, p = .08$. There were significant main effect of cue type, such that medications were given lower judgments than Lithuanian cues, $F(1, 57.34) = 4.72, p = .03$. Also, word pairs that had a non-side-effect target were given larger JOLs, $F(1, 33.19) = 7.57, p = .01$.

Recall. The recall scores were analyzed using the same LMEM as JOLs. The model was fit using the 'lme4' package in R, using maximum likelihood estimation via Laplace approximation. The maximal model did not converge (likely due to low recall scores) and was reduced to a model only including the random intercepts of participant, cue, and target, and the fixed effects. This model was compared to a random-effects-only model, where the fixed effects were removed. The full model with fixed effects yielded a significantly better model fit than the restricted model $\chi^2(3) = 21.3, p < .0001$). The following analyses thus

pertain to the model including fixed effects.

Table 3.5

Random effects regression coefficients for Experiment 3

Group	Effect	Variance	Std. Deviation
Cue	Intercept	0.00	0.00
Target	Intercept	0.202	.045
Participant	Intercept	3.07	1.75

Table 3.6

Fixed effects regression coefficients for Experiment 3

Fixed Effect	Estimate	Std. Error.	Odds	Z
Intercept	-1.38	0.35	0.25	-3.94
Cue Type (Medication - Lithuanian)	-0.08	0.21	0.93	-0.36
Target Type (English - Side-effect)	0.95	0.23	2.59	4.12
Cue Type * Target Type	-0.26	0.29	0.77	-0.92

Table 3.7

Odds ratios for Experiment 3

Condition	Odds
Lithuanian – Side-effect	0.25
Lithuanian – English	0.65
Medication – Side-effect	0.23
Medication – English	0.46

Partitioning the variance into the proportion explained by the random effects versus the fixed effects, the random effects ($R^2 = 0.51$) explain more of the variance than the fixed effects ($R^2 = 0.03$). Tables 3.5 and 3.6 display the estimates of the random and fixed effects. Note that the fixed effects coefficient estimates are given on the log odds ratio (not the response) scale, and the 'Odds' column is the exponentiated estimate. For cue type, the reference group was the Lithuanian words, and for target type the reference was the side-effects. Thus, the intercept represents the coefficient for Lithuanian–side-effect, where the

odds of recalling a pair is 0.19:1. To determine the odds in the other variables, add the estimates and exponentiate, or more simply multiply the odds, i.e. medication–side-effect = $0.25 * 0.93 = 0.23$:1 odds of recall. The odds for each condition are displayed in Table 3.7.

To determine which of the fixed effects significantly predicted recall, Wald χ^2 tests were performed using the 'car' package (Fox & Weisberg, 2011). As with the JOLs, only target type had a significant impact on recall such that pairs with English targets had higher odds of recall than pairs with side-effect targets, $\chi^2(1) = 20.43, p < .001$. Neither a significant main effect of cue-type [$\chi^2(1) = 2.31, p = .13$] nor interaction [$\chi^2(1) = 0.84, p = .36$] were found.

3.1.3.3 Discussion

In this experiment, I split and recombined the pairings found in Experiment 1. The purpose of this modification was to assess the effects of the cues and targets on recall. Two of the conditions in this experiment replicated the design in Experiment 1, and these conditions showed similar differences in recall: the translation pairs had over twice the odds of being recalled compared to the medications. However, this experiment shows that this is not an effect of the cues, but rather an effect of the targets. The targets appear to be easier to recall in the English group than in the side-effect group. It is very likely the case that this is because the side-effect targets are not concrete, and are probably less image-able.

In terms of metacognition, participants were sensitive to both cue type and target type, giving the Lithuanian cues and the English targets higher JOLs. In the case of the English targets, this was expected, again due to the words being concrete common nouns compared to the more abstract feelings of discomfort. The abstraction, and possibly the imagined discomfort, bring a level of disfluency that may be driving the differences in JOLs there. On the other hand, the difference between JOLs for the Lithuanian compared to medications suggests that the medication cues *do* feel harder to learn. This effect was confounded by the pair type in Experiment 1, but here is assessed independently. The reasons for why these

cues are given higher JOLs is likely due to differences in the commonality of the bigrams and letters that are more commonly used in the medications (ax, xa, yp), and are not found in the Lithuanian pairs or everyday English. This is not psycholinguistically examined here, but the explanation would be consistent with other work on the fluency of medication names. For instance people feel drugs with easier to pronounce names are safer to use (Dohle & Montoya, 2017).

3.1.4 Experiment 4

In this final examination of the difference between PAL in these two domains, I consider the notion of perceived purpose and centrality in learning. In Experiment 1 and Experiment 2 participants were under the assumption that they were learning the most common side-effects for medications. Side-effects, however, are unintended consequences of a medication, and are not supposed to happen every time you take the medication (in most cases). As such, it may be that participants find these materials less important to learn than the translations, which seem to have a greater utility. It is very unlikely that participants are expecting that their knowledge of these side-effects will be useful later in life, especially not for all 18-26 medications, but it is easier to imagine situations where knowing some translations may be useful. This experiment examines the differences in recall when the relationships between the word pairs are described as arbitrary, like a random word paired with a foreign word or a medication paired with a side-effect, and when they are described as central, like a translation or the treatment a medication provides.

3.1.4.1 Method

Participants. In this experiment, 58 participants (27 female, $M_{age} = 33.8$) were recruited to participate in this study via Amazon Mechanical Turk. Participants were told that the study would last approximately 25 min and that they would be paid $2.30 for their participation.

Materials and design. This experiment was conducted using a mixed-subjects design across two variables, stimulus type (within: translation, medication) and relationship type (arbitrary, central).

For each participant, a translation list was created by randomly selecting 18 Lithuanian words from Grimaldi et al. (2010), and pairing them with 18 words randomly drawn either from the fruit or animal categories in Van Overschelde et al. (2004). The medication list was created by drawing 18 randomly selected fictitious medications from Appendix A Table A.2 and pairing them with 18 randomly selected side-effects from Appendix A Table A.3

For the relationship type variable, the instructions preceding each list were altered to suggest either an arbitrary relationship between the word pairs or a central relationship. Arbitrary relationships were defined as random pairings for translations, and side-effects for the medications. Central relationships were defined as the correct translation for the translations, and the ailment that a medication treats.

For the translations, the following instructions were used:

> During this study phase you will see Lithuanian words appear on the left and [*random English words / their English translations*] on the right. At the test you will just see the Lithuanian word and will be prompted to write in the [*random English word / English translation*].

Similarly, the following instructions were used for the medications:

> During this study phase you will see medications appear on the left and [*the ailments they treat / common side-effects*] on the right. At the test you will just see the medication and will be prompted to write in [*the illness it treats / the side-effect that it was paired with*].

The order of the medication and side-effect lists were counterbalanced, each participant only saw one of each type, and the instruction–list pairings were counterbalanced across participants.

Procedure. Participants were randomly assigned to one of the counterbalanced stimuli sets. Participants were instructed that in the study they would be studying word pairs with the purpose of recalling them on a later test. After indicating that they understood the instructions, participants were shown the critical instructions. Activity was monitored and all participants in the study maintained browser focus (i.e. they did not navigate away from the web page) and remained on the page at least 12 s.

After reading the instructions they studied each word pair in the first list for 8 s each. During study, each pair was presented in black text, on the center of a white background, separated by a colon (e.g. cue : target). Immediately after studying all 18 pairs in the first list, participants began the test-phase where they were shown the left side of the studied pair (cue) and asked to recall the right side of the pair (target). Participants given a maximum of 20 s to make each response, but could move on sooner if they typed their answer in. After the first study-test phase, participants were told that they would be starting list "2 of 2" and given the next instruction, study, and test sets. Finally, participants answered a series of questions regarding any strategies they may have used and whether there were any problems during the experiment.

Recall was scored by computing the Damerau-Levenshtein edit distance and converting it to a percent deviation, as in Pairings used in Chp. 3 Experiment 1.

3.1.4.2 Results

Recall. To analyze the effect of stimulus type and relationship type, a LMEM was constructed with the fixed experimental factorial effects, along with random intercepts for participants and cues and targets, and random slopes of participants within experimental conditions. This maximally specified random effects model failed to converge, and was reduced to the fixed effects and random intercepts. The model was fit using the 'lme4' package (Bates et al., 2015), using maximum likelihood estimation via Laplace approximation, with the logit link.

First, the model was evaluated against a null model, void of the critical fixed effects, and fit the data significantly better, $\chi^2(3) = 15.70, p = .001$. Next, the variance was partitioned (Nakagawa & Schielzeth, 2013) into the amount explained by the fixed effects ($R^2_m = 0.02$) compared to the random effects ($R^2_c = 0.41$). The regression coefficients for the fixed and random effects are displayed in Tables 3.9 and 3.8, respectively. Note that the fixed effects coefficient estimates are given on the log odds ratio (not the response) scale, and the "Odds" column is the exponentiated estimate. The odds of recall for each group are displayed in Table 3.10.

Table 3.8

Random effects regression coefficients for Experiment 4

Group	Effect	Variance	Std. Deviation
Cue	Intercept	0.12	0.35
Target	Intercept	0.05	0.23
Participant	Intercept	2.04	1.43

Table 3.9

Fixed effects regression coefficients for Experiment 4

Fixed Effect	Estimate	Std. Error.	Odds	Z
Intercept	-1.47	0.25	0.23	-5.83
Stimulus Type (Trans. - Med.)	0.51	0.22	1.66	2.35
Relationship Type (Cen. - Arb.)	-0.35	0.24	0.71	-1.46
Stimulus * Relationship Type	-0.38	0.33	0.68	-1.17

Table 3.10

Odds ratios for Experiment 4

Condition	Odds
Medication – Arbitrary	0.23
Medication – Central	0.16
Translation – Arbitrary	0.38
Translation – Central	0.19

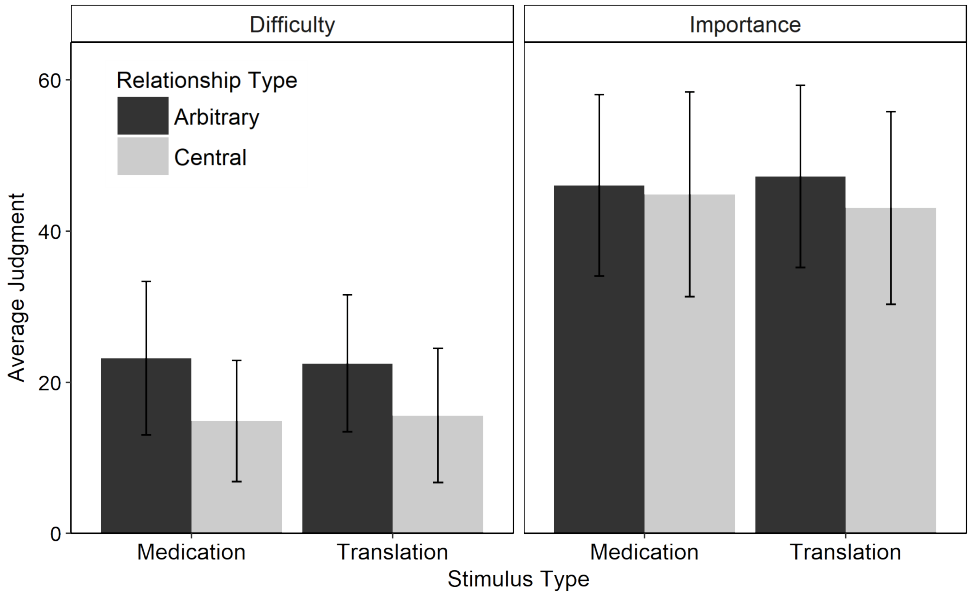


Figure 3.1. Mean judgments of importance and difficulty for pairs in Experiment 4, divided by stimulus type and relationship type. Error bars represent 95% confidence intervals.

A Wald χ^2 analysis of deviance test was performed using the 'car' package (Fox & Weisberg, 2011) to test the fixed effects. The main effect of stimulus type was significant $[\chi^2(1) = 4.64, p = .03]$, as was the main effect of relationship type $[\chi^2(1) = 10.03, p = .002]$[2], though there was no significant interaction between the two, $\chi^2(1) = 1.37, p = .24$.

Metacognition. A separate LMEM was constructed to analyze each of the judgments (importance and difficulty) made after the final recall task. The models included the fully crossed fixed effects of stimulus type and relationship type, as well as random participant intercepts to account for differences in scale usage. The models were fit using the 'lme4' package (Bates et al., 2015). The overall pattern of results from this analysis is displayed in Figure 3.1.

The significance of the fixed effects was analyzed calculated with the 'lmerTest' package

[2] Astute readers might note that two of these conditions replicate the paradigm from Experiment 1: 14 participants saw medication – side-effect (arbitrary) and Lithuanian translation (central) pairs. The pattern of results here is opposite those from Experiment 1—medications were remembered better than translations—but within these 14 participants there were no differences. The lack of differences is likely due to sampling variation with this small subset of participants, and did not affect the overall pattern shown in this experiment.

(Kuznetsova et al., 2017), using Satterthwaite approximations to degrees of freedom. In the importance judgments, there were no significant effects of stimulus type [$F(1, 55.2) = 0.01, p = .92$], relationship type [$F(1, 67.8) = 1.66, p = 0.2$], nor an interaction [$F(1, 69.6) = 0.79, p = .38$]. Thus, the importance judgments are no different than the intercept ($\beta_0 = 49.63$, $SE = 5.12$).

