Memory
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pmem20

# Confidence and memory: Assessing positive and negative correlations

Henry L. Roediger III [a] & K. Andrew DeSoto [a]
[a] Department of Psychology , Washington University in St. Louis , MO , USA
Published online: 30 May 2013.

To cite this article: Henry L. Roediger III & K. Andrew DeSoto , Memory (2013): Confidence and memory: Assessing positive and negative correlations, Memory, DOI: 10.1080/09658211.2013.795974

To link to this article: http://dx.doi.org/10.1080/09658211.2013.795974

*Memory,* 2013
http://dx.doi.org/10.1080/09658211.2013.795974



# Confidence and memory: Assessing positive and negative correlations

**Henry L. Roediger III and K. Andrew DeSoto**

Department of Psychology, Washington University in St. Louis, MO, USA



The capacity to learn and remember surely evolved to help animals solve problems in their quest to reproduce and survive. In humans we assume that metacognitive processes also evolved, so that we know when to trust what we remember (i.e., when we have high confidence in our memories) and when not to (when we have low confidence). However this latter feature has been questioned by researchers, with some finding a high correlation between confidence and accuracy in reports from memory and others finding little to no correlation. In two experiments we report a recognition memory paradigm that, using the same materials (categorised lists), permits the study of positive correlations, zero correlations, and negative correlations between confidence and accuracy within the same procedure. We had subjects study words from semantic categories with the five items most frequently produced in norms omitted from the list; later, subjects were given an old/new recognition test and made confidence ratings on their judgements. Although the correlation between confidence and accuracy for studied items was generally positive, the correlation for the five omitted items was negative in some methods of analysis. We pinpoint the similarity between lures and targets as creating inversions between confidence and accuracy in memory. We argue that, while confidence is generally a useful indicant of accuracy in reports from memory, in certain environmental circumstances even adaptive processes can foster illusions of memory. Thus understanding memory illusions is similar to understanding perceptual illusions: Processes that are usually adaptive can go awry under certain circumstances.

***Keywords:*** Confidence; Memory accuracy; Recognition memory; Categorised lists; Metacognition.

Cognitive processes evolved to keep animals in accurate contact with their environments, both natural and social. Sensing, perceiving, remembering, comprehending, and thinking surely led to adaptive fitness. Yet psychologists are continually confronted and confounded with the challenging puzzles presented by cognitive illusions that arise for each sort of cognitive process. Psychologists have studied illusions of one type or another for over 150 years, but they have been known for millennia; for instance, Aristotle described the moon illusion and an illusion of touch that still bears his name (Benedetti, 1985). Perceptual illusions were studied relentlessly beginning in the mid-1800s (see Coren & Girgus, 1978), and in the past 50 years cognitive and social psychologists have discovered many instances of illusions in thinking and remembering (see Pohl's 2004 edited collection for review chapters on many types of cognitive illusions).

The ubiquity of illusions leads naturally to the question: How can systems that evolved to keep us in touch with our environment (perceiving), to keep track of the past happenings within the

Address correspondence to: Henry L. Roediger III, Department of Psychology, Box 1125, Washington University, One Brookings Drive, St. Louis, MO 63130, USA. E-mail: roediger@wustl.edu

We thank Emily Rosenzweig for her assistance collecting data, Chris Wahlheim for his help creating and implementing the slider, and John Wixted and members of the Memory Lab and the Washington University Behavior, Brain, & Cognition program for their comments.

environment (remembering), and to reason with the information perceived and retained (thinking), be so prone to error? After all, Darwin (1872/1958, p. 447) wrote that "Every complex structure and instinct" should be "useful to the possessor". Natural selection could "never produce in a being anything injurious to itself, for natural selection acts solely by and for the good of each". How can powerful cognitive illusions be reconciled with this statement?

We explore the general form the answer to such a question might take near the end of this paper, but we spend the remainder of the introduction discussing one particular problem that has vexed psychologists for years. The issue concerns the relation between confidence and accuracy in reports from memory. The common assumption (among laypeople, most psychologists, and even the US Supreme Court in Neil v. Biggers, 1972) is that the correlation between confidence and accuracy is relatively strong: When we remember an event or know a fact we can judge our certainty or confidence with reasonable accuracy. We think we know what we know, and we know what we do not know. However, often this assumption is unwarranted. For example, people have been convicted on the basis of highly confident eyewitness testimony but exonerated later when DNA evidence shows they were not the perpetrators of the crime (Garrett, 2012).

Psychologists have issued strikingly different pronouncements about the issue of confidence and accuracy in memory reports. In their book on metacognition Dunlosky and Metcalfe (2009) wrote, "The relative accuracy of people's confidence is high. Higher confidence ratings almost inevitably mean that the item had been previously presented" (p. 176). The research on which this statement was based came largely from standard laboratory paradigms in which students study lists of words or pictures and have their memories tested, usually in a recognition procedure, with a confidence rating given for each judgement. On the other hand, researchers in a different tradition of research—studying recognition of faces in eyewitness memory situations—often have reached a very different conclusion from their work. In a survey of evidence in 1989, Smith, Kassin, and Ellsworth concluded "confidence is neither a useful predictor of the accuracy of a particular witness nor of the accuracy of particular statements made by the same witness" (p. 358). More recently, Odinot, Wolters, and Van Koppen (2009) reiterated this claim and added that the relation between confidence and accuracy is so weak that confidence ratings "should never be allowed as evidence for memory accuracy in the courtroom" (p. 513). (We should note that some eyewitness researchers have revised their opinion that confidence and accuracy are not related when a memory test is given shortly after an event; see Brewer & Wells, 2006.)

From the snippets provided in the previous paragraph one might be tempted to conclude that such wildly different pronouncements have arisen because of differences between the types of materials involved (word lists vs faces) or perhaps procedural differences in the tests given. We do not believe this conclusion is warranted, as we will show in the experiments reported here. Roediger, Wixted, and DeSoto (2012) reviewed the complicated evidence about confidence in reports from memory and concluded that one overriding problem is that a researcher can investigate the confidence–accuracy relation using at least five different analytic techniques and that the conclusions from the various techniques need not agree. For example, one can ask whether people who are generally confident are also generally accurate in their performance (a between-person or between-subjects analysis). On the other hand, one can ask the question: Are events that are generally well remembered also more confidently remembered (a between-events analysis)? The answer to the second question could be *yes* at the same time the answer to the first question is *no*, and vice versa. The same holds for the other techniques.

Given our belief that it is not the type of material that causes differences in confidence and accuracy across studies, the aim of the present research was to develop a technique that would permit us to study various types of confidence–accuracy relationships with the same subjects and the same general materials. If we can develop such procedures then we will be in a position to more carefully assess the confidence–accuracy relationship and to determine under what conditions the two measures are positively related as well as those conditions in which the relationship vanishes. We also sought to determine if we could find confidence–accuracy inversions; that is, conditions in which a greater propensity to make a mistake in reporting from memory is associated with greater confidence accompanying the mistake.

