Psychology of Consciousness: Theory, Research, and Practice

© 2017 American Psychological Association
2326-5523/17/$12.00 http://dx.doi.org/10.1037/cns0000117

Variations in the Relationship Between Memory Confidence and Memory Accuracy: The Effects of Spontaneous Accessibility, List Length, Modality, and Complexity

Benedek Kurdi, Alexander J. Diaz, Caroline A. Wilmuth, Michael C. Friedman, and Mahzarin R. Banaji
Harvard University

Three experiments (total $N = 1{,}058$) were conducted to investigate the relationship between memory accuracy and subjective confidence using thematic lists constructed on the basis of spontaneous accessibility, that is, the frequency with which items are spontaneously generated as category members. After memorizing lists of words and performing a distractor task, participants completed tests of recognition memory and rated confidence in their memory for 4 item types: *studied items* (second highest spontaneous accessibility) and 3 types of nonstudied items, including *strong lures* (highest spontaneous accessibility), *weak lures* (third highest spontaneous accessibility), and *semantically unrelated items*. In Experiments 1 and 2, the items were presented as lists, whereas in Experiment 3, they were embedded in short vignettes. In Experiments 1 and 3, amount of material to be encoded (1 vs. 10 or 8 lists of 15 items), and in Experiment 2, modality of stimulus presentation at encoding and at test (visual–visual, auditory–auditory, and auditory–visual) were varied between participants. Across all experiments, the confidence–accuracy relationship remained consistently positive for studied items. However, increasing the amount of information to be memorized, using inconsistent stimulus modalities across encoding and retrieval, and embedding the items in vignettes resulted in (a) zero or negative confidence–accuracy relationships for weak lures and (b) highly negative confidence–accuracy relationships for strong lures. These results demonstrate that subjective confidence judgments are made on the basis of inferential processes in the course of which spontaneous accessibility is mistaken for memory strength.

Keywords: confidence–accuracy relationship, false memory, memory confidence, metacognition, metamemory

Human memory is both powerful and fallible. Research on memory has revealed highly accurate memory for known faces, with the capacity to detect infinitesimal distortions limited only by visual acuity (Brédart & Devue, 2006; Ge, Luo, Nishimura, & Lee, 2003), the ability to

Benedek Kurdi and Alexander J. Diaz, Department of Psychology, Harvard University; Caroline A. Wilmuth, Harvard Business School and Department of Psychology, Harvard University; Michael C. Friedman, Harvard Initiative for Learning and Teaching, Harvard University; Mahzarin R. Banaji, Department of Psychology, Harvard University.

The data reported in Experiments 2 and 3 were collected as part of a senior thesis conducted by Alexander J. Diaz and supervised by Caroline A. Wilmuth and Mahzarin R. Banaji. The research was supported by grants from the Edmond J. Safra Center for the Study of Ethics and from Harvard University to Mahzarin R. Banaji and a Phi Beta Kappa Patricia King Fellowship from the Harvard College Research Program to Alexander J. Diaz. Some of the data reported in this article were presented at the Fifth Annual Symposium for Boston-Area Graduate Students in Psychology in April 2016 and the 28th Annual Meeting of the Association for Psychological Science in May 2016. All raw data files, analysis scripts, and stimuli used in this article are available for download from the Open Science Framework (https://osf.io/9z2gp). We thank A. DeSoto and H. Roediger for generously sharing stimulus materials and providing useful feedback on previous versions of this project, L. Somerville for insightful comments, as well as A. Amedi, S. Menon, H. Rabasco, and L. Vogel for research assistance.

Correspondence concerning this article should be addressed to Benedek Kurdi, Department of Psychology, Harvard University, William James Hall 1580, 33 Kirkland Street, Cambridge, MA 02138. E-mail: kurdi@g.harvard.edu

remember narrative details of movies weeks and even months after the original viewing (Furman, Dorfman, Hasson, Davachi, & Dudai, 2007), and a long-term memory store exceeding one billion bits of information (Landauer, 1986). Furthermore, accurate recollection of the perceptual details of several thousand studied images has been documented (Standing, 1973), even when lures were not randomly selected, but rather came from the same object category as the targets (Brady, Konkle, Alvarez, & Oliva, 2008). At the same time, the past several decades of research on memory have repeatedly demonstrated the relatively trivial ways in which memory can be distorted, even under consequential conditions (e.g., Hirst & Echterhoff, 2012; Tulving & Craik, 2000; Wells & Olson, 2003; Wixted, 2004). Such distortions include, for instance, source-monitoring errors, the formation of completely erroneous autobiographical memories, and the modification of existing memories upon reactivation.

Even when memory breaks down, failure to remember may be less costly if humans' metacognitive abilities can intervene to provide accurate information on the reliability of each memory trace. In other words, if we were aware of when our memory is accurate and when it is not, we would be able to navigate rationally through life by taking appropriate steps under these differing conditions revealed by our metacognitive abilities. In fact, in surveys of opinions about memory ability, laypeople seem to assume both that their memories are usually veridical and that even when their memory fails, metacognitive monitoring will allow them to track the possibility of memory failure (Kassin, Ellsworth, & Smith, 1989; Kassin, Tubb, Hosch, & Memon, 2001; Lindsay, Wells, & Rumpel, 1981; Penrod & Cutler, 1995).

In the present research, we investigated the *confidence–accuracy relationship*, that is, the robustness of the capacity to distinguish an accurate memory from an inaccurate one or, in other words, the ability of memory confidence to track memory accuracy. When memory accuracy is high, is confidence in that memory also high? And, similarly, when memory accuracy is low, is confidence in that memory also low?

On the one hand, we might find evidence for adaptive metacognitive processes that properly discriminate between accurate and inaccurate memories under a wide range of encoding and retrieval conditions. On the other hand, we might find that the same processes that produce distortion in memory also undermine metacognitive processes that track memory accuracy. If the latter is the case, metacognitive processes may break down when they are most needed (cf. Dunning, Johnson, Ehrlinger, & Kruger, 2003; Hacker, Bol, Horgan, & Rakow, 2000; Kruger & Dunning, 1999), which would make the consequences of memory errors considerably more damning.

The question of the confidence–accuracy relationship has far-reaching theoretical implications. It places fundamental constraints on our understanding of how metacognition works, including whether people can directly access the contents of their minds, or, alternatively, whether memory strength is inferred on the basis of usually accurate but possibly misleading cues (Schwartz, 1994). In addition to its theoretical implications, the question of metacognitive accuracy also bears on a wide range of everyday situations. For instance, if metamemory is usually accurate, people can reasonably assume that the more confident a report is of a prior event, the more likely the report is to be reliable; the stronger the subjective feeling of memory is for studied material, the better one will do on an upcoming exam; the more a manager is skeptical of his or her own ability to remember the past performance of an employee, the more s/he will rely on documentation to arrive at a decision; and the more confidence a doctor has that she or he has recalled the right name of a disease from its symptoms, the more competent s/he will be.

The Confidence–Accuracy Relationship

The confidence–accuracy relationship has been investigated extensively and primarily in the context of eyewitness testimony (for reviews, see N. Brewer & Wells, 2006; Deffenbacher, 1980; Leippe, 1980; Wells & Murray, 1984; for meta-analyses, see Bothwell, Deffenbacher, & Brigham, 1987; Cutler & Penrod, 1989; Sporer, Penrod, Read, & Cutler, 1995). The overall conclusion of this literature is that the relationship between objective accuracy and subjective confidence is, at best, tenuous. In this vein, it has been shown that various manipulations can affect confidence without affecting accuracy. For instance, information about an ostensible co-witness having made the same identification judgment (Luus & Wells, 1994), other forms of confirming feedback (Wells & Bradfield, 1998), and postevent briefing from an attorney (Wells, Ferguson, & Lindsay, 1981) can boost confidence without making participants' memories any more accurate. Other manipulations can affect accuracy without affecting confidence. For instance, optimal recognition conditions are associated with higher levels of accuracy without concomitant changes in subjective confidence (Lindsay et al., 1981).

However, findings from the eyewitness testimony literature might not readily generalize to other kinds of memory. Specifically, research on memory has been characterized by a tension between internal and external validity (Banaji & Crowder, 1989, 1991), with pressures to guarantee both the internal validity and the ecological realism of research procedures. Studies that favor ecological realism and those that favor internal validity have produced differing views of the confidence–accuracy relationship.

In direct contradiction to the findings from the eyewitness memory literature, many experiments involving the simple memorization of word lists have revealed highly positive relationships between memory accuracy and subjective confidence (Arbuckle & Cuddy, 1969; Mandler & Boeck, 1974; Mickes, Hwe, Wais, & Wixted, 2011; Mickes, Wixted, & Wais, 2007). It has also been demonstrated, however, that the inclusion of potentially misleading items in a memory task can disrupt the usually positive confidence–accuracy relationship. For instance, Brewer and colleagues (W. F. Brewer & Sampaio, 2006; Brewer, Sampaio, & Barlow, 2005) investigated the relationship between confidence and accuracy in the cued verbatim recall of sentences. Some sentences were deceptive in that they were likely to produce synonym substitutions (e.g., "The class was difficult," which many participants recalled as "the class was hard"). Whereas a strong positive confidence–accuracy relationship was found for nondeceptive sentences, confidence was unrelated to accuracy for deceptive sentences.

Other studies have used the Deese–Roediger–McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995) to explore the confidence–accuracy relationship, with mixed results. In this paradigm, participants memorize lists of semantically related items and are then administered a recall or recognition task. Crucially, each list converges on a close semantic associate that is never presented at encoding. The measure of the DRM effect is the extent to which participants erroneously recall or recognize the critical lure (i.e., the nonpresented semantic associate). In some cases, a highly positive relationship was found between subjective confidence and accuracy. For instance, McKelvie (1999, 2001) asked participants to study several DRM lists and to complete recall or recognition tasks. Across all studies, participants were more confident about their memory for correctly remembered items than about critical intrusions (see also Payne, Elie, Blackwell, & Neuschatz, 1996; Read, 1996). Pirmoradi and McKelvie (2015), by contrast, found no relationship between confidence and accuracy in the DRM paradigm: Average confidence did not differ across studied items and critical lures.

Roediger and DeSoto (2014a) and DeSoto and Roediger (2014) used another list-learning paradigm involving deceptive items to study the confidence–accuracy relationship. Unlike the DRM task, the list-learning paradigm used in these studies is based on spontaneous accessibility rather than associative strength, the idea being that nonpresented items that constitute good instances of a category (such as "apple" for fruits or "dog" for animals) will be falsely recognized with high levels of subjective certainty because participants mistake chronic accessibility for prior exposure in the context of the experiment (Meade & Roediger, 2006, 2009; S. M. Smith, Ward, Tindell, Sifonis, & Wilkenfeld, 2000). Roediger and DeSoto probed the confidence–accuracy relationship using three different measures: between-participants correlations, between-events correlations, and resolution (relative accuracy within each participant). For studied (i.e., nondeceptive) items, all three correlations were positive, whereas for nonstudied chronically accessible items, the relationship was found to be mostly negative, but sometimes also positive or zero, depending on the analysis used.

The Present Project

Given that prior work has found positive (Arbuckle & Cuddy, 1969; Mandler & Boeck, 1974; Mickes et al., 2007, 2011) and zero (Pirmoradi & McKelvie, 2015) relationships between judgments of subjective confidence and objective measures of accuracy in list-learning paradigms, in the present project we sought to probe the boundary conditions of this phenomenon, relying on high-powered designs to explore a range of possible moderators. The recent studies conducted by Roediger and DeSoto (2014a) and DeSoto and Roediger (2014) were path-breaking in that they were the first ones to demonstrate all three possible kinds of relationships (positive, negative, and zero) between memory confidence and accuracy using the same general kind of material (i.e., thematic lists of words) in a tightly controlled experimental paradigm.