For the difficulty judgments, there was again no main effect of stimulus type [$F(1, 54.3) = 0.04, p = .84$] nor an interaction [$F(1, 61.2) = 0.52, p = .47$]. There was a significant main effect of relationship type such that the arbitrary relationships were judged more difficult to remember ($M = 21.32$, $SE = 3.18$) than the central relationships ($M = 16.44$, $SE = 3.19$), $F(1, 60.8) = 5.70, p = .02$.

3.1.4.3 Discussion

It was expected that in this study recall would be sensitive to how the relationship between the words in each pair was framed. That is, for relationships describing a more central or primary purpose, it was expected that memory would be better, however, the opposite was found. Participants were more likely recall items framed in an arbitrary relationship (random English word, or side-effect) rather than the central relationships (translation, treatment). At face value, this is somewhat surprising. Research on associative strength and memory in PAL would suggest that the opposite is true, i.e. that words with greater associative strength are better remembered (Carroll, Nelson, & Kirwan, 1997; Koriat, 1981). It could be argued that the arbitrarily framed words are less associated, however this is simply *perceived* association, rather than actuality.

These data suggest that perceived associative strength has the opposite effect on learning than true strength. Past research has shown that even weakly associated pairs are much better remembered than those with arbitrary relationships (Koriat & Bjork, 2005). However, Koriat and Bjork (2005) only considers the the broad effects and direction of association, but not differences within pairs with null associations.

Considering the metacognitive post-test judgments sheds some light on the current findings. Participants rated learning of each list as equally important regardless of condition, but they did show differences in what they felt was more difficult. Pairs framed as arbitrarily related were rated as more difficult to learn than the central pairings, which may bely a difference in the amount of effort given to learning each set. In the case of the central pairings, participants rated these pairs as easier to learn which may have led them to give less effort to learning them. Conversely, the arbitrary pairings were perceived as more difficult to learn which may have led participants to engage in more effortful strategies to learn the pairs.

3.1.5 General Discussion

Throughout this study I have shown that medical information is almost consistently more difficult to learn than foreign language vocabulary. In Experiment 1, participants performed significantly worse when attempting to learn the medical materials compared to the translations. Though the side-effects were all from the same category and the translations were varied, Experiment 2 showed that when the categories are made comparable participants still learn the translations better. Some of this effect seems to be explained by the difficulty associated with learning the targets, as the data from Experiment 3 suggests that the translation targets are somewhat better learned. Additionally, some of the variance is explained by the way participants view the pairings: when framed as an arbitrary relationship, participants see the task as more difficult and work harder to learn the pairings (Experiment 4).

Importantly, in the instances where participants made JOLs during the task, these judgments were largely insensitive to learning. In Experiment 1 the average JOL was approximately 34%, did not differ depending on whether the pair was a medication or translation, and overestimated learning in both cases (compare to 0.35:1 odds [25% probability] of recall for translations, and .17:1 odds [14.5%] of recall for medications). Similar levels of overconfidence in JOLs were shown in Experiment 2 as well, along with the insensitivity to experimental condition. The insensitivity and overconfidence is concerning considering the nature of the materials. It is important to understand how these findings generalize to real-

world contexts, as this level of overconfidence could certainly lead to major consequences.

In the case of the translation pairings, Experiment 3 showed that participants both judge and remember the English (non-side-effect) targets better. This is likely due to the fact that these targets are concrete common nouns which are easier to generate mental images for. Perhaps a good metacognitive and mnemonic intervention would be to ask participants to imagine the effects of the side-effect on themselves.

It can be argued learning what medications are used for and the side-effects they may cause is very important compared to learning a new language. This is not supported by the importance ratings in Experiment 4, however it is important to remember that the findings reported here are from learning at home (Amazon Mechanical Turk) or in the lab at school. Judgments of importance may shift drastically depending on context, and people may give more weight to information told to them by their physician. Kessels (2003) shows that regardless of how important people feel that information is, they still forget most of it (up to 80%). The data from the current study shows that poor memory for this information is compounded by overconfidence, which likely will result in less of a drive to restudy it.

The older adult population is confronted with even more of a problem when it comes to learning medical materials. Older adults show general declines across a number of measures of episodic memory (Nilsson et al., 1997). Importantly, they show specific deficits in their ability to bind associated information in memory (Naveh-Benjamin, 2015). This deficit is not an attentional resource problem, nor a general memory problem: older adults remember the component parts of pairs well, but have specific trouble with making the associations between them. Given that a lot of medication adherence is dependent on the ability to remember information associated with specific medications, these are troubling findings.

It is possible that older adults have a framework (or perceived framework) in place for learning new medications. For example, Friedman, McGillivray, Murayama, and Castel (2015) showed that older adults may have some schematic support for learning medications and side-effects. It is postulated that their prior experience with similar materials lends

support during learning. In multiple experiments, older adults performed just as well as younger adults when learning lists of side-effects, when the value of learning those effects was made salient (Friedman et al., 2015). Note that this was learning only for side-effects, not a task where the side-effects were paired with drug names. However, considering the evidence for schematic support for side-effect learning, we may see a similar reprieve from associative deficits with medical materials. Given that most older adults are taking a number of prescription pills (see Qato et al., 2008), research of this nature would be particularly helpful and informative.

Lastly, future studies might focus on the effects of other sources of fluency that could affect judgments of learning for medical materials. In these experiments participants were attempting to study large numbers of pairs at a computer screen. Medical information is commonly presented in pamphlets, on pill bottles, or package inserts. Generally the more important information is highlighted, in uppercase, in boldface, or otherwise pointed out by the pharmacist to patients. These factors are all possible contributions of perceptual fluency which may artificially increase confidence in one's ability to remember the information. Considering this information in context, and not at a computer screen is a critical extension that should be considered. Some pilot studies have been started in lab considering the contextual effects of pill bottles (Nguyentan, Blake, & Castel, 2017) and whether font-size is considered in the construction of JOLs for these (already disfluent) materials, but continuing work is imperative.

CHAPTER 4

Judgments Made About Long-Term Incidental Encoding

In the previous sections I have largely considered metamemory judgments made at the time of encoding. However, there are many instances where we are forced to make judgments about our memory absent any immediately preceding experience of encoding. One classic example of judgments made removed from encoding is the delayed JOL. When participants make delayed JOLs, they usually making a judgment about how well an item will be remembered on a later test, and that judgment is made just after a test of the item. As such, these judgments are remarkably accurate (Nelson & Dunlosky, 1991), a finding which was replicated in Chapter 3 (i.e. global EOLs) even at a much greater delay from test than in Nelson and Dunlosky (1991). Even in these instances though, the experience of encoding is still relatively accessible.

This section examines memory and metamemory for items where encoding is not easily accessed, if accessible at all. These situations are more common and vivid with visual memory than for verbal memory, especially when considering memory for extremely common information that we see every day.

4.1 Study 4[1]

Overall, visual memory tends to be very accurate in humans, such that these memories are stored as distinct and protected from interference; even when hundreds of photos intervene between the first and second appearance of a photo, recognition accuracy is high (Nickerson, 1965). Other research has also shown an immense capacity for visual detail in long-term memory, with high accuracy for over 2,000 images (Brady, Konkle, Alvarez, & Oliva, 2008). However, in a classic study, people were shown to have difficulty recognizing the correct locations of features on a penny (Nickerson & Adams, 1979). Similarly, people often fail to recall the location of previously seen fire extinguishers, despite the fact that fire extinguishers are in high-visibility locations (Castel, Vendetti, & Holyoak, 2012). Explicit memory is poor for items that people interact with daily, such as the keypads of calculators and telephones (Rinck, 1999), computer keyboards (Snyder, Ashitaka, Shimada, Ulrich, & Logan, 2014), the layout of frequently-used elevator buttons (Vendetti, Castel, & Holyoak, 2013), and aspects of road signs (Martin & Jones, 1998), among other items (Castel, Nazarian, & Blake, 2015).

Poor memory for very common objects may be due to a form of attentional saturation, which could then later result in "inattentional amnesia" (Wolfe, 1999). For these types of objects, it becomes unimportant to remember the explicit details due to the frequent presence of those objects in the environment. An extreme case of this is the letter "g." The lowercase "g" is commonly written with a "looptail" (like in the current font) or an open tail (like in print handwriting). Despite massive visual experience, participants show poor awareness of these two forms and have difficulty drawing them (Wong, Wadee, Ellenblum, & McCloskey, 2018).

It can be argued that such inattention is an efficient mental adaptation, and that changing the context of encoding that information may lead people to remember it better. That is, it may be that under intentional learning conditions (e.g. Marmie & Healy, 2004), people

[1] Study 4 is now submitted: Blake, A. B., & Castel, A. D. (under revision). Memory and availability-biased metacognitive illusions for flags of varying familiarity. *Memory & Cognition*

are better able to memorize information, even that associated with objects previously seen many times. However, in naturalistic settings, there is likely no intent to encode the details of various logos and symbols, which leads to an interesting dissociation: increased exposure increases familiarity and confidence but does not reliably affect memory. Despite frequent exposure to simple and often visually pleasing symbols, what we think is memorable may not reflect processes in memory and attention that underlie what is actually memorable (see Castel et al., 2015).

The familiarity of highly-available items may lead people to think they have good memory for the items. In many cases availability is a good diagnostic cue for memory, e.g., multiple presentations of an item will lead to better memory at an immediate test than a single presentation would (Ebbinghaus, 1913). However, in the case of very frequently seen items, familiarity may impair attention to their details: the items saturate the environment so thoroughly that the benefit of having a strong memory for them is minimal—if needed, a representation can be found very quickly. In a study regarding memory and confidence in the Apple logo, participants gave judgments of their confidence in their memory for the Apple logo before and after drawing and choosing the logo from a set of alternatives (Blake, Nazarian, & Castel, 2015). Unlike the prior work with the penny (Nickerson & Adams, 1979), Blake et al. (2015) examined a logo that is prominently advertised, that people attend to frequently, and that was designed to be recognizable. Only one participant was able to draw it with all of the correct features, and roughly half of the participants in the study were unable to pick the correct logo from a set of alternative versions. Participants were overconfident in their memory for the logo when judgments were made prior to both the drawing and recognition tasks (see Iancu & Iancu, 2017, for a replication). The discrepancies between confidence in metamemory and memory indicate that participants were relying on inappropriate strategies or information when assessing their memory. These findings resonate well with work suggesting that judgments of performance are inferred through subjective experiences rather than objective performance (Werth & Strack, 2014).

What are the subjective experiences that may be inflating confidence for highly available

items? A common influence on metacognitive judgments is the ease of processing information at encoding (Begg et al., 1989; Hertzog et al., 2003; Koriat & Ma'ayan, 2005). Generally speaking, easily "learned" information is judged as easy to remember. In particular, when participants are able to generate an image of a to-be-studied item rapidly, they will give higher likelihood judgments of later recall even though this fluency is not well-correlated with recall (Hertzog et al., 2003). Logos, flags, and other brands are designed to be easy to encode and recognize. Advertisers often strive to create minimalistic, simple logos, which are processed more fluently than overly-detailed or complex logos (Janiszewski & Meyvis, 2001).

The design and inherent processing fluency of these highly fluent and available stimuli likely leads to overconfidence in memory for them. The present research utilizes the processing fluency inherent to national flags. Many of the design considerations for national flags regard the speed of recognizing the flag. For example, the number of points on the iconic maple leaf of the Canadian flag was decided following wind tunnel tests of identification and blurriness (Matheson, 1980). Similarly, the design for the flag of the United States of America (US) is based on a naval design where the white stripes were placed on a red background, presumably because a red border is easier to distinguish against a bright sky (Williams Jr., 2012). Additionally, national flags tend to have specific verbalizable rules that are often taught to schoolchildren. For example, the US flag has 13 alternating red and white stripes and a blue field of 50 white stars in the upper left corner. An Italian colleague informed us anecdotally that children in Italy are taught that the green part of the flag should touch the pole. Having both a verbal code and a mental image for a particular object might enhance memory for that object because details are encoded in different ways, resulting a stronger memory trace (Paivio, 1986). However, the presence of both a verbal and visual code may impair metacognition because the verbal code specifically highlights aspects of the object to attend to (e.g., the number of stars that should be present). If those aspects are not critical for identifying the correct flag, they may hinder mnemonic performance and be overconfident prior to the choice and after making the choice.

Another source of metacognitive bias related to overconfidence for common items is the ease of processing at retrieval: when it is easy to retrieve information from memory, that information is judged as better learned than information that takes longer to bring to mind (Kelley & Lindsay, 1993; Koriat & Ma'ayan, 2005; Koriat et al., 2006; Schwarz et al., 1991). For frequently-seen items like logos, it is presumably a relatively fluent experience to generate a vague mental image of the logo. Further, this ease-of-generation may lead to a high confidence in the memory that prevents critical inspection of the mental image: if confidence is at ceiling, there can be no perceived ambiguity in the memory. Indeed, in the study with the Apple logo, participants showed apparent ceiling effects in confidence judgments elicited prior to each memory task (Blake et al., 2015).