Before getting to our experiments we need to acknowledge that we are hardly the first researchers to seek such evidence. Tulving (1981) obtained

a confidence–accuracy inversion when subjects studied pictures of scenes and then had to discriminate between studied scenes and lures that were highly similar to other studied scenes (see also Chandler, 1994; Dobbins, Kroll, & Liu, 1998). He found situations in which subjects made high confidence errors when the lure scene on the test was similar to a different scene that had been studied. Of more direct relevance to our work is that of Brewer, Sampaio, and Barlow (2005) and Brewer and Sampaio (2006). They had subjects study what they termed *deceptive* and *non-deceptive sentences*. The deceptive sentences involved pragmatic implications and were ones like ''The baby stayed awake all night'' that Brewer had used in earlier work (Brewer, 1977; see too Chan & McDermott, 2006; McDermott & Chan, 2006). When subjects were tested on such sentences, they would often remember the sentence as having been ''The baby cried all night'', which is a pragmatic implication of the original sentence but is not correct. In other words, people remembered the inferences they drew. Further, Brewer found negative memory resolution, as measured by Goodman and Kruskal's gamma correlation, for deceptive sentences; the more confident an individual was in his or her responses to these sentences, the more likely he or she was wrong. (Resolution measures the correlation between confidence and accuracy for memory test responses within individual subjects.) On the other hand, for the non-deceptive sentences there was a positive correlation between confidence and accuracy. Brewer and Sampaio (2012) extended this work to semantic memory tasks. In this research the stimuli were deceptive statements such as ''San Antonio, Texas is south of Nogales, Mexico''. This is a true statement, but seems at odds with subjects' general knowledge that ''Mexico is south of Texas''. The mean gamma correlations between confidence and accuracy for these deceptive items were strikingly negative at –.52 and –.58 in two experiments.

Koriat (2008, 2012) has also produced impressive examples of positive and negative correlations between confidence and accuracy in answers to certain types of general knowledge questions. Earlier work by Koriat and Goldsmith (1996) using a wide variety of general knowledge questions showed that, overall, people display good memory resolution for this type of information. For example, when asked ''Who was the first emperor of Rome?'' subjects produced (in a recall version of the experiment) or selected (in a multiple-choice recognition version) the answer (Augustus). Then they were asked to provide a confidence rating for their response. Gamma correlations showed resolution was high, at .87 for the recall version of the experiment and .68 for the recognition version. Many other researchers have obtained similar results (see Perfect, 2002).

In his newer work, however, Koriat (2008) has examined a contrasting set of items that he refers to as *consensually incorrect*, meaning that most people miss these items. For example, if North Americans are asked ''What is the capital of Australia?'' many of them produce ''Sydney'' or select it on a recognition test (Melbourne, Canberra, Sydney, Queensland). Koriat discovered that the gamma correlation for these consensually incorrect items was negative, meaning that subjects were more likely to assign higher confidence ratings to items that had a greater probability of being incorrect. However, for *consensually correct* items (i.e., ones that subjects tended to get correct) or for difficult items (ones to which subjects just did not know the answer), resolution was positive. The consensually incorrect items in Koriat's (2008) experiment thus functioned much like the deceptive items in the research of Brewer and his associates. Note that the measure in all these experiments has generally been resolution.

A time-honoured distinction in the memory literature, dating at least from Bartlett (1932), is between reproductive and reconstructive memory (Roediger & DeSoto, in press). The term *reproductive memory* is intended to cover the Ebbinghaus (1885/1913) research tradition in which subjects learn materials such as nonsense words, digits, or word lists and are thought to mindlessly regurgitate the material later—to merely reproduce it. The *reconstructive memory* research tradition, in contrast, has emphasised use of more naturalistic materials (if we pretend ''The War of the Ghosts'' is natural) and has noted that retrieval processes are rife with errors, as in the results discussed above where people make high confidence errors in their recollections for sentences and general knowledge questions. Thus reconstructive memory is thought to be error prone, unlike reproductive memory for word lists and the like. Such a distinction may be ill founded, however, because similar high error rates can be obtained with simpler materials. Roediger and McDermott (1995) presented subjects with the 12 most common associates to

words (e.g., bed, rest, tired, awake, etc.) produced in norms as associates to words like ''sleep'' (which were never presented). In their first experiment subjects took a recognition test and produced confidence ratings to items called old. The hit rate for studied words (e.g., awake) was .86 and the false alarm rate for the non-studied words from which the lists had been derived (e.g., sleep) was .84. On average, hits received a confidence rating of 3.6 and critical false alarms to words like ''sleep'' were rated at 3.3 (on a 4-point scale, with 4 meaning *sure old*). Further, 58% of the critical lures were assigned the highest possible confidence response. Thus, even in an allegedly ''reproductive'' memory experiment, Roediger and McDermott (1995) found many high confidence false alarms, indicating apparent reconstructive processes operating in list-memory experiments.

The Deese-Roediger-McDermott paradigm just described (DRM; Deese, 1959; Roediger & McDermott, 1995) has been used in many experiments since 1995 (see Gallo, 2006, 2010, for reviews), but relatively few of these studies have employed confidence ratings. Rather, Experiment 2 in the Roediger and McDermott (1995) paper, in which 15-word lists and remember/know judgements were collected, became the standard version of the DRM paradigm. High rates of ''false remembering'' have been obtained in virtually all of these experiments. Because remember judgements and confidence often are highly correlated, it is likely that the high confidence outcome first reported in Experiment 1 of Roediger and McDermott's paper is highly replicable. We further examine the issue of confidence ratings in word list recall in the present research, albeit in a somewhat different paradigm than DRM.

Dewhurst and Anderson (1999) and Smith, Ward, Tindell, Sifonis, and Wilkenfeld (2000) developed materials that were somewhat analogous to DRM lists. These authors used category norms (e.g., Battig & Montague, 1969) in which subjects are given a category name (e.g., vegetable) and asked to produce as many items as they can from the category in a fixed period of time. The norms report the frequency of words produced in this procedure, a measure usually referred to as *output dominance* (so ''carrot'' is high on this dimension, produced by almost everyone, and ''rutabaga'' is low). Dewhurst and Anderson (1999) and Smith et al. (2000) showed that when items high in output dominance are omitted from a list of categorised items presented to subjects, the omitted items tend to be recalled and recognised falsely, akin to the findings of Roediger and McDermott (1995). Much work has been conducted comparing this task to associative false memory tasks like DRM (see Knott, Dewhurst, & Howe, 2012, for a recent discussion).