In the present studies, we aimed to replicate, extend, and further elucidate the results of DeSoto and Roediger (2014) and Roediger and DeSoto (2014a) in the following ways. First, to be able to induce a reasonable degree of variation across trial types, the Roediger–DeSoto experiments used a highly demanding setup: Participants studied and were tested on hundreds of items, and they were required to switch between stimulus modalities across encoding (auditory) and retrieval (visual). As we discuss in more detail below, both of these features of the experiments may have led to poor metacognitive performance. Our present experiments explicitly varied the amount of material to be studied and the modality of stimulus presentation (visual vs. auditory) at encoding and test. Second, Roediger and DeSoto used only lists of words as stimuli in their experiments. These materials have the undeniable advantage of offering superior internal validity; however, findings from such experiments may not generalize to metacognitive effects when the material involves richer stimuli. Therefore, in one experiment we embedded the to-be-memorized items in short narratives to test whether the same result would be obtained if the materials more closely resembled the type of information ordinarily encountered in social communication. Third, the total sample size across the three experiments reported here was more than six times larger than in the experiments conducted by Roediger and DeSoto (N = 166 versus N = 1,058), which—combined with the large number of items completed by each participant—provided the ability to detect any existing relationship between memory confidence and accuracy if such a relationship existed, or to confidently rule out any such relationship if one was not detected. In addition, our studies used samples from general adult populations with a wider range of cognitive abilities and cultural backgrounds. And finally, as explained in more detail in the Results section of Experiment 1 below, our analysis strategy allowed for appropriate treatment of multiple dependencies, offered a model-based, rather than purely descriptive, approach, did not require the exclusion of participants or items with zero variance, and preserved Type-I error rates even in the face of significant item effects.

Moderators of the Confidence–Accuracy Relationship

Spontaneous accessibility. Emulating DeSoto and Roediger (2014) and Roediger and DeSoto's (2014a) procedures, participants in all experiments that we report completed tests of recognition memory for both studied items and nonstudied, yet highly accessible items (lures). In line with prior work, we expected memory performance to be excellent for studied items (DeSoto & Roediger, 2014; Meade & Roediger, 2006, 2009; Roediger & DeSoto, 2014a; S. M. Smith et al., 2000). Crucially, based on the unequivocal results reported by Roediger and DeSoto, we hypothesized that the confidence–accuracy relationship would remain consistently positive for studied items. Furthermore, we anticipated high rates of false alarms for lures (DeSoto & Roediger, 2014; Meade & Roediger, 2006, 2009; Roediger & DeSoto, 2014a; S. M. Smith et al., 2000). However, in spite of poor memory performance for lures, the confidence–accuracy relationship might still remain positive, provided that reliance on metacognitive capacity leads participants to realize that lures can cause memory distortions. On the other hand, if the ability to discern the source of familiarity for an item is compromised, the confidence–accuracy relationship should break down. As mentioned above, Roediger and DeSoto (2014a) and DeSoto and Roediger (2014) reported positive, negative, and nonsignificant confidence–accuracy relationships, depending on the kind of analysis used. Therefore, one goal of the present project was to conclusively establish the nature of the confidence–accuracy relationship for lures by using large samples and adequate statistical analyses. We explored this issue across Experiments 1–3.

Amount of information. The list-length effect, that is, the result that as the amount of material to be studied increases, memory performance goes down, is one of the earliest (Strong, 1912) and most consistently replicated (Gillund & Shiffrin, 1984; Gronlund & Elam, 1994; Murnane & Shiffrin, 1991; Ratcliff, Clark, & Shiffrin, 1990; Shiffrin, Huber, & Marinelli, 1995) findings from the memory literature. In fact, when participants studied 16, rather than six, lists in the DRM paradigm, false recognition increased considerably, and hit rates and false alarm rates became indistinguishable (Roediger & McDermott, 1995). Therefore, in our studies, we expected that increasing the amount of information to be encoded would result in a detriment to memory accuracy. More important, however, was that we sought to investigate whether with the faltering of memory performance, the confidence–accuracy relationship would also break down, or if cognitive and metacognitive performance would remain unrelated. If the former were the case, given their use of large amounts of material to be studied, DeSoto and Roediger may have arrived at an overly pessimistic assessment of the confidence–accuracy relationship (DeSoto & Roediger, 2014; Roediger & DeSoto, 2014a). More specifically, under less cognitively demanding conditions, the confidence–accuracy relationship might be less strongly negative, or even positive, for deceptive items. We explored the effects of list length on metacognitive performance in Experiments 1 and 3.

Stimulus modality. Studies of memory have produced the well-established result that the modality of stimulus presentation at both encoding and retrieval influences memory accuracy. Some research has suggested that auditory presentation of stimuli leads to poorer memory performance than visual presentation, including more intrusions of nonpresented stimuli in both recognition and recall (Beauchamp, 2002; Cleary & Greene, 2002; R. E. Smith & Hunt, 1998; but see Maylor & Mo, 1999). By contrast, other studies have indicated that, in line with the encoding specificity principle (Tulving & Thomson, 1973), it is the match between encoding and retrieval that matters. When modality at encoding and retrieval is the same, superior memory performance should result, as opposed to when the modalities differ (Kellogg, 2001). In addition, evidence has also been provided for the simultaneous operation of both effects. For instance, Gallo, McDermott, Percer, and Roediger (2001) found especially high levels of false recognition for auditory encoding and visual testing and lower, but still considerable, levels of false recognition for visual encoding and auditory testing (see also Israel & Schacter, 1997). Across all their studies, Roediger and DeSoto presented stimuli auditorily at encoding and visually at test. Under all three modality hypotheses, this might have led to worse memory accuracy and metamnemonic performance compared with matched modality at learning and at test. To systematically test the presence and magnitude of modality effects, Experiment 2 included conditions with both matched and mismatched modalities across encoding and retrieval.

Complexity of information. Finally, in Experiment 3, we tested the nature of the confidence–accuracy relationship in a linguistic context that matched everyday social communications, which take the form of narratives rather than lists of words. Here, two possible hypotheses were considered. First, by virtue of creating interconnections between items, narratives have a causal structure (Black & Bern, 1981), which may make memory accuracy superior to list learning. Alternatively, the narrative format of the vignettes might cause memory performance to falter because the narratives (a) draw participants' attention to the shared category of the items included in the text, thereby potentially increasing the incidence of gist intrusions (Brainerd & Reyna, 2002), and (b) may allow task-irrelevant details to interfere with learning. Overall, manipulating complexity of information presented us with another opportunity to probe whether changes in cognitive performance are associated with concomitant changes in metacognitive performance.

Experiment 1

Experiment 1 investigated the relationship between memory confidence and memory accuracy, relying on a modified version of the paradigm used in recent experiments by DeSoto and Roediger (2014) and Roediger and DeSoto (2014a). Participants studied lists of semantically related words (e.g., musical instruments) created on the basis of spontaneous accessibility, that is, the frequency with which each item is spontaneously generated as an instance of the given category. For example, within the category of musical instruments, the item "violin" is highly spontaneously accessible, whereas the item "cymbal" is much less so. Subsequently, participants were tested for recognition memory and provided accompanying confidence ratings for previously studied items as well as some novel items (lures) that systematically varied in spontaneous accessibility.

In addition to the effects of spontaneous accessibility probed by Roediger and DeSoto, Experiment 1 also explicitly explored the effects of list length on the confidence–accuracy relationship. If cognitive and metacognitive performance were tethered to each other, increasing the number of lists to be studied should interfere not only with participants' memory accuracy (Gillund & Shiffrin, 1984; Gronlund & Elam, 1994; Murnane & Shiffrin, 1991; Ratcliff et al., 1990; Shiffrin et al., 1995; Strong, 1912) but also with the relationship between subjective confidence and accuracy. Alternatively, cognitive and metacognitive performance might not be modulated by the same variables, and thus the confidence–accuracy relationship could remain unaffected despite the expected detriment to memory accuracy when more information is to be remembered. Finally, unlike in the DeSoto–Roediger experiments, stimuli were presented in consistent (visual–visual) modalities at encoding and at test, thus ruling out the possibility that the results observed were exclusively due to auditory stimulus presentation (Beauchamp, 2002; Cleary & Greene, 2002; R. E. Smith & Hunt, 1998), mixed modalities across encoding and recognition (Kellogg, 2001), or both (Gallo et al., 2001; Israel & Schacter, 1997).

Method

Participants. 272 users from Amazon.com's Mechanical Turk (MTurk) were paid \$0.50 or \$4.00 (depending on condition) in exchange for their participation. To ensure adequate command of the English language and compliance with task instructions, participation was restricted to users with residence in the United States and an approval rate of at least 90% on previous MTurk assignments. Participants completed the experiment online using their own computers.

Materials. The materials were 10 lists of semantically related words, created by DeSoto and Roediger (2014) by asking participants to list as many members as possible of a given category (e.g., a fruit or a part of the human body).[1] The 25 most highly accessible (i.e., most frequently reported) members of those categories were included in the lists. Specifically, the five most highly accessible category members were used as strong lures, the next 15 most highly accessible category members were used as studied items, and the last five category members were used as weak lures. In addition to the 25 semantically related words, each list included five words that were unrelated to the semantic category of the list, resulting in a total of 30 words per list (i.e., 15 studied items, five strong lures, five weak lures, and five unrelated items).

Design and procedure. The experiment consisted of three phases: a study phase, a distractor task, and a test phase. At the beginning of the experiment, participants were pseudorandomly assigned to one of two between-participants conditions determining how many lists they would be asked to memorize in the study phase and how many items they would be asked to classify as old or new in the test phase. More participants ($N = 208$) were assigned to the one-list condition than to the 10-list condition ($N = 64$), resulting in a comparable number of trials across the two conditions. In the study phase, participants memorized either one list or 10 lists of words, depending on condition. The distractor task required participants to engage in a relatively cognitively demanding but unrelated activity, thus allowing for memory consolidation and forgetting to occur. In the test phase, participants were asked to provide recognition judgments for studied and nonstudied items and to indicate their level of confidence in their recognition memory.

Study phase. Participants in the one-list condition were instructed that they would see a list of words they would have to memorize. On the following screen, they were presented with 15 semantically related items to study. Each

[1] The full list of stimuli is available for download from the Open Science Framework (OSF; https://osf.io/9z2gp).

participant was presented with items from only one list, randomly selected from the 10 lists of semantically related words. The list of words was presented simultaneously, with all 15 words presented in randomized order that was kept consistent across participants. Participants studied the word list at their own pace. Once they proceeded, they were not allowed to return to the previous screen. The 10-list condition was similar, with the exception that participants studied and were tested on all 10 word lists rather than a single randomly selected list. Participants in this condition were presented with the 10 lists in individually randomized order. They advanced to the next list at their own pace.

Distractor task. After studying the list(s) of words, all participants regardless of condition completed a 5-min distractor task, in which they were instructed to list all United States presidents in chronological order (Roediger & Crowder, 1976; Roediger & DeSoto, 2014b). Participants were asked to enter their responses in a textbox while a timer counted down the remaining time. Following the distractor task, the experiment automatically advanced to the instructions for the test phase.

Test phase. After completing the distractor task, participants were instructed that they would be asked to distinguish previously seen words from unseen words by clicking a *Yes* or *No* button, followed by a rating of their confidence in their recognition memory using a sliding scale that ranged from 0 (*lowest confidence*) to 100 (*highest confidence*). For each item, the slider was initially set to 50. Participants entered their responses at their own pace. The test did not proceed to the next screen until both the recognition judgment and the confidence rating were provided. To avoid errors due to inattention, participants who wished to enter 50 as their confidence judgment had to confirm their choice by clicking on the slider. Participants in the one-list condition were presented with 30 items, including the 15 studied items and 15 novel items (five strong lures, five weak lures, and five unrelated words), in individually randomized order. Participants in the 10-list condition were tested on 300 words, 30 per list (15 studied items, five strong lures, five weak lures, and five unrelated items). The presentation order was individually randomized such that words from all 10 lists were mixed together.

After entering their last recognition judgment and confidence rating, participants answered some demographic questions, were debriefed, and compensated.