Finally, the well-documented availability heuristic suggests that people often use cues like relative frequency and recency to guide their judgments (Tversky & Kahneman, 1973). In this research, Tversky and Kahneman (1973) showed that judgments of ecological frequency—how often something occurs in the natural world—for items in a category correlate with the number of examples in a category that participants can bring to mind. For instance, a person's estimate of the class' grade point average might be higher if that person has more friends with higher grades than not. This bias has been explained as an effect of the ease of retrieval for instances rather than the number of instances recalled (Schwarz et al., 1991). Primarily, these availability effects are related to judgments of frequency and probability, but availability may have downstream consequences for metacognitive judgments. We hypothesize that when participants make judgments about highly available items, the judgments are partially influenced by the ease of recalling encounters with the items rather than critically assessing their memories for detail. That is, it may be that when an item is ubiquitous in nature participants substitute a judgment of the frequency recent interactions with the item instead of their memory for it.

The present study approaches this topic from a novel perspective by examining the effects of environmental saturation on metacognition and memory for an object of high personal relevance: the flag of one's own nation, in this case the United States of America (US).

Importantly, relevance and availability of the US flag are highly variable over the course of a year. At the 4th of July and surrounding weeks, the flag is featured prominently in many public venues, leading to a very saturated state of availability in memory. Our first aim in this study was to assess how metacognitive judgments parallel the relative availability of the flag. In the case of lexical materials, participants have better recall and faster response times for words congruent to nearby holidays (e.g. "haunted" in October) than for words not associated with nearby holidays (Coane & Balota, 2009). Though we do not anticipate better memory for the US flag as it is a member of the highly-available items discussed here, it is expected that a more flag-saturated environment will lead participants to be more confident in their ability to recall the flag.

To rectify errors in metacognition and memory that may arise from these biases, Experiment 2 and Experiment 3 direct participants to attend to more relevant, diagnostic cues for memory. Using availability as a cue is useful in many contexts. However, availability is not always a good diagnostic cue for memory, as has been shown in numerous cases in which frequent interaction with an item has not resulted in better recall (Castel et al., 2015). Retrieval fluency is a more diagnostic cue when used in relevant contexts, namely when an attempt is made to recall an item in a manner similar to the test context. For example, when participants were given multiple test events during a study phase, they gave more accurate judgments of how they would perform at a final test (slightly under-confident) than did participants that had no test events at study (grossly overconfident; Roediger & Karpicke, 2006). The leading explanation for this improved accuracy following delayed recall, or test, is that it encourages individuals to attempt to recall a prior memory rather than make a judgment predicated on the current task-difficulty (Nelson & Dunlosky, 1992). If participants are making their judgments of confidence based on the availability of encounters with the US flag instead of retrieving a prior memory, prompting them to think more specifically about their memories for the items and express the details of them, either verbally or visually, is expected to improve metacognitive judgments akin to this testing effect.

Finally, in the case of highly available images, participants maintain overconfidence fol-

lowing free recall and recognition tests, although their overconfidence is attenuated by experiencing the recall episodes (Blake et al., 2015; Iancu & Iancu, 2017). The final aim of the current study was to correct post-recall overconfidence. Metacognitive biases are relatively resilient, and people do not always fully update their knowledge with experience (Mueller et al., 2015). However, overconfidence may be remedied by improving memory through the use of a learning paradigm that specifically highlights errors in memory and corrects those errors. Related research has shown that people are sometimes less overconfident when asked to retrieve specific details of a process before giving their confidence judgment (Keil, 2003; Rozenblit & Keil, 2002).

Generating errors during learning has positive effects on memory when coupled with immediate corrective feedback (Kang et al., 2011; Kornell et al., 2011; Richland, Kornell, and Kao, 2009; Yang, Potts, and Shanks, 2017; but see also Cyr et al., 2015). Notably, research on errorful learning focuses on the learning of new information, specifically for novel word associations though it has been extended to more semantically rich information like trivia questions (Kornell, 2014). However, testing with feedback has also been shown to improve memory for prior-learned information (Fenesi, Sana, and Kim, 2014; see Rowland, 2014, for a meta-analysis). This benefit was limited to practice questions that tested basic retention of facts, which is relevant to memory for visual materials like flags and logos where nearly all features are low-level. In light of this research, it is expected that introducing a recall event prior to study will lead to error-generation that will complement study, improve later recognition of the studied item, and consequently reduce overconfidence (by improving memory to match confidence).

4.1.1 Experiment 1

In the current experiment, we examined the effects of relative availability of flags on memory and metamemory for those flags. We see flags frequently and they likely have a high personal relevance for some people. Comparing memory for flags of different countries to one's own country provides a foundation for understanding how personal relevance and availability

heuristics may affect mnemonic phenomena. Further, national flags have different levels of frequency and extrinsic relevance throughout the year, a point which is integral to this experiment.

If participants make memory judgments based on availability, it is expected that *a priori* confidence in their ability to recognize the flag correctly will be mis-calibrated for very available objects like their country's national flag. Further, it follows that at time-points during which the US flag is more available, overconfidence will increase compared to more neutral time-points. In particular, the US flag is much more prominent and visually available during the weeks surrounding Independence Day in the US (July 4th), which may lead participants to think they can recall its details better—or at least that they "should" be able to due to the flag's increased cultural significance (and, perhaps, participants' increased patriotism) during that holiday. Additionally, it is expected that participants will be less overconfident in their memory for the Canadian (CA) and Mexican (MX) flags, which are not prominently featured on a daily basis in most parts of the US. However, the CA flag is relatively simplistic in its design compared to the more complex MX flag, and presumably this results in higher encoding and retrieval fluencies. The simplicity may foster a sense of confidence in the CA flag compared to the MX flag.

4.1.1.1 Method

Participants and design. Data was collected from 86 participants recruited through Amazon Mechanical Turk, who were paid $6/hr. Participation was limited to people in the United States of America (US) and to workers who had not already participated in any pilot studies involving these or similar materials. Data collection for this experiment occurred at two time-points in 2016: July 4th ($n = 43$, 23 females, $M_{age} = 34.14$, $SD_{age} = 11.68$), and August 6th ($n = 43$, 24 females, $M_{age} = 31.58$, $SD_{age} = 9.10$). This variable was manipulated in a naturalistic, quasi-experimental manner where participants were not randomly assigned to collection date. Participants participated only at one of the time-points, during which their recognition and metamemory performance (pre-recognition confidence and post-recognition

Table 4.1

Altered features for each of the flag stimuli.

Flag	Feature	Correct	Incorrect	% Correct
CA	1	11–pointed maple leaf	15–pointed maple leaf	79.9%
	2	Correctly–sized leaf	Reduced–size leaf	84.3%
	3	Straight leaf stem	Naturally bent leaf stem	60.5%
US	1	Five–pointed star	Six–pointed star	80.0%
	2	50 stars	41 stars	78.1%
	3	Field spans seven lines	Field spans six lines	61.9%
MX	1	Mexican flag emblem	Green US Presidential Seal	61.1%
	2	Emblem facing left	Emblem facing right	75.6%
	3	Green – White – Red	Red – White – Green	50.5%

Note: See Appendix B for the stimuli constructed from these descriptions. The rightmost column "% Correct" is the percentage of people across the relevant experiments that chose a flag with the correct version of that feature.

confidence) for three flags was recorded.

Materials. A set of eight flag stimuli was constructed for each of the flags of the US, MX, and CA. Each set included the correct flag along with seven alternatives created by manipulating key features of the flag. Only three prominent features of each flag were systematically varied for each of the alternatives. These alternatives were informed by pilot studies to yield highly-competitive lures. The correct features and the corresponding alternate features for each flag are detailed in Table 4.1. For each flag the emblem was modified, the layout was altered, and the proportion of the flag taken by the main emblem was changed to create the alternate features. A flag was created for each combination of correct and incorrect features, yielding eight flags per country (see Appendix B for the exact materials used).

Procedure. Participants started the experiment by entering their Amazon Mechanical Turk worker IDs. On the following screen they were instructed that they would be viewing pictures of common objects and would answer questions about them, but importantly they should not look around or navigate away from the page when answering. It was emphasized that their data would be unusable if they were to "cheat," so to speak.

The order and sequence of the flags was counterbalanced, and participants were randomly

assigned to one of the counterbalanced conditions. For each flag, participants were asked to imagine that they would be shown a set of alternatives of the upcoming flag (the country associated with the flag was indicated in the prompt) and were asked to rate their confidence that they could pick out the correct version on a scale from 0 (not confident at all) to 100 (completely confident). Then they were told that they would be shown a set of eight flags as a grid of choices on a neutral gray background (see Appendix B for an example of each grid). To select a choice, a participant would click on the flag that they felt depicted the correct flag for that particular country. A yellow rectangle would show which flag was selected until the participant chose to submit the response. For each participant, the position of each flag in the grid was randomized. Once the response was submitted, participants were again asked their confidence on a 0 to 100 scale, this time regarding whether they chose the correct flag or not. This sequence would repeat until the participant had responded to each of the three flags.

After all of the prompts and flag sets, participants were asked to answer how many stars are on the US flag, how many stripes are on the US flag, their awareness of relevant holidays (Canada Day on July 1st, and Independence Day on July 4th), and a number of demographic questions.

4.1.1.2 Results and Discussion

Figure 4.1 shows the average pre-recognition confidence, recognition accuracy, and post-recognition confidence as a function of flag shown and environmental saturation. The pattern of results shows relatively similar recall for each of the flags that does not differ across time-points. Additionally, the rate of decrease from pre- to post-recognition judgment appears to be stable. However, there is a clear spike in confidence judgments for the US flag at the 4th of July (high environmental saturation), as expected. To test these apparent effects, an analysis of variance (ANOVA) was run for the recognition and confidence judgments.

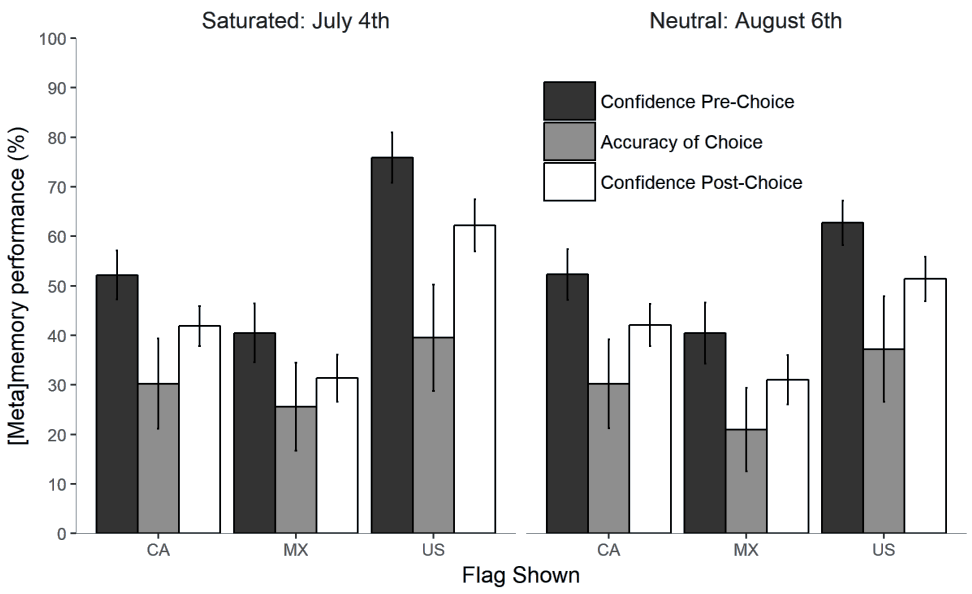


Figure 4.1. Metamemory (confidence) and memory (recognition) performance for each of the CA (Canadian), US (United States), and MX (Mexican) flags at the saturated and neutral time-points. The error bars attached to each of the columns indicate 95% confidence intervals.

Recognition. The percentage of correct choices was analyzed in a 3 (flag: CA, US, MX) x 2 (environmental saturation: low at August 6th, high at July 4th) mixed-factor ANOVA. These data indicated no main effect of environmental saturation on recognition, $F(1, 84) = 0.13, \eta_p^2 = .002, p = .72$, consistent with prior work demonstrating that the high availability of items does not enhance memory for those items (see Castel et al., 2015). Further, there was no interaction between the type of flag shown and environmental saturation, $F(2, 168) = 0.06, \eta_p^2 = .001, p = .94$. Although it was expected that participants might perform more accurately for the US (very familiar, $M = 38.37$, $SD = 48.91$) and CA (less familiar yet very simplistic; $M = 30.23$, $SD = 46.20$) flags over the MX flag (less familiar and complex; $M = 23.26$, $SD = 42.49$), there were no significant differences, $F(2, 168) = 2.67, \eta_p^2 = .031, p = .07$. However, the lack of differences here is not particularly surprising: the flag stimuli alternatives were crafted by the researchers to have only three possible alterations, but due to the nature of flags, these alterations cannot be considered equivalent across flags. Thus,

it is possible that there are differences in recognition based on both memory for the flag and differences in the difficulty of the recognition task across flags, and these differences are canceling one another.