The particular version of the categorised list procedure we used in the present research was developed by Meade and Roediger (2006, 2009; see also Meade, Geraci, & Roediger, 2012). In this procedure subjects were presented with 15-item categorised lists in which the words were items 6–20 from the norms of common categories. The first five items in the categories were omitted from study so that false recall and false recognition of these words could be examined. As others have found, these high output dominance items were falsely recalled and recognised at relatively high rates. Our interest in using this procedure in the present research was to examine the relation between confidence and accuracy of both veridical and false recognition of items from categorised lists. We hypothesised that we could extend the findings of Brewer et al. (2005) and Koriat (2008) without the use of deliberately deceptive items (''The baby stayed awake all night'' or ''Sydney is the capital of Australia''). Rather we chose to use as lures the same type of items subjects studied, just ones higher in output dominance. We presented subjects with categorised lists and later tested them with a free choice (yes/no) recognition test and obtained confidence ratings on a 0–100 scale following each recognition judgement.

Our main interest was measuring the confidence–accuracy relationship in several ways to see if they agreed (Roediger et al., 2012). We used the measure that has been employed in prior work—resolution—or the within-subjects gamma correlation that answers the question: Are subjects more accurate for items that they rate with high confidence than those they rate with low confidence? However, we also used two other methods of assessing that confidence–accuracy relation that were introduced earlier, the ones Roediger et al. (2012) referred to as the between-subjects and the between-events measures. These measures answer the following two questions: First, are subjects who provide higher confidence ratings also more accurate in their judgements? Second, are items for which subjects are more confident also those for which they are more accurate? The answers to the three questions posed in this paragraph need not be the same,

and we shall see here that they are not the same. However, all are valid methods to assess the relation between confidence and accuracy, and the answer to all three questions is of interest. Researchers in the past generally have tended to ask the confidence–accuracy question in memory reports in only one way, however.

## EXPERIMENT 1

Experiment 1 was designed to determine if positive, negative, and zero correlations between confidence and accuracy could be demonstrated using the same set of materials: items from semantic categories. Subjects studied 15 words of intermediate output dominance (i.e., items 6–20 in the norms). After studying these items they then took a recognition test over four different types of items: (1) *studied items* (items 6–20 in the norms), (2) high output dominance members of the category that were never studied (*strongly related lures*; items 1–5 in the norms), (3) lower output dominance members of the category that were never studied (*weakly related lures*; items 21–25), and (4) items from completely new categories (*unrelated lures*). After judging whether each item was old or new, subjects rated their confidence in that judgement on a 0–100 scale.

Based on the work of Brewer and colleagues (Brewer & Sampaio, 2012; Sampaio & Brewer, 2009) and Koriat (2008, 2012), we predicted that across all materials the correlation between confidence and accuracy would be low but positive. However, considering analyses separately among the four classes of test items, we predicted both positive and negative correlations between confidence and accuracy. Specifically, we predicted that the confidence–accuracy correlation would be generally positive for studied items but negative for strongly related lures (those produced most frequently to the category name). This finding would indicate that the modest correlation between confidence and accuracy when collapsed across all items is the result of these underlying and contradictory positive and negative associations. The predictions for the confidence–accuracy relation for the weakly related lures and the unrelated lures were less certain, but at least for the latter type of lures, the expectation from past research was a positive correlation between accuracy and confidence. That is, for unrelated lures the prediction was that items rejected most frequently would be rejected with the highest confidence. For the weakly related lures (those of low output dominance), the expectation was a zero correlation between confidence and accuracy. Once again, we used three analytic methods (between-subjects, between-events, and within-subjects resolution) to assess the confidence–accuracy relation to see if the techniques would provide convergent or divergent answers.

### Method

*Subjects.* A total of 48 Washington University in St. Louis students participated for either course credit or payment.

*Materials.* Ten lists of words were selected from the revised Battig and Montague (1969) category norms (Van Overschelde, Rawson, & Dunlosky, 2004). These norms were obtained by asking a large sample of subjects to generate as many members of a given category (e.g., a bird) as possible. These responses were aggregated, producing lists of category members ordered by output dominance; that is, from the most frequently mentioned category member (eagle) to the least frequently mentioned category member used in the experiment (flamingo, in position 25).

The first 25 items from each of the 10 selected category norms were used as the stimulus set. If any item appeared twice in the stimulus set or could be categorised in another list (e.g., ''squash'' is both a vegetable and a sport), the item was removed from both lists and the appropriate 26th item was appended to each list. An example list can be found in the Appendix. Similar to the manipulation by Meade and Roediger (2006, 2009), the first five items of each list (eagle, robin, bluejay, cardinal, hawk) were never presented and reserved to serve as lures. These 50 items of high output dominance were termed strongly related lures. Items 6–20 (''bluebird'' through ''raven'') made up the studied set; these 150 items were termed studied items. Items 21–25 in output dominance (duck, finch, mockingbird, woodpecker, flamingo) were also never presented and were used as weakly related lures.

A recording was made of a female speaker reading the words into a Logitech desktop microphone in tandem with Apple GarageBand software installed on an Apple MacBook Pro. The speaker read the category name, paused for four seconds, then read each of the category members

in a pre-randomised order at a rate of one word per two seconds. Digital post-processing eliminated any extraneous sounds from the recording.

*Procedure.* During the study phase, subjects were given intentional learning instructions. They were then seated at a computer and listened via headphones to the audio recording of the speaker reading one of the 10 word lists. Once the category name and all 15 words were presented, the procedure was repeated with the remaining lists. The order of lists was randomised for each subject. After all 150 words were presented, subjects engaged in a 5-minute distractor task in which they generated and then chronologically ordered as many United States presidents as possible. This task eliminated contributions from short-term memory in later recall.

During the test phase subjects were given a recognition test of 300 words, one at a time and randomly presented, from the following item sets: the 150 studied items (items 6–20 in terms of output dominance from each of the 10 original category lists), 50 strongly related lures (items 1–5 from the 10 lists), 50 weakly related lures (items 21–25 from the 10 lists), and 50 unrelated lures taken from categories not used in the study lists. Subjects indicated with the numeric keypad whether they believed each word to be old (studied) or new (unstudied). After subjects made this old/new judgement they were prompted to enter their confidence in the judgement, also via numeric keypad, from 0 (labelled *not at all confident*) to 100 (*entirely confident*). Recognition judgements and confidence ratings were self-paced, and the entire experiment took approximately 45 minutes.

## Results

The left side of Table 1 shows the probability of responding "old" to items of each type. The hit rate was .70. As expected, the false alarm rates were greatest for strongly related lures and lowest for unrelated lures, with the weakly related lures intermediate. The confidence ratings for the various item types are shown in the left side of Table 2. We were unable to collect confidence ratings for 2% of responses because subjects either omitted ratings on some items or gave ratings outside the 0–100 scale (e.g., entering 900 when they probably intended to enter 90). We excluded these responses from further analysis.