Results[2]

Descriptive statistics. Descriptive statistics, including accuracy rates (overall and by trial type), discriminability indices (comparing studied items to the three other trial types), and mean confidence levels (overall and by trial type), are reported in Table 1. In line with previous research (DeSoto & Roediger, 2014; Roediger & DeSoto, 2014a), memory in both conditions was most accurate for unrelated items, followed by studied items, weak lures, and strong lures. Also in line with previous research (Gillund & Shiffrin, 1984; Gronlund & Elam, 1994; Murnane & Shiffrin, 1991; Ratcliff et al., 1990; Shiffrin et al., 1995; Strong, 1912), a list-length effect emerged, such that memory accuracy was considerably higher in the one-list condition than in the 10-list condition. In the one-list condition, confidence judgments exhibited the same pattern as accuracy, with the highest confidence ratings given to unrelated items, followed by studied items, weak lures, and strong lures. In the 10-list condition, confidence judgments were overall lower and less differentiated by trial type than in the one-list condition. These descriptive measures confirm the soundness of the experimental design and manipulation.

Modeling strategy. We investigated the confidence–accuracy relationship using mixed-effects modeling (Baayen, Davidson, & Bates, 2008). Mixed-effects models offer at least four clear advantages over the measures of statistical relatedness routinely used in metacognition research, including Goodman–Kruskal's γ (also known as resolution; Goodman & Kruskal, 1954) and Pearson's r.

First, between-subjects and between-events correlations rely on the summary statistic method, with the first one collapsing across item-level dependencies and the second one collapsing across participant-level dependencies in the data, thus losing valuable information. By contrast, mixed-effects models have the ability

[2] All raw data and analysis scripts are available for download from the OSF (https://osf.io/9z2gp).

Table 1
Descriptive Statistics for All Experiments

	Presentation				Accuracy					*d′*			Confidence				
Exp	Encoding	Test	*N*	List	Overall	SI	UR	SL	WL	UR	SL	WL	Overall	SI	UR	SL	WL
1	Visual	Visual	208	1	85%	86%	99%	67%	86%	1.08	1.08	1.41	82 (25)	84 (23)	94 (16)	70 (28)	75 (26)
1	Visual	Visual	64	10	64%	71%	74%	38%	55%	1.24	.22	.80	64 (29)	66 (29)	62 (29)	62 (29)	60 (28)
2	Visual	Visual	146	5	73%	77%	92%	46%	67%	1.78	.61	1.16	73 (27)	76 (26)	74 (27)	69 (27)	66 (27)
2	Auditory	Auditory	200	5	71%	73%	86%	51%	69%	1.90	.66	1.28	72 (28)	76 (27)	72 (30)	67 (28)	65 (28)
2	Auditory	Visual	200	5	70%	71%	92%	46%	67%	1.52	.50	1.19	71 (28)	75 (27)	68 (30)	67 (27)	64 (28)
3	Vignette	Visual	170	1	72%	69%	92%	52%	76%	.97	.50	.88	74 (28)	75 (27)	79 (27)	69 (28)	71 (28)
3	Vignette	Visual	70	8	72%	73%	91%	52%	73%	2.05	.74	1.34	66 (27)	74 (26)	57 (27)	62 (25)	58 (25)

Note. Exp = Experiment; SI = studied items; UR = unrelated items; SL = strong lures; WL = weak lures. For each experimental condition, we report modality of stimulus presentation at encoding and at test, sample size, the number of thematic lists studied, accuracy (overall and broken down by trial type, rounded to integers), discriminability indices (*d′*) comparing studied items to the other trial types (rounded to two decimal places), and confidence ratings (measured on a 0–100 scale, rounded to integers). Accuracy is expressed in terms of percentages, whereas for confidence ratings we report means and, in parentheses, standard deviations.

to detect confidence–accuracy relationships while simultaneously adjusting for item-level dependencies (i.e., the fact that average levels of accuracy may differ across items) and participant-level dependencies (i.e., the fact that participants may differ in terms of their average levels of accuracy and in terms of the extent to which their confidence judgments predict accuracy).

Second, as opposed to purely descriptive measures such as Goodman–Kruskal's γ or Pearson's *r*, mixed-effects modeling offers a model-based approach to data analysis. Thus, using mixed-effects modeling allows researchers to calculate (comparative) model fit indices, as well as point estimates, confidence intervals, and *p* values for individual model parameters, and predicted values of the response variable given a certain configuration of the independent variables. For instance, unlike γ or Pearson's *r*, mixed-effects modeling makes it meaningful to ask questions like what is the best estimate for the conditional probability of an accurate response given that the item is a strong lure with a confidence rating of 56 on a 0–100 scale.

Third, it is impossible to calculate Goodman–Kruskal γ and other descriptive measures of metacognitive accuracy for participants who are either at ceiling or at floor within a given trial type. This issue is far from trivial. For example, in the one-list condition of Experiment 1 (discussed in more detail below), overall resolution could be calculated for only 39 out of 208 participants and, for a certain trial type, no more than one single participant. Thus, this measure conveys a heavily distorted impression of the data. Mixed-effects models, by contrast, have the ability to take into account data from all trials and all participants even if a given item or participant does not have sufficient variability to calculate traditional measures of metacognitive accuracy.

Finally, especially if a significant portion of variance is due to random item effects (which, as we demonstrate below, are characteristic of these data), the inferential tests accompanying traditional measures of metacognitive accuracy, such as the *t* statistic used to establish that a correlation differs from zero, can be prone to elevated Type-I error rates. This problem is only exacerbated if the same data are analyzed repeatedly, using several measures of statistical relatedness, without any adjustment for multiple testing. Mixed-effects models, by contrast, preserve nominal Type-I error rates, even in the presence of significant item effects (Murayama, Sakaki, Yan, & Smith, 2014).

Model fitting. Model fitting was performed using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in the R statistical computing environment. Model selection proceeded stepwise. Across all the models, memory accuracy (accurate vs. inaccurate) was the response variable. A null model containing only a random intercept for participants, controlling for individual differences in terms of memory ability, was used as a baseline. In the second step, a random intercept for items was added, accounting for the possibility that some items may, on average, have been more memorable than others. In the third step, a random slope for confidence was included to allow for the possibility that the relationship between confidence and accuracy might vary across participants. After fitting the random part of the model, fixed effects were entered stepwise. First, a main effect for trial type was added, followed by a main effect for confidence, an interaction between confidence and trial type, a main effect for list length, and finally a three-way interaction between confidence, trial type, and list length.

Across all experiments, a likelihood ratio test was conducted following each modeling step to assess whether entering the given effect led to a significant improvement in model fit (see Table 1 in the supplementary materials provided on the Open Science Framework [OSF], https://osf.io/9z2gp). The best-fitting model for each experiment was the model including random intercepts for participants and items and random slopes for confidence across participants, as well as a three-way interaction between trial type, confidence, and list length (Experiments 1 and 3) or trial type, confidence, and modality (Experiment 2). Therefore, for the remaining experiments, we report and interpret only this best-fitting model, with the details of model fitting provided in OSF Table 1. Overall, the incremental improvements in model fit suggest that as expected, there was a relationship between memory confidence and memory accuracy in all experiments and the confidence–accuracy relationship was modulated by trial type, list length, and modality of stimulus presentation.

Model interpretation. OSF Table 2 (https://osf.io/9z2gp) shows coefficients from the best fitting model. Studied items were selected as the reference category, and mean centering was applied to the confidence variable. Because the model was a mixed-effects logistic regression, regression coefficients express log odds ratios. Given that the predictors are interaction terms between continuous and categorical variables, regression coefficients are devoid of easily accessible intuitive meaning; therefore, the model interpretation provided below relies on a visual summary of the model as well as model-predicted conditional probabilities. Observed and model-predicted conditional probabilities of making an accurate recognition judgment as a function of trial type, confidence, and number of lists studied are shown in Figure 1, which provides a succinct and comprehensive visual summary of the model. To facilitate interpretation of the plot, we report predicted values for the lowest, mean, and highest levels of confidence across list lengths and trial types.

One-list condition. Figure 1 displays the one-list condition in black. As shown in the top left panel of Figure 1, a strong positive relationship was found between memory accuracy and memory confidence for studied items. The predicted probability of giving an accurate response was 21.0% at the lowest level of self-reported subjective confidence, 86.3% at the mean level of confidence, and 96.5% at the highest level of confidence. Semantically unrelated items are generally easy to detect as not belonging to the list and should have been correctly identified as such. Consistent with that expectation, although there seemed to be a slight positive relationship between confidence and accuracy for unrelated items, participants' performance was close to ceiling regardless of their confidence ratings, except for the lowest levels of confidence (see top right panel of Figure 1). Accordingly, the predicted probability of giving an accurate response was 42.9% at the lowest level of self-reported subjective confidence, 98.9% at the mean level of confidence, and 99.9% at the highest level of confidence. For strong lures, participants performed barely above chance at the lowest level of confidence; however, they were considerably more likely to be accurate at the highest level of confidence (see bottom left panel of Figure 1). The predicted probability of giving an accurate response was 59.4% at the lowest level of self-reported subjective confidence, 70.6% at the mean level of confidence, and 75.1% at the

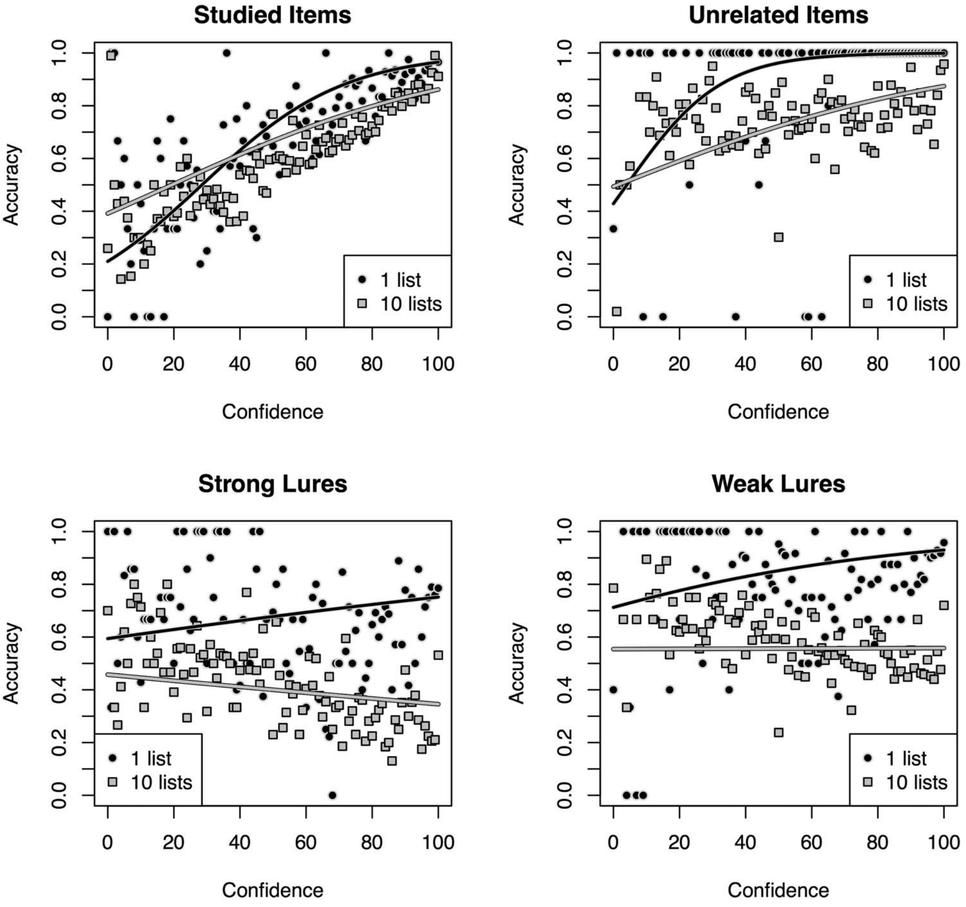

Figure 1. Mixed-effects logistic regression predicting accuracy as a function of a three-way interaction between trial type, confidence, and list length, controlling for item-level and participant-level dependencies (Experiment 1). Even though the model included centered confidence values, for ease of interpretation the *x*-axis shows raw confidence scores (on a scale from 0 to 100) and the *y*-axis shows probability of an accurate response, conditioned on the level of confidence. Each panel shows a different trial type. The lines represent model-predicted conditional probabilities, whereas the dots represent observed conditional probabilities. The black line and black circles correspond to the one-list condition and the gray line and gray squares correspond to the 10-list condition.

highest level of confidence. For weak lures, confidence was a reliable predictor of accuracy (see bottom right panel of Figure 1). Memory performance was well above chance (71.2%), even on trials with low levels of subjective confidence. On trials at the mean levels of subjective confidence, accuracy increased to 88.6% and at the highest levels of subjective confidence, to 93.0%.