Confidence. Participants' metamemory for the flags was compared using a 3 (flag: CA, US, MX) x 2 (environmental saturation: August 6th, July 4th) x 2 (pre- vs. post-recognition) mixed-factor ANOVA. First, considering the three-way interaction, there were no significant differences in the rate of change in confidence from pre- to post-recognition judgments, $F(2, 168) = 0.13, \eta_p^2 = .002, p$ .88. Similarly, there were no significant changes in pre- and post-recognition confidence as a function of flag [$F(2, 168) = 0.66, \eta_p^2 = .008, p$ .52] or date [$F(1, 84) = 0.04, \eta_p^2 = .001, p$ .84]. However, there was a marked decrease in confidence judgments from those given prior to the recognition task ($M = 54.02$, $SD = 31.26$) to those given after the recognition task ($M = 43.32$, $SD = 30.99$), $F(1, 84) = 35.63, \eta_p^2 = .298, p <$ .001.

We next considered the more critical effects regarding the flags shown and the environmental saturation. There was no significant main effect of environmental saturation [$F(1, 84) = 0.70, \eta_p^2 = .008, p = .41$] on confidence judgments, yet there was a significant difference among the flags, [$F(2, 168) = 47.34, \eta_p^2 = .360, p < .001$]. Independent samples t tests comparing the average confidence in memory for each flag indicated that participants gave higher confidence judgments for the US flag ($M = 63.06$, $SD = 25.53$) than the CA flag [$M = 47.11$, $SD = 28.14$; $t(168) = 3.89, d = 0.59, p < .001$], and higher confidence judgments for the CA flag than the MX flag [$M = 35.83$, $SD = 26.98$; $t(168) = 2.68, d = 0.41, p = .008$]. Importantly, confidence judgments for the flags interacted with environmental saturation, $F(2, 168) = 3.04, \eta_p^2 = .035, p = .05$. Independent samples t tests showed that this interaction was driven by the US flag: average confidence in the US flag at the 4th of July ($M =$ 69.07, $SD = 27.20$) was significantly higher than at August 6th ($M = 57.06$, $SD = 22.49$), $t(84) = 2.23, d = 0.48, p = .028$. There were no significant differences between the time-points for either the CA [$t(84) = -0.03, d = -.006, p$ .98] or MX flags [$t(84) = 0.03, d = 0.006, p$ = .98]. Thus, environmental saturation was associated with elevated confidence but no

difference in accuracy.

Finally, looking at the broader comparisons of recognition, participants were highly overconfident in their ability to choose the correct flag both before (about 54%) and after (about 43%) the recognition task (hit rate about 30%). There is a decline between the two confidence judgments, yet participants remain overconfident. One explanation for this overconfidence might be that participants are retrieving improperly stored representations of the flag, and thus the confidence is high because they feel they chose a matching representation. Alternatively, these findings may suggest that participants are not consulting their memory at all for the flags before making their judgments; instead they use the most salient heuristics when making their confidence judgments. This is particularly evident in the case of the US flag where changes in availability of the flag predict changes in participants' confidence regarding it. That is, on the 4th of July when the US flag is relatively saturated in the environment, participants give higher ratings of their confidence for that flag compared to the CA and MX flags, which are not found in great abundance during either time point.

4.1.2 Experiment 2

Given the findings from Experiment 1 indicating relative overconfidence and poor memory for the CA, US, and MX flags, Experiment 2 sought to debias participants' metacognition by prompting them to consider their memory before making confidence judgments. In Experiment 1, it is possible that participants were making their pre-confidence judgments regarding the flags based solely on non-diagnostic factors such as availability (Tversky & Kahneman, 1974) or rote knowledge (e.g., memorized rules) about the flags. Considering that the recognition task is visual in nature, it follows that consulting a mental image of each flag would result in a more accurate metacognitive prediction. By asking participants to describe the flags before making judgments about them, we expected participants to bring to mind the mental images of the flags, creating more diagnostic cues to factor into their judgments. Similarly, when participants describe how to complete a task it lowers their overconfidence in understanding the process (Rozenblit & Keil, 2002) These descriptions are expected to

improve metacognitive performance by decreasing the overconfidence seen in Experiment 1.

4.1.2.1 Method

Participants. Data was collected from 214 participants (114 females, $M_{age} = 36.83$, $SD_{age} = 12.48$), recruited through Amazon Mechanical Turk, who were paid \$6/hr. Participation was limited to people in the United States of America (US) and to workers who had not already participated in Experiment 1 or any pilot studies involving these or similar materials.

Design. Participants were primed with either neutral (workspace-related) or targeted (flag-related) prompts for descriptions. Their recognition memory and metacognitive judgments prior to and after recognition were recorded for the CA, US, and MX flags.

Materials. The flag materials for the US, CA, and MX alternatives in this task were the same as those used in Experiment 1 and can be found in the Appendix B.

Two types of prompts were created for the experiment. Targeted prompts instructed participants to describe each of the US, CA, or MX flag in their own words. For neutral prompts, participants were asked to describe their computer keyboard, the wall behind them, or the chair that they were sitting in. The orders of the flag-targeted prompts and neutral prompts were each counterbalanced.

Procedure. Participants were randomly assigned in equal numbers to either answer targeted or neutral prompts at their own pace. The procedure in this experiment was nearly identical to Experiment 1, with two differences: the data was collected only at one time-point, and prior to each flag sequence (pre-confidence judgment, recognition task, post-confidence judgment) participants answered the prompt assigned to the upcoming flag.

4.1.2.2 Results and Discussion

In Figure 4.2, summaries of the metacognitive and recognition performances are displayed as a function of the flag shown and the priming prompt type. Compared to Experiment 1,

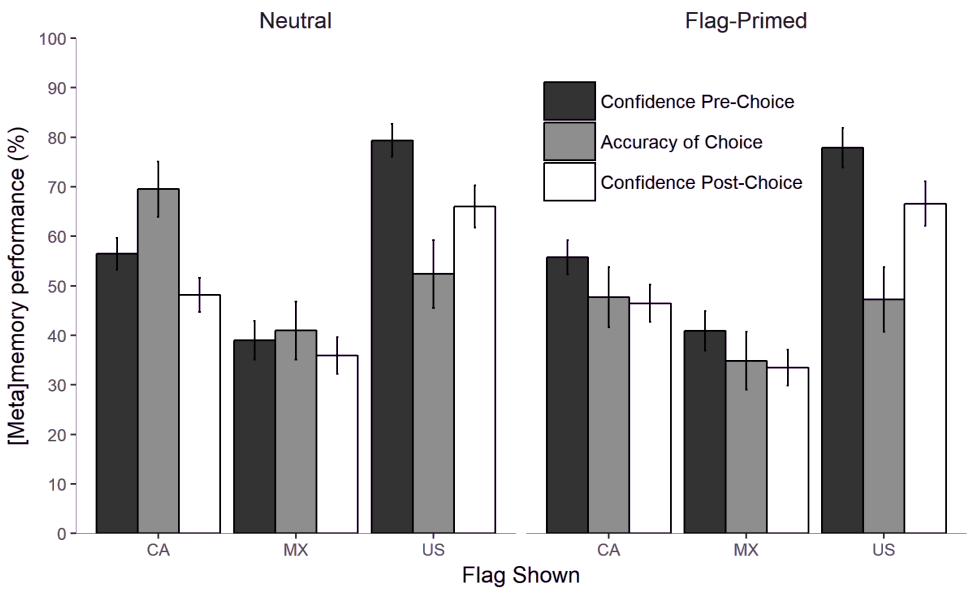


Figure 4.2. Metamemory (confidence) and memory (recognition) performance for each of the CA (Canadian), US (United States), and MX (Mexican) flags in the neutral and targeted priming conditions. The error bars attached to each of the columns indicate 95% confidence intervals.

the confidence scores look very similar in that the US averages are higher than CA, and CA higher than MX. Further, there appears to be a general overconfidence for the US flag that is not seen in the CA or MX flags.

Recognition. A mixed-subjects ANOVA tested the effects of the flags shown (CA, US, MX) and priming prompt (targeted, neutral) on recognition. The ANOVA revealed effects of both the prompts and the flags, but these effects were not qualified by an interaction, $F(1, 422) = 2.40, \eta_p^2 = .011, p = .09$.

Participants in the neutral priming condition ($M = 54.29$, $SD = 49.90$) performed better on the recognition task than participants in the targeted condition ($M = 43.25$, $SD = 49.62$), $F(1, 211) = 6.49, \eta_p^2 = .03, d = 0.18, p = .01$. This finding was somewhat unexpected in that thinking of an object does not usually impair memory for the object. However, research in retrieval-induced forgetting has shown that the act of retrieving some subset of information can reduce memory for other related information (Anderson, Bjork, & Bjork, 1994). It

is possible that the act of verbally retrieving some of the details of the flag simultaneously selected against other non-salient details that would become important at test, thus reducing performance. Similarly, research with the verbal overshadowing effect Schooler and Engstler-Schooler (1990) has shown that identification of previously seen faces is impaired when immediately preceded by a verbal prediction task (Meissner & Brigham, 2001). In the current experiment, the verbal descriptions may have oriented participants to non-discriminative characteristics of the flag foils or caused poor reconstruction of the flag memory.

There was also a significant main effect of flag shown, $F(1, 422) = 10.54, \eta_p^2 = .048, p < .001$. The CA flag ($M = 58.41$, $SD = 49.40$) was not recognized more often than the US flag ($M = 49.77$, $SD = 50.12$) [$t(213) = 1.93, d = 0.13, p_{bonf} = .16$], but was recognized more often than the MX flag ($M = 37.85$, $SD = 48.62$) [$t(213) = 4.57, d = 0.31, p_{bonf} < .001$]. The US flag was also correctly recognized more often than the MX flag, $t(213) = 2.64, d = 0.18, p_{bonf} = .03$. This pattern of results was what was expected in Experiment 1, although the main effect of flag shown was only trending ($p = .07$) in that instance. The same caveats regarding the construction of the materials and possible differences in relative difficulty of recognition still apply, as this experiment uses the same materials. However, it may be that the previous experiment simply had less power to find the recognition effect with these materials, and the larger sample size for this experiment addressed this issue.

Confidence. Participants' confidence scores were compared across flags (within: CA, US, MX), priming condition (within: targeted, neutral) and judgment time-points (between: pre-choice, post-choice) in a mixed-subjects ANOVA. The three-way interaction was not significant [$F(2, 424) = 0.99, \eta_p^2 = .005, p = .37$]. Further, there was no significant interaction between priming prompt and judgment time-point [$F(1, 212) = 0.33, \eta_p^2 = .002, p = .57$], no interaction between priming prompt and flag [$F(4, 424) = 0.02, \eta_p^2 = .001, p = .98$], nor a main effect of priming prompt [$F(1, 212) = 0.09, \eta_p^2 = .001, p = .77$]. The lack of priming effects on confidence is intriguing and defies the *a priori* expectations for this experiment. It was hypothesized that the largely inaccurate metacognitive judgments for Experiment 1 were caused by the use of heuristics instead of memory appraisal through retrieval. However,

even when participants actively recalled the flag, they were unable to make an appropriate judgment of their memory for it.

Despite the lack of priming effects on confidence, there were significant effects on confidence depending on the flag shown, $F(2, 424) = 98.05, \eta_p^2 = .316, p < .001$. Participants were more confident in their memory for the US flag ($M = 72.45$, $SD = 25.45$) compared to the MX flag ($M = 37.32$, $SD = 26.60$) [$t(212) = 13.93, d = 0.95, p_{bonf} < .001$], and compared to the CA flag ($M = 51.0$, $SD = 26.90$) [$t(212) = 8.22, d = 0.56, p_{bonf} < .001$]. Participants were also more confident in their memory for the CA flag compared to the MX flag, $t(213) = 5.70, d = 0.39, p_{bonf} < .001$. This pattern is consistent with Experiment 1 where overall confidence for the US flag was greater than for the CA flag was greater than for the MX flag.

Lastly, participants were more confident in their judgments prior to each choice ($M = 58.23$, $SD = 31.60$) than those after ($M = 49.42$, $SD = 32.84$) the recognition choice was made, $F(1, 212) = 87.28, \eta_p^2 = .292, d = .639, p < .001$. This change in confidence was greater for some of the flags than the others, $F(2, 424) = 4.99, \eta_p^2 = .023, p = .01$. The change in confidence for the US flag from pre- ($M = 78.62$, $SD = 25.10$) to post-recognition choice ($M = 66.29$, $SD = 31.29$) was the largest [$t(213) = 7.20, d = 0.49, p < .001$], followed by the change pre- ($M = 56.12$, $SD = 28.55$) to post-recognition ($M = 47.29$, $SD = 29.88$) for the CA flag [$t(213) = 5.65, d = 0.39, p < .001$], which was trailed by the change pre- ($M = 39.96$, $SD = 28.32$) to post-recognition ($M = 34.69$, $SD = 29.42$) for the MX flag [$t(213) = 3.43, d = 0.24, p < .001$]. In each case, the post-recognition confidence judgment was much better calibrated. The reduced variance from pre- to post-recognition in the MX flag data belies a "confidence in one's own confidence"; that is, participants understand that they have a poor recollection of the MX flag and are able to make a relatively accurate judgment for it that does not change over time. On the other hand, the very available US flag and relatively simplistic CA flag are associated with increased levels of overconfidence prior to the recognition task. This discrepancy in overconfidence suggests that participants are attending to different cues when making their judgments for non-available and unfamiliar

items compared to familiar items.