**TABLE 1**
Probability of saying "old" as a function of item type in Experiment 1 and Experiment 2

| | *Experiment 1* | | *Experiment 2* | |
|---|---|---|---|---|
| *Item type* | *Hits* | *False alarms* | *Hits* | *False alarms* |
| Studied Items | .70 | | .70 | |
| Strongly Related Lures | | .44 | | .43 |
| Weakly Related Lures | | .30 | | .28 |
| Unrelated Lures | | .12 | | .08 |

Although subjects were highly confident when providing hits, the mean confidence ratings for the three types of false alarms were considerably lower and varied significantly from one another, $F(2, 72) = 25.90$, $MSE = 140.07$. Confidence was highest for false alarms to strong lures (68) and weakest for those to unrelated lures (48) with the weak lures of intermediate value (64).

We analysed the confidence–accuracy relationships in three ways (Roediger et al., 2012). First, we examined the between-subjects correlation, which indicates the degree to which subjects who are more confident also tend to be more accurate. Second, we examined the between-events correlation, which indicates the degree to which items responded to more confidently also tend to be responded to more accurately. These were calculated with the Pearson *r*. Finally, we examined the within-subjects correlation, which indicates whether higher confidence was associated with higher accuracy on a subject-by-subject basis. This was calculated with the Goodman-Kruskal gamma ($\gamma$). All results were statistically significant ($p < .05$) unless otherwise noted.

The correlations found in Experiment 1 are summarised in the left half of Table 3. For the between-subjects component of our analysis, confidence and accuracy were averaged across each of the 48 subjects for each item category and a correlation coefficient was computed. The results are in the first column in Table 3 on the far left. A moderate positive correlation was found between the two variables over all item types, $r(46) = .29$. Additionally, a positive correlation was found between confidence and accuracy for the 150 studied items, $r(46) = .42$, but no association significantly different from zero was found between confidence and accuracy for the 50 strongly related lures, $r(46) = -.09$, $p > .05$, or the 50 weakly related lures, $r(46) = -.05$, $p > .05$. However, a positive correlation was found for the 50 unrelated lures, $r(46) = .37$. Thus subjects who

**TABLE 2**
Mean confidence as a function of item type in Experiment 1 and Experiment 2

| | *Experiment 1* | | | | *Experiment 2* | | | |
|---|---|---|---|---|---|---|---|---|
| *Item type* | *Hits* | *Misses* | *False alarms* | *Correct rejections* | *Hits* | *Misses* | *False alarms* | *Correct rejections* |
| Studied Items | 84 | 55 | | | 82 | 49 | | |
| Strongly Related Lures | | | 68 | 60 | | | 61 | 53 |
| Weakly Related Lures | | | 64 | 63 | | | 55 | 55 |
| Unrelated Lures | | | 48 | 67 | | | 46 | 61 |

were more confident for studied items or unrelated lures were also more accurate for those items (in terms of hits or correct rejections, respectively). In contrast, the correlation for related lures was essentially zero between subjects, with no relation between confidence and accuracy for these materials.

Next, for the between-events component of our analysis, confidence and accuracy were averaged across each item for all 300 items and a correlation coefficient was computed to assess the relationship between confidence and accuracy for the entire item set (see the second column in Table 3). A positive correlation was found between the two variables aggregated over all items, $r(298) = .35$. Additionally, a positive correlation was found between confidence and accuracy for the 150 studied items, $r(298) = .70$. On the other hand, a significant negative correlation was found between confidence and accuracy for the 50 strongly related lures, $r(298) = -.54$. A significant positive correlation was found between confidence and accuracy for the 50 weakly related lures, $r(298) = .31$, but no significant correlation was found for the 50 unrelated lures, $r(298) = .21$, $p > .05$. These analyses indicate that studied items and weakly related lures that were recognised more confidently were also recognised more accurately, but for strongly related lures (those like eagle), higher confidence actually predicted lower accuracy.

Finally, for the within-subjects (resolution) component of our analyses, five gamma correlations were calculated for each subject: One across the subject's responses to all items and one across each of the subject's responses to the four item types (studied items, strongly related lures, weakly related lures, and unrelated lures). These gamma correlations were then averaged across subjects. The within-subjects gamma correlation for all items ($\gamma = .38$, $SD = .20$) was significantly greater than zero, $t(47) = 13.29$. This overall positive correlation, however, masked an even stronger significantly positive, $t(47) = 28.00$, within-subjects relationship between confidence and accuracy for studied items ($\gamma = .72$, $SD = .18$). As in the between-events analysis, a significantly negative, $t(47) = 3.76$, confidence − accuracy relationship existed for strongly related lures ($\gamma = -.23$, $SD = .42$). The within-subjects confidence–accuracy correlation for weakly related lures ($\gamma = -.05$, $SD = .39$) was not shown to be different from zero, $p > .05$, but the correlation for unrelated lures ($\gamma = .29$, $SD = .50$) was also positive, $t(36) = 3.56$. These gamma correlations indicate that subjects were more likely to correctly re-

**TABLE 3**
Correlations between confidence and accuracy as a function of item type in Experiment 1 and Experiment 2

| | *Experiment 1* | | | *Experiment 2* | | |
|---|---|---|---|---|---|---|
| *Item type* | *Between subjects* | *Between events* | *Within subjects* | *Between subjects* | *Between events* | *Within subjects* |
| All items | .29* | .35* | .38* | .48* | .12* | .38* |
| Studied items | .42* | .70* | .72* | .62* | .69* | .73* |
| Strongly related lures | −.09 | −.54* | −.23* | −.18 | −.34* | −.23* |
| Weakly related lures | −.05 | .31* | −.05 | .17 | .14 | −.02 |
| Unrelated lures | .37* | .21 | .29* | .44* | .29* | .08 |

Between-subjects and between-events correlations are reported as Pearson correlations. Within-subjects correlations are reported as Goodman-Kruskal gamma correlations. An asterisk (*) denotes that the correlation is significantly different from zero ($p < .05$) as measured by Pearson $r$ (for the between-subjects and between-events analyses) or independent-samples $t$-tests (for the within-subjects analyses).

spond to studied items and unrelated lures that were assigned high confidence ratings. However, when an individual subject responded with high confidence to a strongly related lure, he or she was less likely to be accurate on that item than if confidence were lower. This confidence–accuracy inversion obtained in both the between-events and resolution measures is of particular interest.

## Discussion

The results showed widely different false alarm rates for the three types of lures, with lures of higher output dominance (1–5) producing more false alarms than those of lower output dominance (21–25), but both types of related lures produced higher false alarm rates than the unrelated lures. The results of Experiment 1 show that, for the same data, whether confidence and accuracy are related is sometimes a matter of how the relation is measured, especially for lures highly related to the target items.