10-list condition. In the 10-list condition, displayed in gray in Figure 1, the results were similar to the one-list condition for studied and unrelated items and markedly different for strong and weak lures. Just as in the one-list condition, a strong positive relationship was found between memory accuracy and memory confidence for studied items (see top left panel of Figure 1). The predicted probability of giving an accurate response was 40.5% (i.e., below chance) at the lowest level of self-reported subjective confidence, 75.9% at the mean level of confidence, and 86.6% at the highest level of confidence. For unrelated items, the confidence–accuracy relationship was still positive, but stronger than in the one-list condition (see top right panel of Figure 1). The predicted probability of giving an accurate response was 49.6% at the lowest level of self-reported subjective confidence, 78.7% at the mean level of confidence, and 87.3% at the highest level of confidence. The most likely reason for this difference across conditions is that studying 10 lists placed more cognitive demand on participants than studying one list. Therefore, a higher rate of false alarms was elicited by unrelated items, thus providing more room for an association to emerge between confidence and accuracy. Unlike in the one-list condition, a negative, rather than positive, confidence–accuracy relationship was found for strong lures (see bottom left panel of Figure 1). The predicted probability of giving an accurate response was 42.4% at the lowest level of self-reported subjective confidence, 35.5% at the mean level of confidence, and 32.5% at the highest level of confidence. For weak lures, the positive confidence–accuracy relationship observed in the one-list condition was eliminated and memory accuracy and subjective memory confidence were unrelated. As shown in the bottom right panel of Figure 1, the predicted probability of an accurate response was 55.7% across all levels of confidence.

Discussion

Experiment 1 explored spontaneous accessibility and amount of material to be encoded as possible moderators of the relationship between the objective accuracy of recognition memory judgments and participants' subjective confidence in their memory performance. In the relatively undemanding one-list condition, in which participants studied 15 items and were tested on 30 items, an interaction was observed between confidence and trial type, reflecting varying strengths of the positive relationship between memory confidence and memory accuracy across trial types. For studied items, the relationship was strong, that is, memory accuracy tracked confidence judgments. For weak lures and strong lures, a similar pattern of results emerged; however, the confidence–accuracy relationship was weaker than for studied items. For semantically unrelated items, the accuracy of memory judgments was close to ceiling irrespective of confidence, as expected given their easy discriminability.

This pattern of results proved to be both remarkably stable and surprisingly malleable as the number of lists to be memorized was increased from one to 10. Just as in the one-list condition, a positive confidence–accuracy relationship was observed for studied items and semantically unrelated items. However, the findings for highly accessible nonstudied items stand in sharp contrast with the one-list condition: For strong lures, we obtained a negative confidence–accuracy relationship, and for weak lures, a zero confidence–accuracy relationship. As such, these results remove some interpretational ambiguities from those Roediger–DeSoto studies that found a negative or zero relationship for strong lures and a zero or positive relationship for weak lures, depending on the method of analysis used. Moreover, by virtue of the fact that items were presented in consistent (visual–visual) modalities at encoding and at test, Experiment 1 showed that the negative confidence–accuracy relationship detected by DeSoto and Roediger (2014) and Roediger and DeSoto (2014a) was not (exclusively) due to mixed modalities across encoding and test, but rather due to participants' inability to distinguish chronically high spontaneous accessibility of items from temporarily increased accessibility because of their recent presentation in the context of the experiment.

Experiment 2

Experiment 1 demonstrated that metacognitive performance is tethered to cognitive performance even under the simplest of learning conditions, that is, when presentation modality of the items remains the same across encoding and retrieval and the material to be remembered is a single list of 15 semantically related words. Increasing the number of lists to be encoded from one to 10 led to a decline in the confidence–accuracy relationship, at least for strong and weak lures. In the 10-list condition, we continued to observe a strong positive relationship between subjective confidence and accuracy for studied items; however, the relationship became nonexistent for weak lures, and even markedly negative for strong lures, as the amount of information to be memorized increased. This suggests that, ironically, people's metamnemonic capacities may fail them precisely in those situations when they would need them the most, that is, when their memory becomes error-prone due to heavy cognitive load.

In Experiment 2, we tested an additional moderator of the confidence–accuracy relationship, varying the modality in which the items were presented at encoding and at test. Accordingly, Experiment 2 included three between-participants conditions, with stimuli presented in (a) consistent visual modality across encoding and retrieval (i.e., visual–visual condition), (b) consistent auditory modality across encoding and retrieval (i.e., auditory–auditory condition), and (c) inconsistent auditory and visual modalities across encoding and retrieval (i.e., auditory–visual condition). Just as the increase in list length allows us to build better theories of cognition, as well as to mimic the natural world in which critical and irrelevant information varies in its amount, the introduction of modality was both of theoretical interest and reflected everyday situations in which information, especially of any consequential sort, is rarely merely read, given the auditory nature of much social interaction.

In line with previous work demonstrating detrimental effects to memory as result of auditory presentation (Beauchamp, 2002; Cleary & Greene, 2002; R. E. Smith & Hunt, 1998), mixed modalities across encoding and retrieval (Kellogg, 2001), and both (Gallo et al., 2001; Israel & Schacter, 1997), we expected that memory performance would be superior in the consistent (visual–visual) compared with the mixed (auditory–visual) condition and, possibly, also the auditory–auditory condition. The experiments by Roediger and DeSoto used a mixed-modality design (auditory encoding and visual testing) and thus no evidence is available about the effects of modality match and mismatch on the confidence–accuracy relationship. However, in line with the results of Experiment 1 demonstrating that failures of memory and metamemory tend to co-occur, it is possible that to the extent that memory performance falters, there may be a concomitant decline in metamnemonic performance, that is, the relationship between memory accuracy and confidence could break down as a result of inconsistent modalities at encoding and test.

Method

Participants. 546 volunteers recruited via Socialsci.com were paid $2.00 in exchange for their participation.

Design and procedure. The design of Experiment 2 was similar to that of Experiment 1. However, unlike in Experiment 1, the modality in which stimuli were presented at encoding and at test was varied between participants. 146 participants were pseudorandomly assigned to the visual–visual condition in which stimuli were presented visually both at encoding and at test, 200 participants to the auditory–auditory condition in which stimuli were presented auditorily both at encoding and at test, and 200 participants to the auditory–visual condition in which stimuli were presented auditorily at encoding and visually at test. Unlike in Experiment 1, the number of lists was held constant at five across conditions.

Visual–visual condition. The visual–visual condition was identical to the 10-list condition of Experiment 1, with the sole exception that all participants memorized the same five, rather than 10, thematic lists (body parts, insects, sports, occupations, and weather phenomena).

Auditory–auditory condition. The auditory–auditory condition was identical to the visual–visual condition with the exception that the to-be-memorized items were presented to participants auditorily rather than visually both in the study phase and at test. The word lists were recorded with a text-to-speech converter (http://www.readthewords.com), using a female voice. For each list, the audio file started by stating the theme of the list (e.g., "An occupation or profession"), paused for 4 s, and then proceeded to read the list aloud at a rate of one word per 2 s. The duration of each audio file was 35 s. Items were presented in a randomized order that was held constant across participants. As in Experiment 1, participants studied the lists at their own pace, and were able to pause and rewind the audio file at any time to hear the list again.[3] After encoding the lists of items and completing the distractor task, participants provided recognition judgments for items presented in the au-

[3] Participants in the studies conducted by Roediger and DeSoto (DeSoto & Roediger, 2014; Roediger & DeSoto, 2014a) did not have the option to rewind the audio files. The results of a supplementary study not reported in detail here (see, however, the data file published on OSF, https://osf.io/9z2gp) suggest that this procedural difference probably did not have a major impact on our findings. In the supplementary study, just as in Roediger and DeSoto (2014a), participants studied 10 lists, with auditory presentation at encoding and visual presentation at test. In the data file published on OSF, mean accuracy by trial type was 70% for studied items, 90% for unrelated items, 51% for strong lures, and 71% for weak lures. Roediger and DeSoto (2014a) reported highly similar values, with 70% accuracy for studied items, 88% for unrelated items, 56% for strong lures, and 70% for weak lures.

ditory modality. In line with the self-paced nature of the recognition task in visual–visual condition, participants were allowed to listen to items as many times as they wished before entering their responses.

Auditory–visual condition. The auditory–visual condition was identical to the visual–visual and auditory–auditory conditions with the exception that stimulus presentation modalities were inconsistent across encoding and retrieval, with items presented auditorily in the study phase and visually at test.

Results

Descriptive statistics. Descriptive statistics are reported in Table 1. Overall, the pattern of accuracy and confidence values was quite similar to that found in Experiment 1. With participants studying 5 lists in the present experiment, the mean level of accuracy was between the one-list and 10-list conditions of Experiment 1. Somewhat surprisingly given the preponderance of existing research demonstrating modality effects on memory accuracy (Beauchamp, 2002; Cleary & Greene, 2002; Gallo et al., 2001; Israel & Schacter, 1997; Kellogg, 2001; R. E. Smith & Hunt, 1998), presentation modality did not modulate overall levels of accuracy in the present experiment. However, as in Experiment 1, trial type influenced accuracy. Memory was most accurate for unrelated items, followed by studied items, weak lures, and strong lures. Accuracy levels and confidence judgments did not exhibit the same pattern. Irrespective of presentation modality, the highest confidence ratings were provided for studied items, followed by unrelated items, strong lures, and weak lures.

Model interpretation. The best-fitting model for Experiment 2 included random intercepts for participants and items and random slopes for confidence across participants, as well as a three-way interaction between trial type, confidence, and modality (see OSF Table 1, https://osf.io/9z2gp). Coefficients from the best fitting mixed-effects model are reported in OSF Table 3. Figure 2 provides a succint and comprehensive visual summary of the model by displaying observed and model-predicted conditional probabilities of making an accurate recognition judgment as a function of trial type, confidence, and stimulus presentation modality. To facilitate interpretation of the plot, we report predicted values for the lowest, mean, and highest levels of confidence across stimulus presentation modalities and trial types.

Visual–visual condition. Figure 2 shows the visual–visual condition in black. Just as in both conditions of Experiment 1, a strong positive relationship was found between memory accuracy and memory confidence for studied items (see top left panel of Figure 2). The predicted probability of giving an accurate response was 12.6% (i.e., close to floor) at the lowest level of self-reported subjective confidence, 81.2% at the mean level of confidence, and 94.3% at the highest level of confidence. For unrelated items, the confidence–accuracy relationship exhibited the same moderately positive direction as in Experiment 1 (see top right panel of Figure 2), presumably due to a ceiling effect. The predicted probability of giving an accurate response was 80.6% at the lowest level of self-reported subjective confidence, 93.8% at the mean level of confidence, and 96.1% at the highest level of confidence. For strong lures, similarly to the 10-list condition of Experiment 1, the relationship between memory confidence and memory accuracy was negative (see bottom left panel of Figure 2). The predicted probability of giving an accurate response was 66.0% at the lowest level of self-reported subjective confidence, 51.8% at the mean level of confidence, and 46.0% at the highest level of confidence. For weak lures, in line with the results of the 10-list condition of Experiment 1, a zero relationship was obtained between confidence and accuracy (see bottom right panel of Figure 2). The predicted probability of giving an accurate response was 71.4% at the lowest level of self-reported subjective confidence, 68.9% at the mean level of confidence, and 68.0% at the highest level of confidence. Overall, these results suggest that increasing the amount of information to be memorized from one to five lists is sufficient to observe the magnitude of detriment to metacognitive performance that was found with 10 studied lists in Experiment 1.