4.1.3 Experiment 3

In Experiment 2, participants who typed out the features of the flag before attempting to identify it (priming condition) performed more poorly than participants who completed a control task. Possibilities for this result include the effects of verbal overshadowing (Meissner & Brigham, 2001), retrieval-induced forgetting (Kornell & Bjork, 2009), or a combined false memory explanation whereby participants managed to introduce false details of the flag into their mental image of the flag during the description task, possibly from a faulty or fuzzy memory supplemented by guesswork. Further, the act of describing the flag using verbal codes involves different mental faculties than simply picturing the flag in one's mind (e.g. Paivio, Rogers, & Smythe, 1968). It may be that the verbal code for such a well-known flag invokes different details than a more visual code. Given that the flag is relatively simple in its visual design—as compared to, say, a human face—it may be more helpful to attempt to draw the flag as a method of retrieving the memory. Drawing also has been shown to have strong memorial effects, not unlike the related production effect (MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010), possibly because drawing involves the integration of semantic, visual, and motor memories (Jeffrey D. Wammes, Meade, & Fernandes, 2016) and can aid in producing recollection based memories with more intact source memory (Jeffrey D Wammes, Meade, & Fernandes, 2018).

It is expected that when participants are forced to attempt to draw out the features of the flag, they will be more likely to focus on features they had not considered critical before. For example, participants from the US are very likely know that the US flag has 50 stars and 13 stripes, but may not have considered the arrangement of the stars, the shape of the blue field, and the number of stripes below the blue field. When attempting to draw the flag, these details must be considered in order to create a picture. Generating errors surrounding such details and then providing corrective feedback should benefit memory for those details (Fenesi et al., 2014; Kang et al., 2011). This research on feedback suggests

that if participants were to see the correct flag after they had trouble attempting to draw it, they will develop stronger memories for the flag because they are cued to attend to crucial, overlooked features. However, if participants were to only study the flag, it is unlikely that they would consider such features, as they have not attended to them in countless past viewings of the flag.

Extending these predictions to metamemory judgments, past research shows that participants are frequently overconfident prior to any attempts to retrieve common objects from memory (Blake et al., 2015; Iancu & Iancu, 2017). It is unlikely that simply studying the flag will alter that confidence, as it is already a well-known and commonly seen object. However, attempting to recall the flag and experiencing the disfluency of retrieval for unknown details may force participants to temper their metacognitive judgments (Miller & Geraci, 2014; Rozenblit & Keil, 2002). The more salient cue of retrieval strength should force participants to realize which features they do not have committed to memory and lower their confidence.

4.1.3.1 Method

Participants. Data was collected from 52 participants (35 females, M_{age} = 19.86, SD_{age} = 1.43) through the Psychology Department subject pool at University of California, Los Angeles. Participants received course credits for participating in the study.

Materials and procedure. Participants were randomly assigned in equal numbers to either a study-only condition or a draw-then-study condition. All participants were seated at a desk with a computer which displayed the questions and images in the experiment.

To begin the experiment, all participants were asked to rate how confident they were that they could correctly choose the US flag from a group of alternatives on a scale from 0 (not at all) to 100 (extremely). Then, in the draw-then-study condition, participants were given a sheet of paper and colored pencils and told to draw the US flag on the provided sheet of paper. After 40 s the paper was removed and they were shown a correct image of the US flag on the computer screen to study for 40 s. Participants in the study-only condition were

shown the correct image for 80 s and not asked to draw anything. All participants then made another rating of their confidence that they could choose the correct US flag from a set of similar alternatives.

Prior to the recognition phase of the experiment, all participants completed other laboratory experiments for approximately 20 min. These experiments were primarily word-learning experiments and did not have relevant visual stimuli.

After the intervening experiments were completed, participants again rated their confidence in their ability to choose the correct US flag. Then, they were shown the US flag alternatives (see Appendix B) in a grid and made their choice of the correct flag as in Experiment 1 and Experiment 2. Finally, they were asked to rate their confidence that they chose the correct US flag.

4.1.3.2 Results and Discussion

Recognition. This experiment utilized a drawing task to promote error-generation and facilitate learning. An independent samples t test showed that a larger percentage of participants in the draw-then-study condition ($M = 76.92$, $SD = 42.97$) chose the correct US flag than in the study-only condition ($M = 38.46$, $SD = 49.61$), $t(50) = 2.99, d = 0.83, p = .004$. These data indicate that drawing the flag benefited participants' subsequent learning of the flag at study. This is particularly interesting in that participants in the study-only condition had twice as long to study the flag (80 s) compared to the draw-then-study participants (40 s draw, 40 s study). This is consistent with research on errorful-learning showing that generating an answer is more beneficial than an equivalent amount of study time, even if the answer is incorrect (Kornell & Bjork, 2009). Attempting to retrieve the flag served as a powerful learning event that produced learning beyond intentional study of the flag. We suggest that the benefit in this experiment is derived from the productive failures made during the drawing phase. These failures likely serve to direct participants' attention to study the features of the flag that they were unsure of when attempting to draw it. This explanation

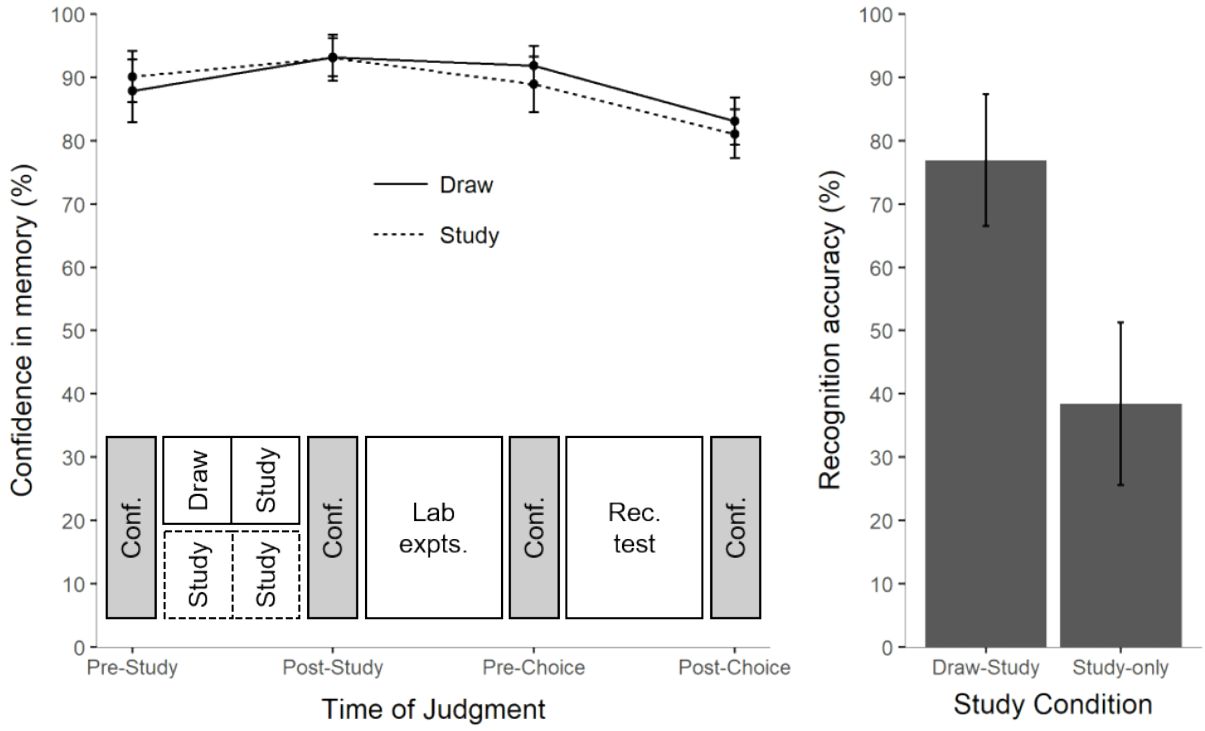


Figure 4.3. Confidence in memory for the flag of the United States at each of the four metacognitive-judgment time-points (left panel) compared to the recognition accuracy as a percentage of correct responses (right panel). A study outline diagram is overlaid in the left panel to clarify when each measure was taken. The error bars attached to each of the columns indicate 95% confidence intervals.

fits with other data showing that immediate feedback improves memory for fact-based information (Fenesi et al., 2014), even when errors are made at test (Kang et al., 2011). Lastly, these data complement recent research on the benefits of drawing information rather than just studying or restudying it (Jeffrey D. Wammes et al., 2016).

Confidence. In the last two experiments, metacognitive judgments have been poorly calibrated in that participants overestimated their performance on the recognition task. Additionally, the recognition test has acted as something of a metacognitive intervention where judgments made post-test have been lower, indicating a lower overconfidence. Figure 4.3 shows the confidence at each judgment time in this experiment (pre-study, post-study, pre-

choice, post-choice) with separate lines indicating the condition (draw, study) that is being summarized. The general pattern suggests that participants were very confident in their ability to recognize the US flag and that this confidence seems to be unaffected by the condition they were in. This lack of difference would be in sharp contrast to Rozenblit and Keil (2002) where participants were less overconfident when they had attempted to recall a process than when they had not. Further, there appears to be drop in the final judgment which would indicate that the test is acting as a debiasing event in this experiment as well.

A 4 (judgment time-point: pre-study, post-study, pre-choice, post-choice) x 2 (condition: study only, draw-study) mixed-subjects ANOVA was used to analyze these apparent effects. The interaction between judgment time-point and condition was indeed non-significant, $F(3, 150) = 0.73, \eta_p^2 = .014, p = .54$. Similarly, participants who were asked to draw the flag during the study phase ($M = 86.58$, $SD = 24.32$) were no more confident that participants who only studied the flag ($M = 78.32$, $SD = 32.08$), $F(1, 50) = 0.04, \eta_p^2 = .001, p = .83$. However, there was a main effect of judgment time-point, $F(3, 150) = 12.17, \eta_p^2 = .196, p < .001$.

Multiple paired-samples t tests were run to elucidate the nature of the time-point main effect. From pre-study ($M = 88.98$, $SD = 15.04$) to post-study ($M = 93.14$, $SD = 12.65$) participants became more confident, $t(51) = -2.85, d = -0.40, p = .006$. This is likely due to the fact that the participants have just seen a perfect rendition of the flag, which is a common object, and thus the judgment is made in a very fluent retrieval context compared to pre-study. In the interval during which participants completed other lab tasks, confidence dropped from post-study to pre-choice ($M = 90.37$, $SD = 14.46$), $t(51) = 2.89, d = 0.40, p = .006$. Presumably, the intervening tasks degraded participants' retrieval fluency of the flag, making it harder to pull the details to mind, and reducing their confidence in turn. The values at pre-study, when the flag has not been seen in the lab yet, and at pre-choice, when the intervening tasks had just completed, were similar; participants went back to baseline after the intervening tasks, $t(51) = -0.76, d = -0.11, p = .45$. Lastly, from pre-choice ($M = 90.37$, $SD = 14.46$) to post-choice ($M = 82.08$, $SD = 15.89$) participants again were

debiased somewhat by the recognition task, which is "harder than [they] thought" as some participants reported, $t(51) = 2.914, d = -0.40, p = .005$.

It was expected that participants would be less confident in their memory for the flag after the study opportunity in the draw-then-study condition. The act of drawing likely highlighted many participants' missing or false memories for details of the flag, as was the case in other related studies (Blake et al., 2015; Iancu & Iancu, 2017). The salience of these errors generally decreases confidence in memory, and some research has shown that participants are unaware of the benefits of error-generation on learning (Yang et al., 2017). However, in this experiment the opposite was true: participants in the drawing condition did not show a decrease in confidence following the study phase, and all participants were more confident at the post-study judgment. A likely explanation is that the feedback portion of the draw-then-study condition was enough for participants to evaluate and rectify problems with their memory for the flag and thus re-inflate their confidence. Indeed, post-study confidence appears to be exhibiting a ceiling effect.

Finally, participants in the draw-then-study condition showed better metacognitive calibration overall. That is, their metacognitive judgments showed less of a deviation from their recognition score than did the study-only group. This is evident from a cursory comparison of recognition and metacognition judgments, but an exploratory analysis of the data was performed to confirm the apparent effect. For each participant, the average of all metacognitive judgments was computed and then subtracted from the recognition score. An independent samples t test showed that participants in the drawing condition (M = -12.07, SD = 41.43) gave less extreme judgments than in the study-only condition (M = -49.83, SD = 50.39), $t(50) = 2.95, d = 0.82, p = .005$. It is unclear whether participants are truly debiased, however, as this discrepancy may be driven entirely by the increased ability to recognize the US flag as opposed to improved ability to judge one's own memory. Nonetheless, participants in the drawing group did show less biased metacognitive judgments.

4.1.4 General Discussion

In this study, memory and metamemory for flags was examined across three experiments. A clear thread in these three studies is that participants remained relatively overconfident, especially prior to test. In each case, the testing event served to debias metacognitive judgments significantly: once faced with the test, participants were made aware that their memory was not as accurate as they thought. The only apparent changes in metacognition prior to the recognition task were seen in Experiment 1 where participants in August showed less overconfidence than in July (presumably as a function of flag availability), and in Experiment 3 where the study event increased overconfidence.