For studied items a relatively strong correlation between confidence and accuracy was found for all three techniques (between-subjects, between-events, and within-subjects). When all item types were included in the analysis the correlation was still positive but less strong, because the correlation between confidence and accuracy was much weaker or even negative for strongly related lures (in two methods of analysis) and also generally weaker for the other types of lures. The variable relation between confidence and accuracy with lure items therefore offset the strongly positive correlation between confidence and accuracy for studied items and made the correlation across all item types weaker.

The variation between confidence and accuracy across three types of lures is interesting. When analysed across the 50 strongly related lures in the between-events analysis the correlation was rather strongly negative ($-.54$); the items eliciting the greatest proportion of false alarms also led subjects to make these errors with high confidence. The same negative relation was observed for the strongly related lures when resolution was measured, with individual subjects' judgements of confidence being negatively related ($-.23$) to their accuracy. However, when measured between subjects the correlation for these highly related lures did not differ significantly from zero ($-.09$). For the more weakly related lures (items with output dominance 21–25), the confidence–accuracy relation hovered around zero except for the analysis between events, where it was modestly positive. Thus, even with two types of lures that were both members of the studied categories, the relation between confidence and accuracy varied depending on the type of analysis. We defer discussion of these results until we report a second experiment that eliminated the problem with measuring confidence using the keypad that caused some responses to be excluded in Experiment 1.

# EXPERIMENT 2

Experiment 1 successfully demonstrated that it is possible to show positive, zero, and negative correlations between confidence and accuracy within the same general class of materials—items from categorised lists—depending on the subset of items analysed and the method used. Because a small fraction of the data (2%) was missing, however, we sought to replicate Experiment 1 with another procedure that used an on-screen slider, rather than a numeric keypad entry system, to collect confidence ratings. The slider technique was less susceptible to errors or omissions of responding (subjects could not proceed with the experiment before moving the slider). Of course, because our findings are novel, replication is in order anyway. To our knowledge, no one has previously obtained negative correlations between confidence and accuracy in a list-learning task nor considered the various methods of measuring confidence–accuracy relations on the same data.

## Method

*Subjects and materials.* A total of 48 Washington University in St. Louis students participated for either course credit or payment. The same materials used in Experiment 1 were used in Experiment 2 and were presented the same way.

*Procedure.* The procedure was identical to that used in Experiment 1, except for one change: Instead of entering confidence judgements into the computer via numeric keypad after making old/new judgements, subjects reported confidence

by clicking and dragging with the mouse cursor on an on-screen slider. Subjects could not progress to the next word until this rating was made. This change was made so that a confidence rating from 0–100 was collected for every recognition judgement and submitting scores outside the permissible range was impossible. The cursor was placed on 50 to begin each rating and subjects moved it to the desired position on the scale.

## Results

Results were analysed in the same manner as in Experiment 1. The right panels of Tables 1 and 2 show, for different categories of test item, the probability of responding "old" to an item type (Table 1) and average confidence assigned to "old" and "new" responses to different item types (Table 2). The data on the right side of Table 1 show almost exactly the same values as those on the left, showing a highly replicable pattern at this aggregate level. The hit rates and three false alarm rates are all within a few percentage points of those obtained in Experiment 1. The confidence ratings on the right side of Table 2 were generally lower than those on the left, but the same patterns were observed for false alarms: highest confidence for false alarms to strongly related lures (61) and lowest confidence for false alarms to unrelated lures (46), with the confidence for weakly related false alarms at an intermediate level (55), $F(2, 60) = 12.33$, $MSE = 138.17$.

Again, our main interest was in the three types of confidence–accuracy analyses for the four item types, as shown on the right side of Table 3. For the between-subjects analysis, confidence and accuracy were averaged across each of the 48 subjects for each item category and Pearson $r$ was computed to assess the relation between confidence and accuracy. A positive correlation was found between the two variables when averaged across all items, $r(46) = .48$. Additionally, a positive correlation was found between confidence and accuracy for the 150 studied items, $r(46) = .62$, but no significant correlation was found between confidence and accuracy for the 50 strongly related lures, $r(46) = -.18$, $p > .05$, or the 50 weakly related lures, $r(46) = .17$, $p > .05$. A positive correlation was found for unrelated lures, however, $r(46) = .44$. Thus subjects who were more confident when responding to studied items or unrelated lures were also more accurate for those items, with no significant relation existing for the strongly and weakly related lures. These results are quite consistent with similar analyses in Experiment 1.

Next, for the between-events component of our analysis confidence and accuracy were averaged across each item for each type of item (studied items or the three lure types). When Pearson $r$ was computed between confidence and accuracy for all 300 words, a low but significant result obtained, $r(298) = .12$. Considering the same analysis for the 150 studied items, the correlation was large and positive, as in Experiment 1, $r(298) = .69$. However, a negative correlation was again obtained between confidence and accuracy for the 50 strongly related lures, $r(298) = -.34$. No significant correlation was found between confidence and accuracy for the 50 weakly related lures, $r(298) = .14$, $p > .05$, but a positive correlation was found for the 50 unrelated lures, $r(298) = .29$. Thus, as in Experiment 1, the between-events analysis showed a significant positive correlation for the studied items and unrelated lures. However, we replicated the finding of a significant negative correlation for the strongly related lures. Unlike Experiment 1 we did not obtain a significant positive correlation for the weakly related lures but we did find one for the unrelated lures.

Finally, for the within-subjects (resolution) component of our analysis, gamma correlations were again calculated for each subject for each item type and collapsed across subjects. As in Experiment 1 the mean gamma correlation for subjects across all item types ($\gamma = .38$, $SD = .19$) was modestly (but significantly) positive, $t(47) = 13.85$. The correlation between confidence and accuracy for studied items ($\gamma = .73$, $SD = .20$) was strongly positive, $t(47) = 26.54$. However, we replicated the negative value of resolution, $t(47) = 3.77$, for strongly related lures ($\gamma = -.23$, $SD = .42$). In contrast, the within-subjects confidence–accuracy correlations for weakly related lures ($\gamma = -.02$, $SD = .36$) and unrelated lures ($\gamma = .08$, $SD = .57$) were not shown to be statistically different from zero, $p > .05$. As in Experiment 1 resolution was quite good for studied words but poor for lures. In fact, for the 50 high output dominance lures, the more likely subjects were to false alarm, the more confidence they had in their erroneous responses.