Auditory–auditory condition. The auditory–auditory condition, displayed in gray in Figure 2, produced essentially the same results as the visual–visual condition for each of the four trial types. That is, a highly positive confidence–accuracy relationship was observed for studied items, a moderately positive relation-

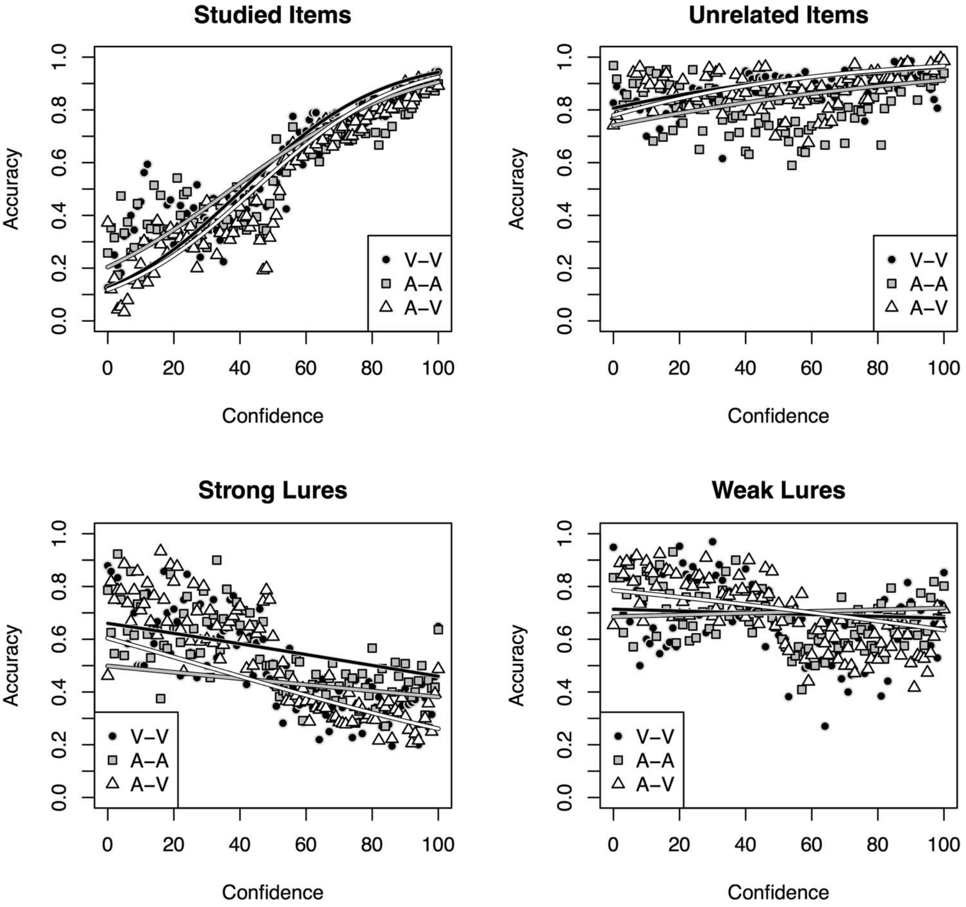

Figure 2. Mixed-effects logistic regression predicting accuracy as a function of a three-way interaction between trial type, confidence, and modality of stimulus presentation, controlling for item-level and participant-level dependencies (Experiment 2). Even though the model included centered confidence values, for ease of interpretation the *x*-axis shows raw confidence scores (on a scale from 0 to 100) and the *y*-axis shows probability of an accurate response, conditioned on the level of confidence. Each panel shows a different trial type. The lines represent model-predicted conditional probabilities, whereas the dots represent observed conditional probabilities. The black line and black circles correspond to the the visual–visual (V–V) condition, the gray line and gray squares correspond to the auditory–auditory (A–A) condition, and the white line and white triangles correspond to the auditory–visual (A–V) condition.

ship for unrelated items, a negative relationship for strong lures, and a zero relationship for weak lures. For studied items, the predicted probability of giving an accurate response was 20.4% (i.e., below chance) at the lowest level of self-reported subjective confidence, 77.5% at the mean level of confidence, and 90.5% at the highest level of confidence (see top left panel of Figure 2). For unrelated items, the predicted probability of giving an accurate response was 74.1% at the lowest level of self-reported subjective confidence, 87.9% at the mean level of confidence, and 91.3% at the highest level of confidence (see top right panel of Figure 2). For strong lures, the predicted probability of giving an accurate response was 49.9% at the lowest level of self-reported subjective confidence, 41.4% at the mean level of confidence, and 38.1% at the highest level of confidence (see bottom left panel of Figure 2). For weak lures, the predicted probability of giving an accurate response was 68.6% at the lowest level of self-reported subjective confidence, 70.9% at the mean level of confidence, and 71.8% at the highest level of confidence (see bottom right panel of Figure 2).

Auditory–visual condition. Compared with both conditions involving consistent stimulus presentation modalities, the auditory–visual condition (displayed in white in Figure 2) yielded highly similar results for studied and unrelated items; however, the confidence–accuracy relationship declined markedly for both strong and weak lures. Just as in the visual–visual and auditory–auditory conditions, a highly positive confidence–accuracy relationship was observed for studied items (top left panel of Figure 2) and a moderately positive relationship for unrelated items (top right panel of Figure 2). For the former, the predicted probability of giving an accurate response was 11.8% (i.e., close to floor) at the lowest level of self-reported subjective confidence, 77.5% at the mean level of confidence, and 92.5% at the highest level of confidence; for the latter, the predicted probability of giving an accurate response was 77.9% at the lowest level of self-reported subjective confidence, 94.1% at the mean level of confidence, and 96.6% at the highest level of confidence. For strong lures, the confidence–accuracy relationship was even more strongly negative than in the consistent modality conditions (see bottom left panel of Figure 2). The predicted probability of giving an accurate response was 60.4% at the lowest level of self-reported subjective confidence, 34.8% at the mean level of confidence, and 26.1% at the highest level of confidence, corresponding to a 34-percentage point accuracy gap between the lowest and highest levels of confidence in the negative direction. Finally, instead of the zero relationship that we observed in the visual–visual and auditory–auditory conditions, a negative confidence–accuracy relationship was found for weak lures (bottom right panel of Figure 2). The predicted probability of giving an accurate response was 78.6% at the lowest level of self-reported subjective confidence, 68.2% at the mean level of confidence, and 63.5% at the highest level of confidence.

Discussion

Experiment 2 tested the effects of the modality of stimulus presentation on the confidence–accuracy relationship. On the basis of extensive previous work demonstrating modality effects (Beauchamp, 2002; Cleary & Greene, 2002; Gallo et al., 2001; Israel & Schacter, 1997; Kellogg, 2001; R. E. Smith & Hunt, 1998), we had predicted that presenting items in mixed modalities across encoding and retrieval, and possibly in consistent auditory modality, would result in worse memory performance and, concomitantly, in a detriment to the confidence–accuracy relationship. Contrary to this expectation, modality of stimulus presentation did not affect memory accuracy. On the other hand, in line with our prediction, we observed an attenuation of metacognitive performance due to mixed modalities at encoding and retrieval, whereas the auditory modality did not seem to be inherently inferior to the visual modality. However, it should be noted that metacognitive performance did not decline equally across trial types. Similarly to Experiment 1, increasing the cognitive demands of the memory task left metacognitive accuracy intact for studied and unrelated items; at the same time, inconsistent modalities across encoding and retrieval resulted in a highly negative confidence–accuracy relationship for strong lures and a negative confidence–accuracy relationship for weak lures. Thus, Experiment 2 further reinforces our conclusion from Experiment 1 according to which Roediger and DeSoto (DeSoto & Roediger, 2014; Roediger & DeSoto, 2014a) may have underestimated participants' metacognitive capacities given the highly demanding nature of the memory tasks that they included in their studies, involving large numbers of lists and inconsistent stimulus presentation modalities. Crucially, this pattern of results also underlines that, at least for deceptive items (i.e., strong and weak lures in the present studies), poor cognitive performance goes hand in hand with poor metacognitive performance.

Experiment 3

Experiments 1 and 2 were conducted using lists of words because presenting stimuli in this manner allows for complete control over participants' learning and ensures high levels of internal validity. However, in experimental research, there is often a tradeoff between internal and external validity. In the context of the present studies, one might raise the objection that in everyday situations, people rarely memorize lists of words outside some meaningful narrative context. Therefore, in Experiment 3 we create a more naturalistic set of materials to be memorized so as to observe whether the results obtained using word lists will replicate under more ecologically realistic conditions. Thus, in the present study, stimuli were presented to participants at encoding as part of short legal vignettes rather than as decontextualized lexical items. On the one hand, a prediction could be made that the switch to narratives will improve memory by creating a causal structure through which items are connected to each other (Black & Bern, 1981). On the other hand, memory performance under these conditions might decline further because of an increase in gist intrusions as a result of making shared category membership of the stimuli more salient (Brainerd & Reyna, 2002). Be that as it may, the aims of Experiment 3 were twofold: first, we sought to replicate the metamnemonic effect of list length obtained in Experiment 1 under ecologically more realistic conditions and second, through a cross-study comparison with Experiment 1, we sought to explore the effects of presenting items as word lists versus embedded in short narratives on the confidence–accuracy relationship.

Method

Participants. 170 volunteers recruited via SocialSci.com were paid $2.00 in exchange for their participation and 70 undergraduate students participated in exchange for course credit. SocialSci.com participants completed the experiment using their own computers, whereas the undergraduate participants completed the experiment at a laboratory. SocialSci.com participants were assigned to the one-list condition and undergraduate participants were assigned to the eight-list condition (see below). The fact that condition and participant pool were confounded may be considered less than ideal; however, because the group with superior memory skills[4], more proneness to demand effects (Sears, 1986), and subject to more tightly controlled experimental conditions (Hamby & Taylor, 2016) was assigned to complete the more demanding task, if anything, this design should result in conservative estimates about the effects of list length on metamnemonic performance.

Design and procedure. The design of Experiment 3 was similar to that of Experiment 1. However, unlike Experiment 1, the to-be-memorized items were embedded in brief legal narratives rather than presented as mere word lists. Participants were assigned to a one-list condition or an eight-list condition.

One-list condition. The materials and procedure used in the one-list condition of Experiment 3 were similar to those used in Experiment 1 but, instead of word lists to memorize, studied items were embedded in a brief legal narrative.[5] For instance, the bird list was presented in the context of a legal narrative on bird keeping and extinction.

> Recent increases in the popularity of bird keeping have led to a dangerous decline in certain bird populations in a local county park. This decline would ultimately lead to extinction of rare species of birds unless action is pursued. The endangered birds due to the increase in bird keeping are *bluebirds*, *hummingbirds*, *seagulls*, *penguins*, *parrots*, *parakeets*, *canaries* and *doves*. Importantly, these are not the only bird populations that are endangered. The populations of more predatory birds that rely on the existence of these captured birds are also decreasing. These predatory birds include *crows*, *sparrows*, *ravens*, *falcons*, *ostriches*, *pigeons*, and *owls*. The local environmental agency has called

[4] Scores on the Scholastic Aptitude Test, used heavily in the undergraduate admissions process, are highly correlated with working memory ability (Daneman & Hannon, 2001). In addition, the decline of episodic memory over the life span is well-documented (for a recent review see Shing et al., 2010) and many activities routinely performed by undergraduates involve memorization and testing, which also suggests that they may be superior at tasks like the one included in the present study compared with general adult populations. Crucially, in a supplementary study not reported in detail here (see, however, the data file published on OSF, https://osf.io/9z2gp), undergraduate participants memorized 10 lists auditorily and were then tested visually. Mean accuracy for studied items was 70%, for unrelated items 90%, for weak lures 71% and for strong lures 51%. In the auditory–visual condition of Experiment 2, where SocialSci.com participants memorized five, rather than 10, lists, mean accuracy for studied items was 71%, for unrelated items 92%, for weak lures 67% and for strong lures 46%. That is, in spite of the fact that the undergraduate participants were asked to memorize twice as many items, their memory performance was slightly better than the memory performance of SocialSci.com participants.