Interpreting this time of year effect in light of past research on metacognitive biases, the results of these experiments support a theory of cue utilization (Koriat, 1997), but also show how availability can bias use of certain familiar cues. The US flag is a highly salient national symbol: students learn about it early in school, it is part of history, and Americans likely feel that we should know it well. Additionally, as discussed in the introduction, national flags are often constructed to be relatively simple and easy to encode which may lead to biased metacognition due to the processing fluency. In these studies, participants showed particular overconfidence in the US flag that correlated with natural changes in the environmental saturation of the flag. The US flag tends to be more available in the days surrounding July 4th than a baseline comparison at August 6th, as the flag appears in Independence Day-themed advertisements, social media posts, and even apparel and lawn decorations. Participants tested at the saturated time-point showed a stronger tendency for overconfidence than did those at the neutral time-point, suggesting that environmental saturation and availability may play a large role in how people make their judgments about these types of frequently seen items.

This differs somewhat from other research on prospective judgments of learning and cue-utilization in metacognition. Particularly, when participants are asked to retrieve information from memory, this is usually a very salient diagnostic cue of that memory which

results in very accurate judgments (Nelson & Dunlosky, 1991). In these studies, participants were also asked to judge their memory for the US flag and others when the flag was not present in memory, and yet were unable to produce accurate confidence judgments in their memory. One explanation for the inaccuracy may be that participants were simply unable to comprehend the difficulty of the recognition test despite it being described to them. We discount this notion with the counterargument that participants in both Experiment 1 and Experiment 2 were given a chance to experience the test three separate times, and this did not appear to reduce overconfidence. Additionally, the increase in overconfidence across time-point in Experiment 1 suggests that this overconfidence is related, in part, to the relative availability of recent interactions or sightings with the flag. We suggest that participants are not only considering that they should know the flag due to its ease-of-encoding, cultural significance, etc., but also because of the ease of recalling encounters with it.

In Experiment 2, we sought to bridge the discrepancy and resolve the overconfidence issue by forcing participants to describe the items verbally before making judgments or recognition attempts. If participants were making their judgments more on availability rules than on assessments of retrieval fluency, their metacognitive judgments would likely improve by making the diagnostic retrieval cues more salient. This description task had no effect on metacognitive judgments, suggesting that participants already attempt to bring the item to mind when making their judgment. Further, participants exhibited a verbal overshadowing effect (Schooler & Engstler-Schooler, 1990) where their recognition performance was weakened by the description task, and incidentally resulted in more overconfidence in that condition as a result of the attenuation.

Finally, Experiment 3 demonstrated a strong debiasing intervention that appears to operate solely by improving memory. By asking participants to attempt to draw the US flag, rather than report a verbal code for it as in Experiment 2, memory for the flag was improved beyond that of study alone. This improvement in memory reduced the overconfidence effect drastically and complements findings in extant literature showing that drawing (Jeffrey D. Wammes et al., 2016) can enhance memory via generation (Slamecka & Graf, 1978) and

production-like (MacLeod et al., 2010) effects. Though this drawing effect was shown for enhancing recall of verbal materials, the current study effectively extends the effect to restudy of visual materials. Further, we suggest that the act of drawing made errors in participants' memories for the flag salient, allowing the feedback to have a stronger effect like in other errorful-learning research (e.g. Kornell & Bjork, 2009; Richland et al., 2009).

Failure to recall all the details of a flag is somewhat more opaque than errors in recalling a list of words. When participants are asked to recall as many words as possible from a list, it is very clear when the recalled words number fewer than the studied words. In the case of visual materials, unless the image is drawn participants may not be able to assess fully how many and which details are missing. It is possible that participants are partially relying on semantic rather than visual memory for the flags, and that matches with semantic knowledge inflate confidence. For example, the semantic knowledge that the CA flag has a maple leaf in the center increases confidence, but participants do not readily employ mental strategies to reduce confidence and are unable to assess how detailed that maple leaf should be until they put pen to paper. In essence, the act of drawing the mental representation forces the learner to produce their failures, a process that invokes retrieval dynamics and diagnostic monitoring that result in better memory for the items than additional study alone does.

In sum, this collection of studies shows the improper utilization of cues in metacognitive judgments about national flags, which are highly-available, frequently-seen objects that we often feel we should be able to remember. Further, there is a non-trivial relationship between overconfidence and environmental availability of items (Expt. 1) as well as recent mnemonic activity (Expt. 2). It is clear that these types of items are special in their ability to bias metacognitive faculties across a number of domains and everyday settings (Castel et al., 2015), but this study also shows that this bias is not insurmountable. We demonstrate a powerful metacognitive debiasing intervention and learning tool (Expt. 3) that can be useful in a variety of contexts: the draw-then-study method for invoking productive failure. This work extends theories of errorful learning and generation to visual materials, and highlights the role of productive failures in focusing attention toward previously overlooked features. We

suggest further research to address the long-term effects of this method, and the application of this method to rectify other everyday memory failures, some of which can be very important like the location of the nearest fire extinguisher (Castel, Vendetti, & Holyoak, 2012). As metamemory both matures and changes across the lifespan (see Blake & Castel, 2015, for an overview), it is also of interest to examine the effects of the variables considered here in children and older adults. Flags are especially interesting to study across the lifespan as national attitudes and even the representation of the flag shift over time.

CHAPTER 5

General Conclusions

Throughout, I have talked about the broad components driving prospective memory judgments, how memory and metamemory patterns can shift across domains, and some of the factors affecting metacognitive judgments that are far removed from encoding.

As a type of metacognitive monitoring, JOLs almost certainly do not have any direct access to memory. Instead these judgments are often informed through peripheral, erroneous cues and beliefs which distort the actual level of learning that has occurred. Past research has strongly suggested that these cues and their effects on metacognitive monitoring are driven by fluency (Baddeley & Longman, 1978; Begg et al., 1989; Benjamin, 2003; Kelley & Rhodes, 2002; Koriat & Bjork, 2005; Koriat & Ma'ayan, 2005; Rhodes & Castel, 2008a, 2009). Whether it is fluency or beliefs about fluency (Mueller et al., 2014) that drive JOLs is uncertain and may be difficult to de-couple.

However, the studies in Chapter 2 attempt to show that people appear to be integrating both sources of information to form their judgments. In particular, Blake and Castel (2018, in Study 1) demonstrates that changing participants' beliefs is not enough to reverse the font-size effect, suggesting there are separate effects of perceptual fluency. Study 2 extends this finding by showing that even when paired with a very diagnostic cue (value) font-size still biases learners.

The fluency-belief debate aside, it is abundantly clear that very often JOLs are associated more with performance than learning (see Soderstrom & Bjork, 2015, for a review). In fact, it is clear that for a non-trivial portion of the research, JOLs are dissociated with learning entirely. Even in those cases where the pattern matches recall (e.g. Arbuckle & Cuddy,

1969; Castel et al., 2007), it is unlikely that these judgments do any more than track current performance. The most accurate judgment of future performance that a person can get is to make a JOL at a delay sufficient to have eluded the effects of cues at encoding. In Experiment 1 and Experiment 4, delayed EOLs almost perfectly tracked recall performance, but immediate judgments were not sensitive to learning. The theme of JOLs tracking current performance or difficulty runs through Chapter 3, but more importantly these effects are generalized to PAL for medications and their side-effects.

Though delayed metacognitive judgments are generally very accurate, there are many times when people are susceptible to metacognitive illusions when encoding is far, far in the past. These illusions arise for the same reasons that we see them crop up at encoding: participants attend to the wrong cues either because diagnostic cues are not present or because participants do not think to call on them. In the case of the Apple logo (Blake et al., 2015) and the US flag (Blake & Castel, under revision), participants are very overconfident in their estimation of how well they know the items. Experiment 2 shows that when participants think of the item before attempting to identify it, they sometimes have poorer performance than when they just attempt the identification. It is unclear as yet how thinking about the item affects metacognitive predictions, but Blake et al. (2015) might suggest it is safe to say their overconfidence will increase. Experiment 3 proposes a method for both helping participants calibrate their metacognitive judgments correctly and may promote better learning by forcing participants to recognize what they do not know.

Lastly, consider the important role of *disfluency* in these studies. Though fluency and attention to fluency has been considered at each point, disfluency has been the primary drive in facilitating accurate judgments of learning and confidence. In Chapter 2 the tests between lists serve as a source of disfluency. Calibration of metacognitive judgments is much better over the second list than the first list. In Chapter 3 it was shown that people are sensitive to conceptual disfluency in terms of framing: information framed in a way to to make the relationships seem arbitrary is judged as harder to learn (disfluent) leading to better recall and better metamemory. Finally, a core consideration in Judgments Made About Long-Term

Incidental Encoding Experiment 3 was creating disfluency to foster better memory through error-generation. Though much research addresses how participants use "positive" fluency, people seem to be more receptive to disfluency as a debiasing agent.

Why are people more receptive to disfluency? When considering positively biased effects (i.e. whether you have *learned* something) it is hard to make a good judgment without a true test of your knowledge. However, when there is immediate feedback, even if only subjective and self-assessed, it is much more clear when you have done something wrong. This sense of "I was wrong" feels subjectively stronger than "I think I was correct."

In all, these studies examine the ways in which people fall into metacognitive traps. People often need to make estimations of their future performance and we know there are causal links between JOLs and control of study Metcalfe (2009), Metcalfe and Finn (2008), Nelson and Narens (1990) and that these judgments have consequences on learning (e.g. Rhodes & Castel, 2009; Thiede et al., 2003). As Mazzoni and Nelson so aptly put, "The overall pattern of findings suggests that JOLs are theoretically rich and are based on more than whatever underlies the likelihood of recall" (Mazzoni & Nelson, 1995, p. 1263). Understanding this richness is the important task that this line of research approaches.

APPENDIX A

From Chapter 3

Table A.1

Pairings used in Chp. 3 Experiment 1

Drug	Side-Effect	Lithuanian	English
Clavosec	weight gain	tvora	fence
Tretonin	nausea	stalas	table
Durapam	anxiety	ziedas	ring
Avaridol	insomnia	puodelis	cup
Efixar	numbness	plaukas	hair
Dypraxa	dry mouth	turgus	market
Promazatol	itching	sokis	dance
Volexum	blurred vision	kambarys	room
Gambutrol	arm pain	sriuba	soup
Phedrazone	hives	urvas	cave
Metazine	cough	kunigas	priest
Doloxan	clumsiness	kareivis	soldier
Cordrazine	fatigue	padanga	tire
Baclofen	flushing	pyragas	cake
Lamictal	fever	riteris	knight
Deltasone	calm	lietus	rain
Scopalomine	loss of appetite	arbata	tea
Antivert	pain in gums	masina	car
Fioricet	fainting	langas	window
Zelnorm	swelling	vinis	nail
Aldactone	vomiting	medus	honey
Glucophage	diarrhea	adata	needle
Cordarone	short breath	zirkles	scissors
Nardil	loss of strength	augalas	plant
Synthroid	bloating	kumpis	ham
Norvasc	painful urination	vilkas	wolf

Table A.2

Fictitious medications used in Chp. 3

Name	Familiarity	Realness
Blovestra	1.79	2.69
Imobatine	1.87	2.69
Declavyl	1.85	2.74
Lipal	2.05	2.74
Taralin	1.93	2.83
Melapor	1.88	2.85
Kalocin	1.61	2.86
Rolidal	2.00	2.89
Qualex	1.91	2.91
Azanox	1.83	2.91
Lypamol	2.14	3.00
Pylacor	2.05	3.11
Byphodine	1.79	3.12
Volapram	2.11	3.14
Efixar	2.71	3.29
Axorbil	1.86	3.32
Tretonin	2.53	3.37
Clavosec	2.29	3.38
Durapam	2.45	3.45
Dypraxa	2.54	3.54
Metazine	2.93	3.6
Gambutrol	2.78	3.63
Avaridol	2.33	3.64
Phedrazone	2.68	3.81
Volexum	2.52	3.82
Cordrazine	3.00	3.91
Promazatol	2.24	3.94
Doloxan	2.66	4.06
Ambibutrol	3.17	4.14
Hydroxil	3.89	4.52

Table A.3

Side effects used in Chp. 3

Name	Concern	Familiarity
calm	1.06	3.82
dry mouth	1.82	3.82
itching	2.18	3.85
cough	2.18	3.94
flushing	2.29	2.82
fatigue	2.41	4.03
bloating	2.44	3.47
pain in gums	2.47	2.62
loss of appetite	2.47	3.59
clumsiness	2.47	2.88
nausea	2.5	4.03
diarrhea	2.68	4.03
arm pain	2.71	2.82
insomnia	2.76	3.76
hives	2.85	3.09
anxiety	2.88	3.56
fever	3.00	3.68
vomiting	3.09	2.90
swelling	3.09	3.15
numbness	3.15	2.88
loss of strength	3.29	2.59
painful urination	3.29	3.00
weight gain	3.35	2.71
blurred vision	3.50	2.88
short breath	3.82	2.82
fainting	3.82	2.53

APPENDIX B

From Chapter 4

Below are the flag stimuli used in Chapter 4.1. A label is given below each flag indicating which features are correct or incorrect (see Table 4.1). For example, FTF is short for false–true–false indicating that the first and third features are incorrect, but the second is correct. (Flags start on next page.)

Figure B.1. Flag alternatives for the United States of America.