## Discussion

The results of Experiment 2 were quite consistent with the results of Experiment 1, at least for studied items and strongly related lures (the two item types of most interest). Subjects were more likely to incorrectly respond "old" (i.e., commit a false alarm) to high output dominance lures relative to other lure types. Considering only the 150 studied items, all three types of analysis revealed positive correlations between confidence and accuracy. However, for the three types of lures, confidence–accuracy correlations varied from modestly positive (for unrelated lures) to negative (for two analyses of the high output dominance lures). The data concerning strong lures (those of high output dominance) were of most interest. When analysing across the 50 items, both experiments showed negative correlations between confidence and accuracy. Further, we replicated the finding of negative resolution for the high dominance lures.

# GENERAL DISCUSSION

The categorised list procedure used in our experiments led to high levels of false recognition for the most frequently produced items according to category norms (those high in output dominance). In addition, the main goals of our experiments were achieved. We had hypothesised that when we used different methods of analysis to assess the relation between confidence and accuracy of memory reports, we would discover positive, zero, and negative correlations depending on the measure and the type of items used. All three patterns were obtained. We discuss these various findings in turn.

## False recognition of categorised list items

We used a procedure similar to those developed over a decade ago by Dewhurst and Anderson (1999) and Smith et al. (2000), although the specific version we used in which the five items highest in output dominance were omitted was that of Meade and Roediger (2006, 2009). In our experiments we assessed false recognition for these items (50, five each across ten categories) as well as for other lures of much lower output dominance (50 additional lures that were items 21–25 in the norms). Relative to 50 unrelated lures (with a .10 false alarm rate, averaged across the two experiments), lures high in output dominance produced a .44 false alarm rate whereas those lower in output dominance still created a sizeable false alarm rate of .29. Although these false alarm rates for related lures are not as high as those typically seen in the DRM paradigm, they are still quite large when one considers that the materials are word lists and the retention interval is relatively short. Because error rates like these are often taken as the hallmark of reconstructive rather than reproductive memory processes, it is clear that even recognition of word lists seems to display reconstructive features.

To return to the questions raised at the outset of this paper—asking how one can reconcile an error-prone memory with adaptive fitness—we can provide a tentative answer. Our memory systems seem tuned to reconstruct plausible events that might have happened in the past. These can include the most readily accessible members of categories (as in our experiments and those of others), words associated with members of a list (as in the DRM paradigm; Roediger & McDermott, 1995), implications of statements that might have been inferred during encoding (Brewer, 1977) or material that fits a schema (e.g., Bartlett, 1932; Brewer & Treyens, 1981). In general, remembering is adjusted to the broad parameters of how events were experienced in the environment, and people use these parameters in helping to reconstruct what happened in the past. We can only suppose (but not prove) that this reconstructive process usually improves our recollections of the past even if it can lead us astray on occasion.

In general, the confidence data for false alarms in Table 2 follow in an orderly way the proportion of false alarms observed in Table 1. When combined across experiments, confidence ratings were greatest for false alarms to strongly related items (65), next for weakly related false alarms (60), and lowest for false alarms to unrelated items (47). Thus, at the aggregate level, confidence is positively correlated with the proportion of errors across the three types of lures; stated differently, there is a negative correlation between confidence and accuracy. However, these aggregate judgements of confidence across the three types of lures are not particularly incisive about addressing issues of confidence–accuracy relations, for which our other measures are more

informative. Future investigations may benefit from examining confidence ratings as a function of output dominance rather than lure type as in the current experiments. The primary interest in the current experiments, however, was in the three ways of assessing confidence–accuracy relations rather than overall performance. We turn to these next.

## Confidence–accuracy relations

As noted in the introduction, researchers have tended to ask about confidence–accuracy relations as though only one answer is possible, and the answers typically given (depending on the research) have been that the two are positively correlated (e.g., Dunlosky & Metcalfe, 2009) or that they are not correlated (e.g., Smith et al., 1989). Prior research has shown both these patterns and, for deceptive sentences or consensually incorrect items of general knowledge, even negative correlations have been obtained (Brewer & Sampaio, 2006; Koriat, 2008). In the current experiments we show all three possible patterns using the same general types of items (words from categorised lists), which is the first time such observations have been reported (see Roediger et al., 2012, for review).

*Between-subjects analysis.* This type of analysis asks if subjects who are more accurate are also more confident (and, obviously, vice versa; the issue can be framed either way). When the question is posed across all item types, the answer is generally yes (.29 and .48 in the two experiments; see Table 3). However, when the data are decomposed into the various item types, quite different answers were obtained. For studied items and unrelated lures a relatively strong positive correlation was shown between confidence and accuracy in both experiments; both hits and correct rejections of unrelated lures are correlated with confidence across subjects. On the other hand, the situation is quite different for the strong and weak lures. The correlation between confidence and accuracy across subjects for strong lures is about zero in Experiment 1 and slightly negative in Experiment 2. For the weak lures the corresponding values are about zero and slightly positive in the two experiments. Obviously no strong conclusions can be drawn regarding the strong and weak lures, but between subjects, confidence and accuracy are positively related for both studied items and for the unrelated lures.

Relatively few other experiments examining the confidence–accuracy relation between subjects exist in the literature (see Roediger et al., 2012, pp. 106–108). Perfect, Watson, and Wagstaff (1993) reported positive confidence–accuracy correlations between subjects for general knowledge questions but not for episodic memory for movie scenes which are likely more relevant for eyewitness testimony. In the latter case the correlation between subjects for confidence and accuracy was about zero (see too Perfect, Hollins, & Hunt, 2000; Robinson & Johnson, 1996).

*Between-events correlations.* This type of analysis has rarely been examined in the literature on confidence and accuracy, leading Roediger et al. (2012) to declare it "surprisingly overlooked" (p. 103). However, in one study Sampaio and Brewer (2009) examined non-deceptive and deceptive sentences and obtained a modest positive correlation between the confidence and accuracy for non-deceptive sentences while reporting a strongly negative correlation for the deceptive sentences (–.61). Our analysis across events (with four types of items) conceptually replicates their work. For our studied words (items 6–20 in the norms) we found strongly positive correlations (.70 and .69 in the two experiments). The relation for these 150 words (15 words in 10 categories, with data combined across the two experiments) is shown in Figure 1. Looking only at these data one would conclude (as many have in the list-learning tradition) that the correlation between confidence and accuracy is high. The confidence–accuracy correlation is also significantly positive, if more modest, for the unrelated lures in both experiments.

The data for related lures tell a much different story, however. For strongly related lures (as for the deceptive sentences used by Sampaio & Brewer, 2009) there was a sizeable negative correlation between confidence and accuracy ($-.54$ and $-.34$ in the two experiments). Figure 2 shows the data for these 50 lures collapsed across experiments for greater power. Interestingly this negative correlation held only for the strongly related lures; for weakly related lures, results from both experiments showed modest positive correlations between confidence and accuracy.