[5] The full set of narrative stimuli is available for download from OSF (https://osf.io/9z2gp).

> on the city council to pass an ordinance prohibiting the capture of birds from any local park by any unauthorized entity. This would also call for the creation of bird capturing permits that may cause unneeded burden on the already depleted city hall public committee resources. However, these actions must be pursued if these bird populations are to be saved.

In the example presented above, the items on which participants were later tested are highlighted for clarity. However, this was not the case in the experiment: Although participants were explicitly instructed to memorize the details of the case, they were not told to focus on specific words. As in prior experiments, participants completed the study phase at their own pace. The distractor task and the testing phase were identical to Experiment 1, with the exception that in addition to providing recognition judgments and confidence ratings, participants also answered four yes or no questions regarding specific details of the narrative. These questions were exploratory in nature and will not be discussed here.

Eight-list condition. The materials and procedure used in the eight-list condition were identical to those used in the one-list condition, but participants studied and were tested on the content from eight different legal narratives. The narratives were presented in an individually randomized order. In addition to providing recognition judgments and confidence ratings, participants also answered 32 yes or no questions regarding specific details of the narratives. These questions were exploratory in nature and will not be discussed here.

Results

Descriptive statistics. Descriptive statistics are reported in Table 1. Participants in the one-list condition were markedly less accurate than in the one-list condition of Experiment 1, which involved memorizing the same number of items presented as lists. This drop in performance affected all trial types but was most pronounced for strong lures for which recognition judgments were at chance in the present experiment. This suggests that embedding the items in a vignette impeded rather than facilitated memory performance, presumably because participants did not know what aspect of the text to focus on, or because the narrative structure might have given rise to more gist intrusions than presenting items as part of a list. Confidence ratings, on the other hand, were comparable to the ones measured in the one-list condition of Experiment 1. On average, participants were most confident for unrelated items, followed by studied items, weak lures, and strong lures.

In the eight-list condition, the accuracy of recognition judgments remained practically unchanged compared with the one-list condition, thus lending credence to our conjecture above that undergraduates' memory performance may be superior compared with the general population. As in the one-list condition, participants were most accurate for unrelated items, followed by studied items and weak lures with comparable values and, finally, strong lures. Confidence ratings, on the other hand, decreased markedly vis-à-vis the one-list condition, especially for unrelated items and weak lures.

Model interpretation. As in Experiment 1, the best-fitting model for Experiment 3 included random intercepts for participants and items and random slopes for confidence across participants, as well as a three-way interaction between trial type, confidence, and list length (see OSF Table 1, https://osf.io/9z2gp). Coefficients from the best fitting mixed-effects model are reported in OSF Table 4. Figure 3 provides a comprehensive and succinct summary of the model by showing observed and model-predicted conditional probabilities of making an accurate recognition judgment as a function of trial type, confidence, and list length. To facilitate interpretation of the plot, as for Experiments 1 and 2, we report predicted values for the lowest, mean, and highest levels of confidence across list lengths and trial types.

One-list condition. Figure 3 shows the one-list condition in black. For studied items, similarly to the consistent results of both previous experiments, a strong positive relationship was found between confidence and accuracy (see top left panel of Figure 3). The predicted probability of giving an accurate response was 17.5% at the lowest level of self-reported subjective confidence, 68.7% at the mean level of confidence, and 86.9% at the highest level of confidence. For unrelated items, in line with the findings of Experiments 1 and 2, no strong confidence–accuracy relationship could emerge due to a ceiling effect (see top right panel of Figure 3). The predicted probability of giving an accurate response was 62.2% at the lowest level of self-

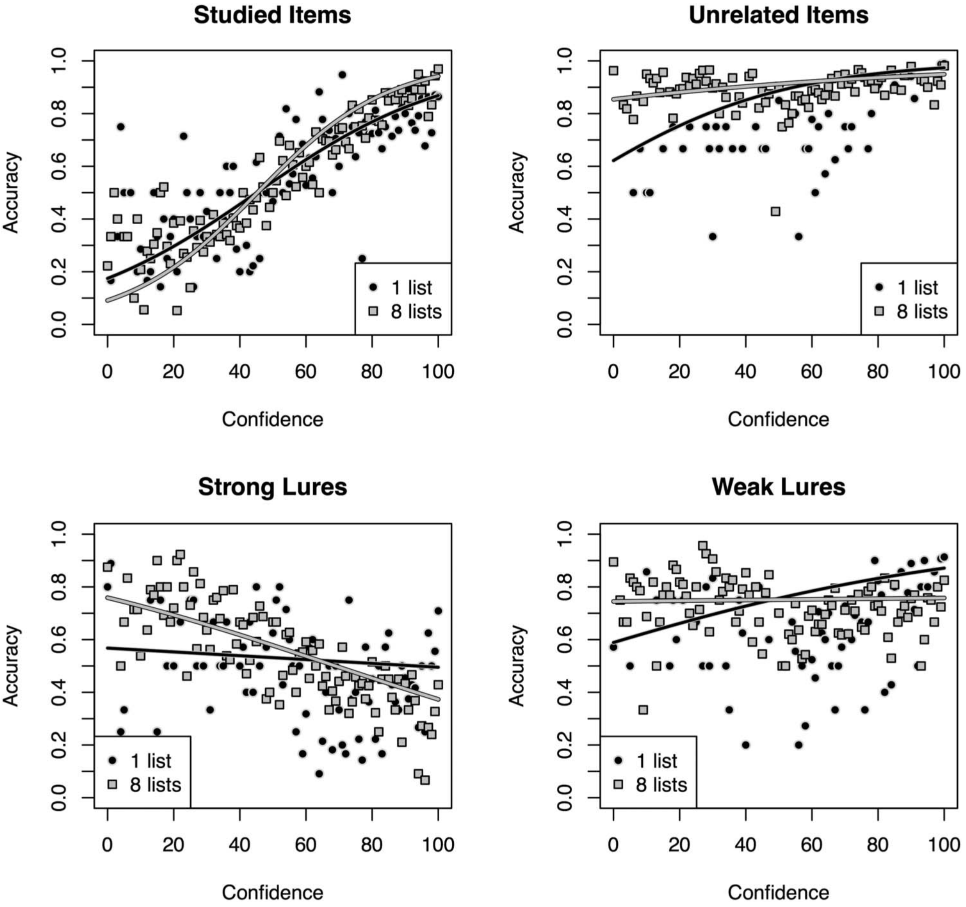

Figure 3. Mixed-effects logistic regression predicting accuracy as a function of a three-way interaction between trial type, confidence, and list length, controlling for item-level and participant-level dependencies (Experiment 3). Even though the model included centered confidence values, for ease of interpretation the *x*-axis shows raw confidence scores (on a scale from 0 to 100) and the *y*-axis shows probability of an accurate response, conditioned on the level of confidence. Each panel shows a different trial type. The lines represent model-predicted conditional probabilities, whereas the dots represent observed conditional probabilities. The black line and black circles correspond to the the one-list condition and the gray line and gray squares correspond to the eight-list condition.

reported subjective confidence, 93.2% at the mean level of confidence, and 97.4% at the highest level of confidence. For strong lures, unlike in Experiment 1, where the same number of items were presented as lists, a negative, rather than positive, confidence–accuracy relationship was obtained (see bottom left panel of Figure 3). The predicted probability of giving an accurate response was 56.8% at the lowest level of self-reported subjective confidence, 51.8% at the mean level of confidence, and 49.5% at the highest level of confidence. For weak lures, the confidence–accuracy relationship was positive, although less strongly so than in Experiment 1 (see bottom right panel of Figure 3). The predicted probability of giving an accurate response was 58.9% at the lowest level of self-reported subjective confidence, 80.5% at the mean level of confidence, and 87.1% at the highest level of confidence.

Eight-list condition. Figure 3 shows the eight-list condition in gray. In line with previous results, a tight alignment between confidence and accuracy was found for studied items (see top left panel of Figure 3). The predicted probability of giving an accurate response was 9.1% at the lowest level of self-reported subjective confidence, 76.0% at the mean level of confidence, and 94.2% at the highest level of confidence. For unrelated items, in spite of high levels of subjective uncertainty, recognition performance remained near ceiling for all levels of confidence (see top right panel of Figure 3). The predicted probability of giving an accurate response was 85.4% at the lowest level of self-reported subjective confidence, 92.8% at the mean level of confidence, and 94.9% at the highest level of confidence. For strong lures, the relationship between confidence and accuracy was markedly more negative than either in the one-list condition of the present experiment or in the 10-list condition of Experiment 1 (see bottom left panel of Figure 3). The predicted probability of giving an accurate response was 75.9% at the lowest level of self-reported subjective confidence, 50.4% at the mean level of confidence, and 37.3% at the highest level of confidence, corresponding to a 39-point accuracy gap in the negative direction. For weak lures, just as in the 10-list condition of Experiment 1, accuracy was essentially unrelated to confidence (see bottom right panel of Figure 3). The predicted probability of giving an accurate response was 74.5% at the lowest level of self-reported subjective confidence, 75.4% at the mean level of confidence, and 75.8% at the highest level of confidence.

Discussion

Unlike in the first two experiments, the study phase of Experiment 3 involved memorizing short legal vignettes rather than mere lists of words, mimicking more closely the nature of information as encountered in the social world. In the one-list condition, a negative confidence–accuracy relationship was observed for strong lures, even though in the equivalent condition of Experiment 1, where items were presented as lists at encoding, the same relationship was positive. Moreover, as the amount of material to be memorized was increased in the eight-list condition, the confidence–accuracy relationship for strong lures became even more markedly negative, and for weak lures, the formerly positive relationship broke down. At the same time, in line with both previous experiments, no detriment to metacognitive performance was found for studied or semantically unrelated items.

The results of Experiment 3 may generalize to more ecologically valid settings in which people are rarely exposed to lists of semantically related words without any further context but are often required to read short vignettes like the ones encountered in this study. However, it should be noted that similarly to both previous experiments, participants still received an instruction to memorize the content of the vignette (although not the items themselves), suggesting that the results reported here may represent the upper bounds of metacognitive performance in similar real-world situations, where learning is often incidental. Moreover, as a further potential limitation of this experiment, participant pools differed across the one-list and eight-list conditions, and this difference may have created a possible confound. We believe, however, that because the more cognitively demanding task was completed under more tightly controlled conditions (Hamby & Taylor, 2016) by college participants likely to possess superior memory skills (Daneman & Hannon, 2001; Shing et al., 2010) and more prone to experimenter demand (Sears, 1986), this study, if anything, underestimated the detrimental effects of large amounts of to-be-encoded material on metacognitive performance under more ecologically realistic conditions. Nonetheless, to conclusively eliminate this confound, future work should replicate the study with participant pools (and thus all demographic characteristics, including age) held constant across list length conditions.

General Discussion

In the three experiments reported above, we explored four possible moderators of the relationship between objective accuracy on a recognition memory task and participants' subjective judgments of memory strength: spontaneous accessibility, amount of material to be memorized, presentation modality at encoding and at test, and information complexity (i.e., presenting items embedded in narratives vs. as decontextualized word lists).