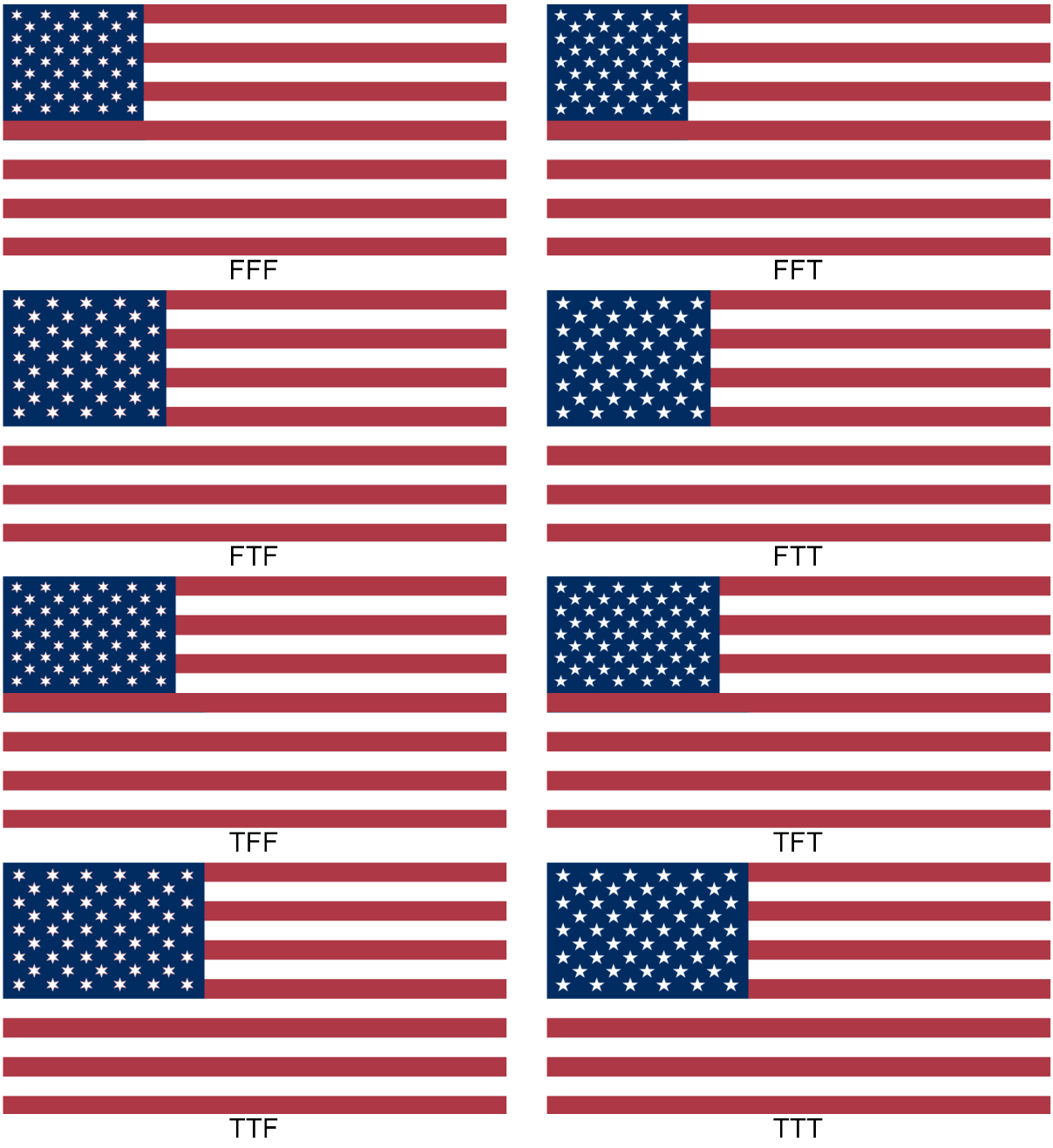


Figure B.2. Flag alternatives for Mexico.

Figure B.3. Flag alternatives for Canada.

REFERENCES

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235. doi:10.1177/1088868309341564

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063–1087. doi:10.1037/0278-7393.20.5.1063

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, *81*(1), 126–131. doi:10.1037/h0027455

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General*, *138*(3), 432–447. doi:10.1037/a0015928

Aubuchon, V. (2015). Prescription drug side effects summary. Retrieved November 20, 2016, from http://www.vaughns-1-pagers.com/medicine/prescription-drug-side-effects.htm

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press. Retrieved from http://www.langtoninfo.com/web%7B%5C_%7Dcontent/9780521709187%7B%5C_%7Dfrontmatter.pdf

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi:10.1016/j.jml.2007.12.005. arXiv: 1011.1669v3

Baddeley, A. D., & Longman, D. J. A. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics*, *21*(8), 627–635.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459. doi:10.3758/BF03193014

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi:10.1016/j.jml.2012.11.001. arXiv: 1207.1916

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). doi:10.18637/jss.v067.i01

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, *28*(5), 610–632. doi:10.1016/0749-596X(89)90016-8

Begg, I., Vinski, E., Frankovich, L., & Holgate, B. (1991). Generating makes words memorable, but so does effective reading. *Memory & Cognition*, *19*(5), 487–497.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition*, *31*(2), 297–305. doi:10.3758/BF03194388

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. doi:10.1037/0096-3445.127.1.55

Besken, M., & Mulligan, N. W. (2013). Easily perceived, easily remembered? Perceptual interference produces a double dissociation between metamemory and memory performance. *Memory & Cognition*, *41*(6), 897–903. doi:10.3758/s13421-013-0307-8

Besken, M., & Mulligan, N. W. (2014). Perceptual fluency, auditory generation, and metamemory: Analyzing the perceptual fluency hypothesis in the auditory modality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 429–440. doi:10.1037/a0034407

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 55–64).

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (Chap. 9, pp. 185–205). Cambridge, MA: MIT Press.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In A. Koriat & D. Gopher (Eds.), *Attention and performance xvii: Cognitive regulation of performance: Interaction of theory and application* (Chap. 15, pp. 435–459). Cambridge, MA: MIT Press.

Blake, A. B., & Castel, A. D. (under revision). Memory and availability-biased metacognitive illusions for flags of varying familiarity. *Memory & Cognition*.

Blake, A. B., & Castel, A. D. (2015). Metamemory. In S. K. Whitbourne (Ed.), *The encyclopedia of adulthood and aging* (1st ed., pp. 1–5). doi:10.1002/9781118521373.wbeaa043

Blake, A. B., & Castel, A. D. (2018). On belief and fluency in the construction of judgments of learning: Assessing and altering the direct effects of belief. *Acta Psychologica*, *186*(May 2018), 27–38. doi:10.1016/j.actpsy.2018.04.004

Blake, A. B., Nazarian, M., & Castel, A. D. (2015). The Apple of the mind's eye: Everyday attention, metamemory, and reconstructive memory for the Apple logo. *The Quarterly Journal of Experimental Psychology*, *68*(5), 858–865. doi:10.1080/17470218.2014.1002798

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14325–9. doi:10.1073/pnas.0803390105

Brill, E., & Moore, R. C. (2000). An improved error model for noisy channel spelling correction. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, (ACL '00), 286–293. doi:10.3115/1075218.1075255

Camerer, C., Babcock, L., Loewenstein, G., & Thaler, R. (1997). Labor supply of New York City cabdrivers: One day at a time. *The Quarterly Journal of Economics*, *112*(2), 407–441. doi:10.1162/003355397555244

Carroll, M., Nelson, T. O., & Kirwan, A. (1997). Tradeoff of semantic relatedness and degree of overlearning: Differential effects on metamemory and on long-term retention. *Acta Psychologica*, *95*(3), 239–253. doi:10.1016/S0001-6918(96)00040-6

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition*, *36*(2), 429–437. doi:10.3758/MC.36.2.429

Castel, A. D., McCabe, D. P., & Roediger, H. L. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin & Review*, *14*(1), 107–111. doi:10.3758/bf03194036

Castel, A. D., McGillivray, S., & Friedman, M. C. (2012). Metamemory and memory efficiency in older adults: Learning about the benefits of priority processing and value-directed remembering. In M. Naveh-Benjamin & N. Ohta (Eds.), *Memory and aging: Current issues and future directions* (pp. 245–270). New York, NY: Psychology Press.

Castel, A. D., Nazarian, M., & Blake, A. B. (2015). Attention and incidental memory in everyday settings. In J. M. Fawcett, E. F. Risko, & A. Kingstone (Eds.), *The handbook of attention* (pp. 463–483). Cambridge, MA: MIT Press.

Castel, A. D., Vendetti, M., & Holyoak, K. J. (2012). Fire drill: Inattentional blindness and amnesia for the location of fire extinguishers. *Attention, Perception & Psychophysics*, *74*(7), 1391–6. doi:10.3758/s13414-012-0355-3

Catlin, A., Cowan, C., Hartman, M., Heffler, S., Barron, M. C., Lassman, D., ... Whittle, L. (2008). National health spending in 2006: A year of change for prescription drugs. *Health Affairs*, *27*(1), 14–29. doi:10.1377/hlthaff.27.1.14

Coane, J. H., & Balota, D. A. (2009). Priming the holiday spirit: Persistent activation due to extraexperimental experiences. *Psychonomic Bulletin & Review*, *16*(6), 1124–1128. doi:10.3758/PBR.16.6.1124

Crawford, V. P., & Meng, J. (2011). New York city cab drivers' labor supply revisited: Reference-dependent preferences with rational-expectations targets for hours and income. *American Economic Review*, *101*(5), 1912–1932. doi:10.1257/aer.101.5.1912

Cyr, A.-A., & Anderson, N. D. (2015). Mistakes as stepping stones: Effects of errors on episodic memory among younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 841–850. doi:10.1037/xlm0000073

Dohle, S., & Montoya, A. K. (2017). The dark side of fluency: Fluent names increase drug dosing. *Journal of Experimental Psychology: Applied*, *23*(3), 231–239. doi:10.1037/xap0000131

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality and development*. Philadelphia, PA: Taylor & Francis.

Ebbinghaus, H. (1913). Retention as a function of the number of repetitions (H. A. Ruger & C. E. Bussenius, Trans.). In, *Memory: A contribution to experimental psychology* (pp. 52–61). doi:10.1037/10011-006

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign-language vocabulary learning. *Language Learning*, *43*(December), 559–617. doi:DOI10.1111/j.1467-1770.1993.tb00627.x

Fenesi, B., Sana, F., & Kim, J. A. (2014). Evaluating the effectiveness of combining the use of corrective feedback and high-level practice questions. *Teaching of Psychology*, *41*(2), 135–143. doi:10.1177/0098628314530344

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language*, *58*(1), 19–34. doi:10.1016/j.jml.2007.03.006

Fox, J., & Weisberg, S. (2011). *Companion to applied regression*. Thousand Oaks, CA: Sage. Retrieved from http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Frank, D. J., & Kuhlmann, B. G. (2017). More than just beliefs: Experience and beliefs jointly contribute to volume effects on metacognitive judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(5), 680–693. doi:10.1037/xlm0000332

Friedman, M. C., McGillivray, S., Murayama, K., & Castel, A. D. (2015). Memory for medication side effects in younger and older adults: The role of subjective and objective importance. *Memory & Cognition*, *43*(2), 206–215. doi:10.3758/s13421-014-0476-0

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin*, *119*(1), 159–165. doi:10.1037//0033-2909.119.1.159

Grimaldi, P. J., Pyc, M. A., & Rawson, K. A. (2010). Normative multitrial recall performance, metacognitive judgments, and retrieval latencies for Lithuanian-English paired associates. *Behavior Research Methods*, *42*(3), 634–42. doi:10.3758/BRM.42.3.634

Groninger, L. D. (1979). Predicting recall: The "feeling-that-I-will-know" phenomenon. *The American Journal of Psychology*, *92*(1), 45–58. doi:10.2307/1421478

Hanczakowski, M., Zawadzka, K., Pasek, T., & Higham, P. A. (2013). Calibration of metacognitive judgments: Insights from the underconfidence-with-practice effect. *Journal of Memory and Language*, *69*(3), 429–444. doi:10.1016/j.jml.2013.05.003

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 22–34. doi:10.1037/0278-7393.29.1.22

Hertzog, C., Hines, J. C., & Touron, D. R. (2013). Judgments of learning are influenced by multiple cues in addition to memory for past test accuracy. *Archives of Scientific Psychology*, *1*(1), 23–32. doi:10.1037/arc0000003

Iancu, I., & Iancu, B. (2017). Recall and recognition on minimalism. A replication of the case study on the Apple logo. *Kome*, *5*(2), 57–70. doi:10.17646/KOME.2017.24

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. doi:10.1016/j.jml.2007.11.007

Janiszewski, C., & Meyvis, T. (2001). Effects of brand logo complexity, repetition, and spacing on processing fluency and judgment. *Journal of Consumer Research*, *28*(1), 18–32. doi:10.1086/321945

Jemstedt, A., Schwartz, B. L., & Jönsson, F. U. (2017). Ease-of-learning judgments are based on both processing fluency and beliefs. *Memory*, *8211*, 1–9. doi:10.1080/09658211.2017.1410849

Kang, S. H., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology*, *103*(1), 48–59. doi:10.1037/a0021977

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, *7*(8), 368–373. doi:10.1016/S1364-6613(03)00158-X

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, *32*(1), 1–24. doi:10.1006/jmla.1993.1001