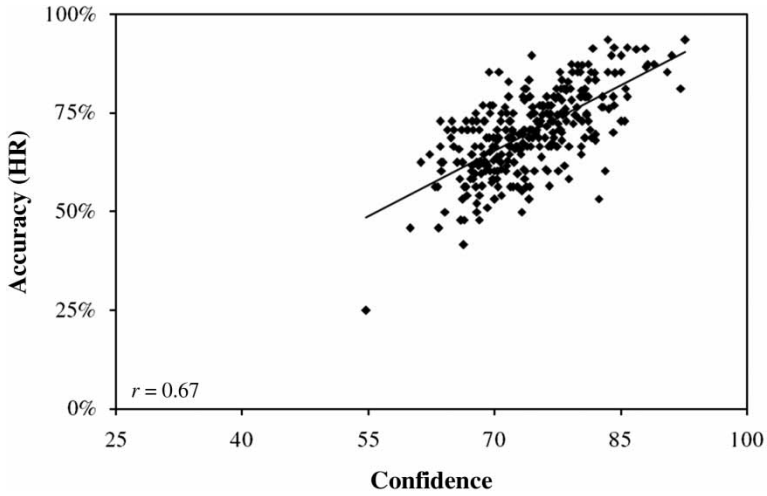


**Figure 1.** The between-events plot for studied items in Experiments 1 and 2. This plot shows that those studied items to which subjects are more likely to correctly respond ''old'' (i.e., hit) are also more likely to be assigned higher confidence ratings.

The findings with strongly related lures may have implications for the problem of high confidence mistakes in eyewitness memory situations. Similarity of lures and targets clearly matters for recognition. In general one may surmise that the greater the similarity between a person in a lineup and the actual suspect, the more likely (all else being equal) that a false identification will occur with high confidence. Although the point seems obvious, we can find few experiments that actually demonstrate it, perhaps because confidence ratings are not generally taken in eyewitness research (but see Brewer & Wells, 2006).

Our results, along with those of Sampaio and Brewer (2009), show that type of events to be remembered is critical in determining confidence–accuracy correlations. If a researcher examined all items in our experiment (without decomposition into various types), he or she would have concluded that there is a modest but significant correlation between the two variables across events. However, the fact that different patterns are obtained for subclasses of items shows that this conclusion would be misleading.

We argued in the introduction that illusions of memory, like illusions of perception, may arise when normally adaptive mental processes face unusual environmental challenges. In perceptual illusions often depth cues induced by a two-dimensional image to make it appear three-

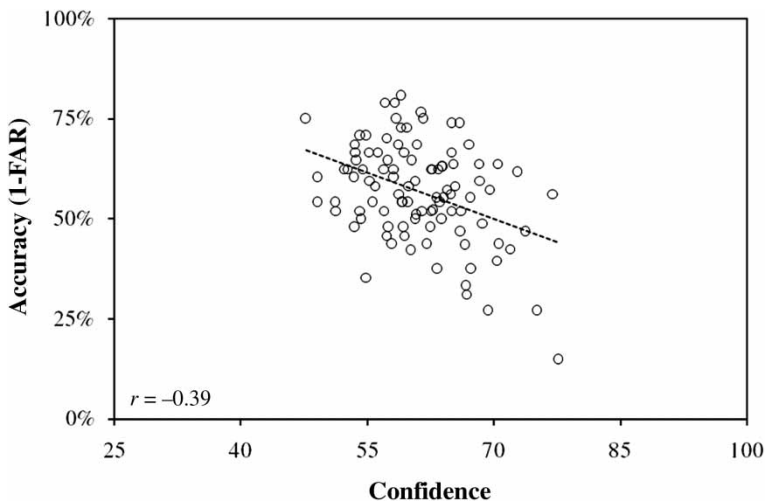


**Figure 2.** The between-events plot for strongly related lures in Experiments 1 and 2. This plot shows that those strong lure items to which subjects are more likely to correctly respond ''new'' (i.e., correctly reject) are also more likely to be assigned *lower* confidence ratings.

dimensional lead to illusions of size: The same image is seen as larger if it appears farther away than another version that appears to be closer. This is the general basis for the well-known Ponzo (or railroad) illusion and its many varieties. Similarity seems to provide the same general role in memory illusions. Just as humans learn about depth cues over time, so too do they learn to be sensitive to metacognitive variables such as when to be confident in their memories. Mickes, Hwe, Wais, and Wixted (2011) noted that the metacognition literature shows that young adults generally have better metacognitive accuracy than children. This fact probably indicates that as we age we gain experience with, and learn to attend to, the internal cues that are diagnostic of accuracy. Thus, as many experiments (including the ones reported here) show, young adults can show excellent metacognitive accuracy, as in the .70 correlation between confidence and accuracy for studied items in the current research. However, when the environment provides cues that usually indicate good metacognitive accuracy but that now occur in a misleading context, subjects' memory decisions and their confidence in their judgements can be misapplied, leading to errors. The strongly related lures in the present experiments provide just this sort of misleading cue, leading to strong confidence–accuracy inversions. ''Eagle'' and similar dominant category members create the illusion of remembering, and of doing so with high confidence. The DRM procedure similarly leads to highly confident but erroneous judgements due to great associative activation accruing for items that were never presented (Gallo, 2010).

*Within-subjects correlations (resolution).* The gamma correlation to measure resolution has been the most common way to assess confidence–accuracy correlations. Using categorised lists, we have conceptually replicated Brewer et al.'s (2012) and Koriat's (2008, 2012) findings using non-deceptive or consensually correct materials as well as those that are deceptive or consensually incorrect. That is, for studied items and unrelated lures we find positive gamma correlations, so that the more accurate subjects are, the more confident they are. However, for the strong lures high in output dominance, a negative correlation exists—the more false alarms subjects make, the more confident they are (as with deceptive materials). Further, for the weakly related lures a zero correlation exists. Thus, with four types of items, we show three different patterns of resolution – positive, zero, and negative. There is no simple relation between confidence and accuracy in measures of resolution.

*A puzzle for the future.* Signal detection theorists, along with most recognition memory researchers, assume that various values of ''strength of evidence'' are all positively correlated with confidence. As Wixted and Mickes (2010) put it, ''The concept of memory strength applies naturally to a variety of behavioral measures that tend to covary, such as confidence, accuracy, and reaction time. Generally speaking, memories are said to be strong when they are associated with relatively high confidence, high accuracy, and fast reaction times'' (p. 1025). We can certainly understand how signal detection theory can account for false memories as shown in our general pattern in Table 1 (e.g., Wixted & Stretch, 2000). One simply assumes that the distribution of studied items is highest, then strongly related lures, then weakly related lures, then unrelated lures. The interesting question is how the theory can account for confidence–accuracy inversions as we have shown (using two different methods) for strongly related lures or the most commonly produced items in a category.