Consistently across the three experiments, participants' recognition memory was most accurate for items unrelated to the presented semantic categories, followed by studied items, weak lures (i.e., nonstudied items with relatively low spontaneous accessibility), and strong lures (i.e., nonstudied items with high spontaneous accessibility). In fact, accuracy never went below 74% for unrelated items and 69% for studied items, whereas performance was barely above chance (55%) for weak lures in one condition and for strong lures in several experiments, dropping as low as 38% in the 10-list condition of Experiment 1. For studied items, subjective confidence tracked accuracy across all experiments: The higher participants' confidence judgments, the more likely recognition memory was to be correct. For unrelated items, recognition memory tended to be highly accurate, presumably because detecting that the semantic category to which the item belonged was not presented at encoding was sufficient for a correct rejection to occur. Due to the excellent overall recognition memory performance (i.e., the restricted range of accuracy), confidence and accuracy were barely associated with each other for unrelated items. As discussed below, the confidence–accuracy relationship for weak and strong lures varied quite substantially across the experiments and conditions as a function of additional moderator variables, ranging from positive to highly negative; however, within each study, participants' metacognitive performance was consistently better for studied items than for weak lures and for weak lures than for strong lures.

In combination with previous work using the DRM paradigm (Deese, 1959; Gallo, 2010; Roediger & McDermott, 1995) and recent studies conducted by Roediger and DeSoto using the same paradigm as the present work (DeSoto & Roediger, 2014; Roediger & DeSoto, 2014a), these findings militate strongly against the view that list learning involves reproductive memory (Leippe, 1980). Rather, it seems that similar reconstructive processes might be at work in the case of mere lists of lexical items, on the one hand, and richer and ecologically more valid scenarios customarily used in eyewitness memory research, on the other hand. Moreover, the findings of this project also call into question direct-access theories of metamnemonic judgments (Hart, 1965; Schwartz, 1994), according to which humans have privileged metacognitive access to the contents of their minds on the basis of which they can provide accurate verbal reports on the strength of memory traces. If that view were warranted, participants should have exhibited excellent metamnemonic performance across the board or, at least, the confidence–accuracy relationship should not have been systematically modulated by trial type. This was evidently not the case.

As such, our results corroborate existing work showing that metamnemonic judgments are subserved by inferential processes (Benjamin, Bjork, & Schwartz, 1998; Brewer et al., 2005; W. F. Brewer & Sampaio, 2006, 2012; Jacoby & Whitehouse, 1989; Kelley & Lindsay, 1993; Koriat, 1995, 2012; Rhodes & Castel, 2008; Stone, Luminet, & Hirst, 2013). More specifically, they provide evidence for Koriat's (2008) consensuality principle, according to which ratings of subjective memory confidence track the consensuality rather than the accuracy of the response. In line with this view, a strong positive relationship emerged between subjective confidence and objective accuracy for consensually correct studied items, whereas the relationship was almost always zero or negative for consensually wrong strong lures. According to Koriat (1997), the inferences subserving judgments of subjective confidence are based on three main kinds of information: intrinsic cues related to the stimulus, including associative relatedness or imagery value, extrinsic cues related to stimulus presentation or processing, including the number of items studied or levels of processing, and internal mnemonic cues, that is, the phenomenological experience accompanying information processing. Our findings demonstrate that the intrinsic, stimulus-driven, cues upon which participants base confidence ratings include spontaneous accessibility, that is, the frequency with which the given item is spontaneously reported as a member of its category. In other words, when nonpresented but highly accessible members of a category are encountered in a recognition task, the strong phenomenological experience of familiarity may be misattributed to prior exposure within the experiment.

In addition to spontaneous accessibility, some further potential moderators of the confidence–accuracy relationship were also investigated. Based on extensive previous work (Gillund & Shiffrin, 1984; Gronlund & Elam, 1994; Murnane & Shiffrin, 1991; Ratcliff et al., 1990;

Shiffrin et al., 1995; Strong, 1912), we had predicted that with increasing amounts of information to be encoded, participants' memory performance would falter, especially for lures (Roediger & McDermott, 1995), and concomitantly, the relationship between accuracy and confidence would be impaired. In line with our expectations, memory accuracy decreased markedly from the one-list condition to the 10-list condition in Experiment 1. This decrease in memory accuracy was accompanied by a decline in metamnemonic performance for deceptive items (i.e., strong and weak lures). However, for nondeceptive items (i.e., studied and unrelated words), the positive confidence–accuracy relationship remained impervious to the amount of information to be encoded.

Furthermore, our experiments explored the moderating effects of varying stimulus presentation modality at encoding and retrieval. Prior work has yielded somewhat equivocal findings regarding the effects of modality on memory accuracy. Visual modality at encoding might produce superior memory (Beauchamp, 2002; Cleary & Greene, 2002; R. E. Smith & Hunt, 1998), better memory might be associated with matching modalities across encoding and retrieval (Kellogg, 2001), or both (Gallo et al., 2001; Israel & Schacter, 1997). Interestingly, in these studies, we did not observe strong effects of switching from consistent (visual–visual or auditory–auditory) to inconsistent (auditory–visual) modalities across encoding and retrieval on memory accuracy. The effects of modality on the confidence–accuracy relationship were considerably more noticeable, with higher levels of confidence associated with *lower* probabilities of an accurate response for strong and weak lures. However, similarly to Experiment 1, the metacognitive performance for studied and unrelated items remained unaffected by the increased cognitive load. Moreover, Experiment 2 offered strong evidence that the detrimental effects on metacognitive accuracy were due to mixed modalities rather than auditory presentation at encoding, considering that the confidence–accuracy relationship was virtually identical across visual–visual and auditory–auditory presentations. This pattern of results lends further credence to inferential theories of metamemory given that conditions usually associated with poor memory performance led to poor metamnemonic performance. Finally, these results suggest that given the design of their studies, Roediger and DeSoto (DeSoto & Roediger, 2014; Roediger & DeSoto, 2014a) may have offered an overly pessimistic assessment of the confidence–accuracy relationship for deceptive items given that their participants memorized large numbers of items, and modalities were different across encoding and retrieval.

In Experiment 3, we probed the effects of embedding the items in short legal narratives to improve the external validity of our findings (Banaji & Crowder, 1989, 1991). When participants were tested on one vignette, metamnemonic performance for strong lures declined considerably compared with Experiment 1, in which stimuli were presented as mere lists. Moreover, when the number of items to be memorized was increased eightfold, we obtained a highly negative confidence–accuracy relationship for strong lures, on par with the results observed for five lists and inconsistent modalities across encoding and retrieval in Experiment 2. Thus, our findings seem to suggest that similar processes might be in place in more ecologically valid settings where lexical items are usually embedded in coherent texts rather than encountered as lists of words devoid of any coherence or causal structure.

Overall, the theoretical implications of these findings are unequivocal. Even supposedly simple list-learning scenarios involve reconstructive memory (DeSoto & Roediger, 2014; Roediger & DeSoto, 2014a; Roediger & McDermott, 1995). Moreover, metacognitive processes are inferential (Koriat, 1997, 2012): When the cues used for judging memory strength are imperfectly correlated with accuracy (e.g., when participants mistake spontaneous accessibility for accessibility as a result of prior exposure), the usually positive relationship between subjective confidence and accuracy can break down. This problem is exacerbated even further if the conditions of encoding or retrieval are suboptimal, including when large amounts of information are to be retained, presentation modalities differ across learning and testing, or the stimuli are embedded in narratives that focus attention on the shared category of the items, thus leading to gist intrusions.

In terms of practical implications, these findings suggest that episodic memory may be subject to something akin to the Dunning–Krueger effect (Dunning et al., 2003; Kruger & Dunning, 1999). That is, ironically, metacognitive monitoring might break down precisely in those situations when one would need it most because one's memory performance is wanting. This has important consequences for lay conceptions of memory, eyewitness testimony, and educational settings. According to naïve theories of memory, memory confidence tracks memory accuracy, that is, highly confident memories are more likely to be accurate (Kassin, 1985; Kassin, Rigby, & Castillo, 1991; Lindsay et al., 1981; Luus & Wells, 1994). In line with findings from the eyewitness memory literature (Bothwell et al., 1987; Sporer et al., 1995), our results confirm that this need not necessarily be the case. To give a specific example, to the extent that Black individuals are highly spontaneously accessible to a perceiver as members of the category "criminal" (Eberhardt, Goff, Purdie, & Davies, 2004), that perceiver might be more likely to falsely recognize a Black foil to have been present at the scene of a crime. Moreover, the person might make the recognition judgment with a high level of subjective confidence, mistaking chronically high spontaneous accessibility of Black individuals for an episodic memory of a specific encounter. In educational settings, students' self-assessment of their memory performance might not be in line with their actual performance (Dunning et al., 2003). Such failures may be especially likely to occur when students need to retain large amounts of information, when stimulus modalities change from encoding to retrieval, and when the to-be-remembered information is encountered under conditions that encourage gist-based processing.

Finally, our hope is that this project will encourage researchers in the field of metacognition to use mixed-effects models in their own work rather than descriptive indices of association such as Pearson's r or Goodman–Kruskal γ (Goodman & Kruskal, 1954). As demonstrated by the analyses presented above, mixed-effects modeling is superior to correlation measures in a number of ways. Unlike with correlations, researchers can perform stepwise model fitting, conduct inferential tests of individual parameters, and make predictions about the value of the response variable given a certain configuration of the independent variables. Crucially, mixed-effects models avoid the specification problems and elevated false positive rates inherent in between-participants, between-event, and within-participant correlations in that they allow for the inclusion of multiple random effects explicitly accounting for different kinds of (participant-level and item-level) dependencies in the data without any loss of information.

Summary

Three experiments were conducted to investigate moderators of the relationship between subjective confidence and the accuracy of recognition memory in several variations of a list-learning task. Subjective confidence was a consistently excellent predictor of memory accuracy for studied items; however, for nonstudied items representing good instances of their respective semantic categories (such as "dog" for animals or "head" for body parts), participants mistakenly relied on high levels of chronic accessibility as a cue for memory strength, especially when they had to encode large amounts of information, stimulus presentation modalities differed across encoding and retrieval, and stimuli were embedded in narratives encouraging gist-based processing. These findings are irreconcilable with the view that humans have privileged access to the strength of memory traces and provide support for theories positing that judgments of subjective confidence are the output of, sometimes error-prone, inferential processes. Moreover, our results are in stark contradiction with lay theories of memory according to which high levels of confidence always indicate high levels of memory accuracy.

References

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

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,* 390–412. http://dx.doi.org/10.1016/j.jml.2007.12.005

Banaji, M. R., & Crowder, R. G. (1989). The bankruptcy of everyday memory. *American Psychologist, 44,* 1185–1193. http://dx.doi.org/10.1037/0003-066X.44.9.1185

Banaji, M. R., & Crowder, R. G. (1991). Some everyday thoughts on ecologically valid methods.