Kelley, C. M., & Rhodes, M. G. (2002). Making sense and nonsense of experience: Attributions in memory and judgment. *Psychology of Learning and Motivation*, *41*, 293–320. doi:10.1016/S0079-7421(02)80010-X

Kessels, R. P. C. (2003). Patients' memory for medical information. *Journal of the Royal Society of Medicine*, *96*(5), 219–222. doi:10.1177/014107680309600504

Koriat, A. (1981). Semantic facilitation in lexical decision as a function of prime-target association. *Memory & Cognition*, *9*(6), 587–598. doi:10.3758/BF03202353

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. doi:10.1037//0096-3445.126.4.349

Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory & Cognition*, *36*(2), 416–428. doi:10.3758/MC.36.2.416

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 187–94. doi:10.1037/0278-7393.31.2.187

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, *133*(4), 643–656. doi:10.1037/0096-3445.133.4.643

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, *52*(4), 478–492. doi:10.1016/j.jml.2005.01.001

Koriat, A., Ma'ayan, H., Sheffer, L., & Bjork, R. A. (2006). Exploring a mnemonic debiasing account of the underconfidence-with-practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 595–608. doi:10.1037/0278-7393.32.3.595

Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(1), 106–114. doi:10.1037/a0033699

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592. doi:10.1111/j.1467-9280.2008.02127.x

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, *138*(4), 449–468. doi:10.1037/a0017350

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science*, *22*(6), 787–794. doi:10.1177/0956797611407929

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Dartmouth Publishing Group.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). doi:10.18637/jss.v082.i13

Lenth, R. (2018). emmeans: Estimated marginal means, aka least-squares means. Retrieved from https://cran.r-project.org/package=emmeans

Ley, P. (1988). *Communicating with Patients. Improving Communication, Satisfaction and Compliance*. doi:10.1080/10550880903203378

Luna, K., Martín-Luengo, B., & Albuquerque, P. B. (2017). Do delayed judgments of learning reduce metamemory illusions? A meta-analysis. *The Quarterly Journal of Experimental Psychology*, 1–11. doi:10.1080/17470218.2017.1343362

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(3), 671–685. doi:10.1037/a0018785

Marmie, W. R., & Healy, A. F. (2004). Memory for common objects: brief intentional study is sufficient to overcome poor recall of US coin features. *Applied Cognitive Psychology*, *18*(4), 445–453. doi:10.1002/acp.994

Martin, M., & Jones, G. V. (1998). Generalizing everyday memory: Signs and handedness. *Memory & Cognition*, *26*(2), 193–200. doi:10.3758/BF03201132

Matheson, J. R. (1980). *Canada's flag: A search for a country*. Boston, Mass.: GK Hall.

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs? *Memory & Cognition*, *29*(2), 222–233. doi:10.3758/BF03194916

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition*, *18*(2), 196–204. doi:10.3758/BF03197095

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1263–1274. doi:10.1037//0278-7393.21.5.1263

McGillivray, S., & Castel, A. D. (2011). Betting on memory leads to metacognitive improvement by younger and older adults. *Psychology and Aging*, *26*(1), 137–42. doi:10.1037/a0022681

McGuire, L. C. (1996). Remembering what the doctor said: Organization and adults' memory for medical information. *Experimental Aging Research*, *22*(4), 403–428. doi:10.1080/03610739608254020

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology*, *15*(6), 603–616. doi:10.1002/acp.728

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, *18*(3), 159–163. doi:10.1111/j.1467-8721.2009.01628.x

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174–179. doi:10.3758/PBR.15.1.174

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language*, *52*(4), 463–477. doi:10.1016/j.jml.2004.12.001

Miele, D. B., Finn, B., & Molden, D. C. (2011). Does easily learned mean easily remembered?: It depends on your beliefs about intelligence. *Psychological Science*, *22*(3), 320–324. doi:10.1177/0956797610397954

Miller, T. M., & Geraci, L. (2014). Improving metacognitive accuracy: How failing to retrieve practice items reduces overconfidence. *Consciousness and Cognition*, *29*, 131–140. doi:10.1016/j.concog.2014.08.008

Mueller, M. L., & Dunlosky, J. (2017). How beliefs can impact judgments of learning: Evaluating analytic processing theory with beliefs about fluency. *Journal of Memory and Language*, *93*, 245–258. doi:10.1016/j.jml.2016.10.008

Mueller, M. L., Dunlosky, J., & Tauber, S. K. (2015). Why is knowledge updating after task experience incomplete? Contributions of encoding experience, scaling artifact, and inferential deficit. *Memory & Cognition*, *43*(2), 180–192. doi:10.3758/s13421-014-0474-2

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*, *70*, 1–12. doi:10.1016/j.jml.2013.09.007

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*, *20*(2), 378–384. doi:10.3758/s13423-012-0343-6

Murayama, K., Blake, A. B., Kerr, T., & Castel, A. D. (2016). When enough is not enough: Information overload and metacognitive decisions to stop studying information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(6), 914–924. doi:10.1037/xlm0000213

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. doi:10.1111/j.2041-210x.2012.00261.x. arXiv: 2746

Naveh-Benjamin, M. (2015). Associative Deficit Hypothesis. In *The encyclopedia of adulthood and aging* (pp. 1–5). doi:10.1002/9781118521373.wbeaa287

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science*, *2*(4), 267–270. doi:10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Dunlosky, J. (1992). How shall we explain the delayed-judgment-of-learning effect? *Psychological Science*, *3*(5), 317–318. doi:10.1111/j.1467-9280.1992.tb00681.x

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science*, *5*(4), 207–213. doi:10.1111/j.1467-9280.1994.tb00502.x

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 676–686. doi:10.1037//0278-7393.14.4.676

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 26, pp. 125–173). doi:10.1016/S0079-7421(08)60053-5

Nguyentan, C., Blake, A. B., & Castel, A. D. (2017). *Memory and metamemory for medication side effects in context and out of context in younger adults*. Los Angeles, CA.

Nickerson, R. S. (1965). Short-term memory for complex meaningful visual configurations: A demonstration of capacity. *Canadian Journal of Psychology*, *19*(2), 155–160. doi:10.1037/h0082899

Nickerson, R. S., & Adams, M. J. (1979). Long-term memory for a common object. *Cognitive Psychology*, *11*(3), 287–307. doi:10.1016/0010-0285(79)90013-6

Nilsson, L., Adolfsson, R., Bäckman, L., Frias, C. M., Molander, B., & Nyberg, L. (1997). The betula prospective cohort study: Memory, health, and aging. *Aging, Neuropsychology, and Cognition*, *11*(2-3), 134–148. doi:10.1080/13825589708256633

Paivio, A. (1986). *Mental representations: A dual coding approach*. New York, NY: Oxford University Press.

Paivio, A., Rogers, T. B., & Smythe, P. C. (1968). Why are pictures easier to recall than words? *Psychonomic Science*, *11*(4), 137–138. doi:10.3758/BF03331011

Qato, D. M., Caleb, A. G., Conti, R. M., Johnson, M., Schumm, P., & Lindau, S. T. (2008). Use of prescription and over-the-counter medications and dietary supplements among older adults in the United States. *JAMA*, *300*(24), 2867–2878. doi:10.1001/jama.2008.892

Quené, H., & Van Den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, *43*(1-2), 103–121. doi:10.1016/j.specom.2004.02.004

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413–425. doi:10.1016/j.jml.2008.02.002

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.r-project.org/

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky & S. K. Tauber (Eds.), *Oxford handbook of metamemory* (pp. 65–80). New York: Oxford University Press.

Rhodes, M. G., & Castel, A. D. (2008a). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, *137*(4), 615–625. doi:10.1037/a0013684

Rhodes, M. G., & Castel, A. D. (2008b). Metacognition and part-set cuing: can interference be predicted at retrieval? *Memory & Cognition*, *36*(8), 1429–1438. doi:10.3758/MC.36.8.1429

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review*, *16*(3), 550–554. doi:10.3758/PBR.16.3.550

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin*, *137*(1), 131–48. doi:10.1037/a0021705

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. doi:10.1037/a0016496

Rinck, M. (1999). Memory for everyday objects: where are the digits on numerical keypads? *Applied Cognitive Psychology*, *13*(4), 329–350. doi:10.1002/(SICI)1099-0720(199908)13:4⟨329::AID-ACP583⟩3.0.CO;2-3

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., & Schmidt, S. R. (1980). Output interference in the recall of categorized and paired-associate lists. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(1), 91–105. doi:10.1037/0278-7393.6.1.91

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. doi:10.1037/a0037559

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. doi:10.1016/S0364-0213(02)00078-2

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*(1), 36–71. doi:10.1016/0010-0285(90)90003-M

Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1997). The inferential and experiential bases of metamemory. *Current Directions in Psychological Science*, *6*(5), 132–137. doi:10.1111/1467-8721.ep10772899

Schwartz, B. L., & Efklides, A. (2012). Metamemory and memory efficiency: Implications for student learning. *Journal of Applied Research in Memory and Cognition*, *1*(3), 145–151. doi:10.1016/j.jarmac.2012.06.002

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, *61*(2), 195–202. doi:10.1037//0022-3514.61.2.195

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning & Memory*, *4*(6), 592–604. doi:10.1037/0278-7393.4.6.592

Snyder, K. M., Ashitaka, Y., Shimada, H., Ulrich, J. E., & Logan, G. D. (2014). What skilled typists don't know about the QWERTY keyboard. *Attention, perception & psychophysics*, *76*(1), 162–71. doi:10.3758/s13414-013-0548-4

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176–199.

Soderstrom, N. C., & McCabe, D. P. (2011). The interplay between value and relatedness as bases for metacognitive monitoring and control: Evidence for agenda-based monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1236–1242. doi:10.1037/a0023548

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science*, *3*(5), 315–316. doi:10.1111/j.1467-9280.1992.tb00680.x

Sungkhasettee, V. W., Friedman, M. C., & Castel, A. D. (2011). Memory and metamemory for inverted words: illusions of competency and desirable difficulties. *Psychonomic Bulletin & Review*, *18*(5), 973–978. doi:10.3758/s13423-011-0114-9

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*(1), 66–73. doi:10.1037/0022-0663.95.1.66

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *25*(4), 1024–1037. doi:10.1037//0278-7393.25.4.1024

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, *41*(3), 429–442. doi:10.3758/s13421-012-0274-5

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. doi:10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. doi:10.1126/science.185.4157.1124

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289–335. doi:10.1016/j.jml.2003.10.003

Van Zee, A. (2009). The promotion and marketing of oxycontin: Commercial triumph, public health tragedy. *American Journal of Public Health*, *99*(2), 221–227. doi:10.2105/AJPH.2007.131714

Vendetti, M., Castel, A. D., & Holyoak, K. J. (2013). The floor effect: Impoverished spatial memory for elevator buttons. *Attention, Perception & Psychophysics*, *75*(4), 636–43. doi:10.3758/s13414-013-0448-7

Von Restorff, H. (1933). Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychologische Forschung*, *18*(1), 299–342. doi:10.1007/bf02409636

Wammes, J. D. [Jeffrey D.], Meade, M. E., & Fernandes, M. A. (2016). The drawing effect: Evidence for reliable and robust memory benefits in free recall. *Quarterly Journal of Experimental Psychology*, *69*(9), 1752–1776. doi:10.1080/17470218.2015.1094494

Wammes, J. D. [Jeffrey D], Meade, M. E., & Fernandes, M. A. (2018). Creating a recollection-based memory through drawing. *44*(5), 734–751. doi:10.1037/xlm0000445

Werth, L., & Strack, F. (2014). An inferential approach to the knew-it-all-along phenomenon. *Memory*, *11*(4-5), 411–9. doi:10.1080/09658210244000586

Wickens, D. D. (1970). Encoding categories of words: An empirical approach to meaning. *Psychological Review*, *77*(1), 1–15. doi:10.1037/h0028569

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *2*(5-6), 440–445. doi:10.1016/S0022-5371(63)80045-6

Williams Jr., E. P. (2012). Did Francis Hopkinson design two flags? *The Quarterly Newsletter of the North American Vexillogical Association*, *216*, 7–9. Retrieved from http: //www.flagguys.com/pdf/NAVANews%7B%5C_%7D2012%7B%5C_%7Dno216.pdf

Wolfe, J. M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting memories* (pp. 71–94). Cambridge, MA: MIT Press.

Wong, K., Wadee, F., Ellenblum, G., & McCloskey, M. (2018). The Devil's in the g-tails: Deficient letter-shape knowledge and awareness despite massive visual experience. *Journal of Experimental Psychology: Human Perception and Performance*. doi:10.1037/xhp0000532

Yang, C., Potts, R., & Shanks, D. R. (2017). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *43*(7), 1073–1092. doi:10.1037/xlm0000363

Yue, C. L., Castel, A. D., & Bjork, R. A. (2013). When disfluency is—and is not—a desirable difficulty: the influence of typeface clarity on metacognitive judgments and memory. *Memory & Cognition*, *41*(2), 229–241. doi:10.3758/s13421-012-0255-8

Zaragoza, M. S., Payment, K. E., Ackil, J. K., Drivdahl, S. B., & Beck, M. (2001). Interviewing witnesses: Forced confabulation and confirmatory feedback increase false memories. *Psychological Science*, *12*(6), 473–477. doi:10.1111/1467-9280.00388