American Psychologist, 46, 78–79. http://dx.doi.org/10.1037/0003-066X.46.1.78

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using 'Eigen' and S4 [Software]. Available from http://CRAN.R-project.org/package=lme4

Beauchamp, H. M. (2002). Aural, visual, and pictorial stimulus formats in false recall. *Psychological Reports, 91,* 941–951. http://dx.doi.org/10.2466/pr0.2002.91.3.941

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55–68. http://dx.doi.org/10.1037/0096-3445.127.1.55

Black, J. B., & Bern, H. (1981). Causal coherence and memory for events in narratives. *Journal of Verbal Learning & Verbal Behavior, 20,* 267–275. http://dx.doi.org/10.1016/S0022-5371(81)90417-5

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology, 72,* 691–695. http://dx.doi.org/10.1037/0021-9010.72.4.691

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,* 14325–14329. http://dx.doi.org/10.1073/pnas.0803390105

Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science, 11,* 164–169. http://dx.doi.org/10.1111/1467-8721.00192

Brédart, S., & Devue, C. (2006). The accuracy of memory for faces of personally known individuals. *Perception, 35,* 101–106. http://dx.doi.org/10.1068/p5382

Brewer, N., & Wells, G. L. (2006). The confidence–accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12,* 11–30. http://dx.doi.org/10.1037/1076-898X.12.1.11

Brewer, W. F., & Sampaio, C. (2006). Processes leading to confidence and accuracy in sentence recognition: A metamemory approach. *Memory, 14,* 540–552. http://dx.doi.org/10.1080/09658210600590302

Brewer, W. F., & Sampaio, C. (2012). The metamemory approach to confidence: A test using semantic memory. *Journal of Memory and Language, 67,* 59–77. http://dx.doi.org/10.1016/j.jml.2012.04.002

Brewer, W. F., Sampaio, C., & Barlow, M. R. (2005). Confidence and accuracy in the recall of deceptive and nondeceptive sentences. *Journal of Memory and Language, 52,* 618–627. http://dx.doi.org/10.1016/j.jml.2005.01.017

Cleary, A. M., & Greene, R. L. (2002). Paradoxical effects of presentation modality on false memory. *Memory, 10,* 55–61. http://dx.doi.org/10.1080/09658210143000236

Cutler, B. L., & Penrod, S. D. (1989). Forensically relevant moderators of the relation between eyewitness identification accuracy and confidence. *Journal of Applied Psychology, 74,* 650–652. http://dx.doi.org/10.1037/0021-9010.74.4.650

Daneman, M., & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology: General, 130,* 208–223. http://dx.doi.org/10.1037/0096-3445.130.2.208

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58,* 17–22. http://dx.doi.org/10.1037/h0046671

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior, 4,* 243–260. http://dx.doi.org/10.1007/BF01040617

DeSoto, K. A., & Roediger, H. L., III. (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. *Psychological Science, 25,* 781–788. http://dx.doi.org/10.1177/0956797613516149

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12,* 83–87. http://dx.doi.org/10.1111/1467-8721.01235

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology, 87,* 876–893. http://dx.doi.org/10.1037/0022-3514.87.6.876

Furman, O., Dorfman, N., Hasson, U., Davachi, L., & Dudai, Y. (2007). They saw a movie: Long-term memory for an extended audiovisual narrative. *Learning & Memory, 14,* 457–467. http://dx.doi.org/10.1101/lm.550407

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition, 38,* 833–848. http://dx.doi.org/10.3758/MC.38.7.833

Gallo, D. A., McDermott, K. B., Percer, J. M., & Roediger, H. L., III. (2001). Modality effects in false recall and false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 339–353. http://dx.doi.org/10.1037/0278-7393.27.2.339

Ge, L., Luo, J., Nishimura, M., & Lee, K. (2003). The lasting impression of chairman Mao: Hyperfidelity of familiar-face memory. *Perception, 32,* 601–614. http://dx.doi.org/10.1068/p5022

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67. http://dx.doi.org/10.1037/0033-295X.91.1.1

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49,* 732–764. http://dx.doi.org/10.1080/01621459.1954.10501231

Gronlund, S. D., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1355–1369. http://dx.doi.org/10.1037/0278-7393.20.6.1355

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92,* 160–170. http://dx.doi.org/10.1037/0022-0663.92.1.160

Hamby, T., & Taylor, W. (2016). Survey satisficing inflates reliability and validity measures: An experimental comparison of college and Amazon Mechanical Turk samples. *Educational and Psychological Measurement, 76,* 912–932. http://dx.doi.org/10.1177/0013164415627349

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56,* 208–216. http://dx.doi.org/10.1037/h0022263

Hirst, W., & Echterhoff, G. (2012). Remembering in conversations: The social sharing and reshaping of memories. *Annual Review of Psychology, 63,* 55–79. http://dx.doi.org/10.1146/annurev-psych-120710-100340

Israel, L., & Schacter, D. L. (1997). Pictorial encoding reduces false recognition of semantic associates. *Psychonomic Bulletin & Review, 4,* 577–581. http://dx.doi.org/10.3758/BF03214352

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General, 118,* 126–135. http://dx.doi.org/10.1037/0096-3445.118.2.126

Kassin, S. M. (1985). Eyewitness identification: Retrospective self-awareness and the accuracy–confidence correlation. *Journal of Personality and Social Psychology, 49,* 878–893. http://dx.doi.org/10.1037/0022-3514.49.4.878

Kassin, S. M., Ellsworth, P. C., & Smith, V. L. (1989). The "general acceptance" of psychological research on eyewitness testimony: A survey of the experts. *American Psychologist, 44,* 1089–1098. http://dx.doi.org/10.1037/0003-066X.44.8.1089

Kassin, S. M., Rigby, S., & Castillo, S. R. (1991). The accuracy–confidence correlation in eyewitness testimony: Limits and extensions of the retrospective self-awareness effect. *Journal of Personality and Social Psychology, 61,* 698–707. http://dx.doi.org/10.1037/0022-3514.61.5.698

Kassin, S. M., Tubb, V. A., Hosch, H. M., & Memon, A. (2001). On the "general acceptance" of eyewitness testimony research. A new survey of the experts. *American Psychologist, 56,* 405–416. http://dx.doi.org/10.1037/0003-066X.56.5.405

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,* 1–24. http://dx.doi.org/10.1006/jmla.1993.1001

Kellogg, R. T. (2001). Presentation modality and mode of recall in verbal false memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 913–919. http://dx.doi.org/10.1037/0278-7393.27.4.913

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124,* 311–333. http://dx.doi.org/10.1037/0096-3445.124.3.311

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

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 945–959. http://dx.doi.org/10.1037/0278-7393.34.4.945

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119,* 80–113. http://dx.doi.org/10.1037/a0025648

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121–1134. http://dx.doi.org/10.1037/0022-3514.77.6.1121

Landauer, T. K. (1986). How much do people remember? Some estimates of the quantity of learned information in long-term memory. *Cognitive Science, 10,* 477–493. http://dx.doi.org/10.1207/s15516709cog1004_4

Leippe, M. R. (1980). Effects of integrative memorial and cognitive processes on the correspondence of eyewitness accuracy and confidence. *Law and Human Behavior, 4,* 261–274. http://dx.doi.org/10.1007/BF01040618

Lindsay, R. C. L., Wells, G. L., & Rumpel, C. M. (1981). Can people detect eyewitness-identification accuracy within and across situations? *Journal* *of Applied Psychology, 66,* 79–89. http://dx.doi.org/10.1037/0021-9010.66.1.79

Luus, C. A. E., & Wells, G. L. (1994). The malleability of eyewitness confidence: Co-witness and perseverance effects. *Journal of Applied Psychology, 79,* 714–723. http://dx.doi.org/10.1037/0021-9010.79.5.714

Mandler, G., & Boeck, W. J. (1974). Retrieval processes in recognition. *Memory & Cognition, 2,* 613–615. http://dx.doi.org/10.3758/BF03198129

Maylor, E. A., & Mo, A. (1999). Effects of study–test modality on false recognition. *British Journal of Psychology, 90,* 477–493. http://dx.doi.org/10.1348/000712699161567

McKelvie, S. J. (1999). Effect of retrieval instructions on false recall. *Perceptual and Motor Skills, 88,* 876–878. http://dx.doi.org/10.2466/pms.1999.88.3.876

McKelvie, S. J. (2001). Effects of free and forced retrieval instructions on false recall and recognition. *Journal of General Psychology, 128,* 261–278. http://dx.doi.org/10.1080/00221300109598912

Meade, M. L., & Roediger, H. L., III. (2006). The effect of forced recall on illusory recollection in younger and older adults. *The American Journal of Psychology, 119,* 433–462. http://dx.doi.org/10.2307/20445352

Meade, M. L., & Roediger, H. L., III. (2009). Age differences in collaborative memory: The role of retrieval manipulations. *Memory & Cognition, 37,* 962–975. http://dx.doi.org/10.3758/MC.37.7.962

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General, 140,* 239–257. http://dx.doi.org/10.1037/a0023007

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review, 14,* 858–865. http://dx.doi.org/10.3758/BF03194112

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1287–1306. http://dx.doi.org/10.1037/a0036914

Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 855–874. http://dx.doi.org/10.1037/0278-7393.17.5.855

Payne, D. G., Elie, C. J., Blackwell, J. M., & Neuschatz, J. S. (1996). Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Memory and Language, 35,* 261–285. http://dx.doi.org/10.1006/jmla.1996.0015

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law, 1,* 817–845. http://dx.doi.org/10.1037/1076-8971.1.4.817

Pirmoradi, M., & McKelvie, S. (2014). Feedback, confidence, and false recall in the DRMRS Procedure. *Current Psychology, 34,* 248–267. http://dx.doi.org/10.1007/s12144-014-9255-0

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 163–178. http://dx.doi.org/10.1037/0278-7393.16.2.163

Read, J. D. (1996). From a passing thought to a false memory in 2 minutes: Confusing real and illusory events. *Psychonomic Bulletin & Review, 3,* 105–111. http://dx.doi.org/10.3758/BF03210749

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

Roediger, H. L., III, & Crowder, R. G. (1976). A serial position effect in recall of United States presidents. *Bulletin of the Psychonomic Society, 8,* 275–278. http://dx.doi.org/10.3758/BF03335138

Roediger, H. L., III, & DeSoto, K. A. (2014a). Confidence and memory: Assessing positive and negative correlations. *Memory, 22,* 76–91. http://dx.doi.org/10.1080/09658211.2013.795974

Roediger, H. L., III, & DeSoto, K. A. (2014b). Forgetting the presidents. *Science, 346,* 1106–1109. http://dx.doi.org/10.1126/science.1259627

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814. http://dx.doi.org/10.1037/0278-7393.21.4.803

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin & Review, 1,* 357–375. http://dx.doi.org/10.3758/BF03213977

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51,* 515–530. http://dx.doi.org/10.1037/0022-3514.51.3.515

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 267–287. http://dx.doi.org/10.1037/0278-7393.21.2.267

Shing, Y. L., Werkle-Bergner, M., Brehmer, Y., Müller, V., Li, S.-C., & Lindenberger, U. (2010). Episodic memory across the lifespan: The contributions of associative and strategic components. *Neuroscience and Biobehavioral Reviews, 34,* 1080–1091. http://dx.doi.org/10.1016/j.neubiorev.2009.11.002

Smith, R. E., & Hunt, R. R. (1998). Presentation modality affects false memory. *Psychonomic Bulletin & Review, 5,* 710–715. http://dx.doi.org/10.3758/BF03208850

Smith, S. M., Ward, T. B., Tindell, D. R., Sifonis, C. M., & Wilkenfeld, M. J. (2000). Category structure and created memories. *Memory & Cognition, 28,* 386–395. http://dx.doi.org/10.3758/BF03198554

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence–accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118,* 315–327. http://dx.doi.org/10.1037/0033-2909.118.3.315

Standing, L. (1973). Learning 10,000 pictures. *The Quarterly Journal of Experimental Psychology, 25,* 207–222. http://dx.doi.org/10.1080/14640747308400340

Stone, C. B., Luminet, O., & Hirst, W. (2013). Induced forgetting and reduced confidence in our personal past? The consequences of selectively retrieving emotional autobiographical memories. *Acta Psychologica, 144,* 250–257. http://dx.doi.org/10.1016/j.actpsy.2013.06.019

Strong, E. K. J. (1912). The effect of length of series upon recognition memory. *Psychological Review, 19,* 447–462. http://dx.doi.org/10.1037/h0069812

Tulving, E., & Craik, F. I. (2000). *The Oxford handbook of memory*. Oxford, UK: Oxford University Press.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352–373. http://dx.doi.org/10.1037/h0020071

Wells, G. L., & Bradfield, A. L. (1998). “Good, you identified the suspect:” Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83,* 360–376. http://dx.doi.org/10.1037/0021-9010.83.3.360

Wells, G. L., Ferguson, T. J., & Lindsay, R. C. L. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology, 66,* 688–696. http://dx.doi.org/10.1037/0021-9010.66.6.688

Wells, G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 155–170). Cambridge, UK: Cambridge University Press.

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology, 54,* 277–295. http://dx.doi.org/10.1146/annurev.psych.54.101601.145028

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology, 55,* 235–269. http://dx.doi.org/10.1146/annurev.psych.55.090902.141555

Received August 1, 2016
Revision received December 20, 2016
Accepted January 29, 2017 ■