One general approach to this issue, suggested by John Wixted (personal communication, 19 March 2013), is that when several different types of lures are used (as in our experiment, but unlike standard recognition procedures), one must consider the lure distributions as different among themselves in order to apply SDT. In our experiments the four different item classes (studied items, strongly related lures, weakly related lures, and unrelated lures) can be conceptualised as distributions of items centred on different points on the ''strength of evidence'' axis. The confidence ratings assigned to items of each item class suggest that studied items had a greater strength of evidence (i.e., were greater in ''oldness'') than strongly related lures, strongly related lures had greater strength than weakly related lures, and weakly related lures had greater strength than unrelated lures.

If each subject's criterion—that is, the value on the strength of evidence axis at which he or she feels justified in calling an item at test ''old''—is assumed to fall amidst these four distributions, it is likely that for studied items, a large proportion of the distribution is called ''old'' but only a small

proportion is called ''new''. Also, because the ''old'' portion of the studied item distribution stretches further into the ''old'' half of the strength of evidence axis than the ''new'' half (as divided by the criterion), ''old'' responses are given higher confidence ratings, on average, than ''new'' responses. Thus, due to the position of the studied item distribution, subjects are more likely to (1) respond to these items correctly than incorrectly and (2) assign correct responses a greater confidence rating than incorrect responses on a recognition test. Thus a positive confidence–accuracy relation emerges.

The strongly related lure distribution also stretches beyond the criterion into the ''old'' half of the strength of evidence axis. Just as for studied items, ''old'' responses are likely for strongly related lures and are also assigned higher confidence than ''new'' responses. In contrast to studied items, however, ''old'' responses to strongly related lures are incorrect. This means that due to the position of the strongly related lure distribution, subjects are more likely to (1) respond to these items incorrectly and (2) assign incorrect responses a higher confidence rating at test. Thus a negative confidence–accuracy relation for these items emerges. Because the distributions for weakly related lures and unrelated lures do not stretch as far into the ''old'' portion of the strength axis, positive or zero correlations obtain for these items, depending on the precise location of the distributions.

This theoretical explanation permits us to conclude that SDT may provide a satisfactory account of confidence–accuracy inversions, but a further explication of this point is beyond the scope of this paper.

## CONCLUSION

The question of ''how are confidence and accuracy related in reports from memory?'' has often been asked in research reports, but it has no general answer. As with most problems in the psychology of memory, the answer is ''it depends'' (Roediger, 2008). We used a categorised list procedure and three different methods of analysis to show that strongly positive, strongly negative, and zero correlations can be obtained with the same set of data and using straightforward materials—categorised word lists. This means that in some circumstances, one can rely on confidence as a proxy for ''strength of evidence that a memory is correct'' but in other situations the opposite holds true and caution is warranted. Unfortunately we can provide no firm answer to the question that began our paper: How are confidence–accuracy inversions possible in an evolved memory system that is adaptive? The best answer we can provide, as often maintained for perceptual illusions, is that such errors represent cases of a normally adaptive system going awry in a particular situation due to environmental circumstances. In our experiments the strongly related lures were quite similar to the target items (or perhaps represented the more prototypical members of the same class of items) and hence led to false memories that were held with high confidence. For these items, responses lower in confidence were more accurate than those higher in confidence, the reverse of the usual intuition that people have learned from experience.

Manuscript received 5 February 2013
Manuscript accepted 11 April 2013
First published online 21 May 2013

## REFERENCES

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, UK: Cambridge University Press.

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut norms. *Journal of Experimental Psychology*, *80*, 1–46.

Benedetti, F. (1985). Processing of tactile spatial information with crossed fingers. *Journal of Experimental Psychology: Human Perception and Performance*, *11*, 517–525.

Brewer, N., & Wells, G. L. (2006). The confidence–accuracy relationship in eyewitness identification: Effect of line-up instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*, 11–30.

Brewer, W. F. (1977). Memory for the pragmatic implications of sentences. *Memory & Cognition*, *5*, 673–678.

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

Brewer, W. F., & Sampaio, C. (2012). The metamemory approach to confidence: A test using semantic memory. *Journal of Memory and Language*, *67*, 59–77.

Brewer, W. F., Sampaio, C., & Barlow, M. R. (2005). Confidence and accuracy in the recall of deceptive and non-deceptive sentences. *Journal of Memory and Language*, *52*, 618–627.

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*, *13*, 207–230.

Chan, J. C. K., & McDermott, K. B. (2006). Remembering pragmatic inferences. *Applied Cognitive Psychology*, *20*, 633–639.

Chandler, C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition*, *22*, 273–280.

Coren, S., & Girgus, J. S. (1978). *Seeing is deceiving: The psychology of visual illusions*. Oxford, UK: Erlbaum.

Darwin, C. (1872/1958). *The origin of species (6th ed)*. New York, NY: New American Library, Signet.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17–22.

Dewhurst, S. A., & Anderson, S. J. (1999). Effects of exact and category repetition in true and false recognition memory. *Memory & Cognition*, *27*, 665–673.

Dobbins, I. G., Kroll, N. E. A., & Liu, Q. (1998). Confidence–accuracy inversions in scene recognition: A remember/know analysis. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24*, 1306–1315.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: Sage Publications.

Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology*. New York, NY: Teachers College, Columbia University.

Gallo, D. A. (2006). *Associative illusions of memory: False memory research in DRM and related tasks*. New York, NY: Psychology Press.

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, *38*, 833–848.

Garrett, B. L. (2012). *Convicting the innocent: Where criminal prosecutions go wrong*. Cambridge, MA: Harvard University Press.

Knott, L. M., Dewhurst, S. A., & Howe, M. L. (2012). What factors underlie associative and categorical memory illusions? The roles of backward associative strength and interitem connectivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 229–239.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 945–959.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*, 80–113.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490–517.

McDermott, K. B., & Chan, J. C. K. (2006). Effects of repetition on memory for pragmatic influences. *Memory & Cognition*, *34*, 1273–1284.

Meade, M. L., Geraci, L. D., & Roediger, H. L. (2012). Neuropsychological status in older adults influences susceptibility to false memories. *The American Journal of Psychology*, *125*, 449–467.

Meade, M. L., & Roediger, H. L. (2006). The effect of forced recall on illusory recollection in younger and older adults. *The American Journal of Psychology*, *119*, 433–462.

Meade, M. L., & Roediger, H. L. (2009). Age differences in collaborative memory: The role of retrieval manipulations. *Memory & Cognition*, *37*, 962–975.

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.

*Neil v. Biggers, 409 U.S. 188*. (1972). U.S. Supreme Court Decision.

Odinot, G., Wolters, G., & van Koppen, P. J. (2009). Eyewitness memory of a supermarket robbery: A case study of accuracy and confidence after 3 months. *Law and Human Behavior*, *33*, 506–514.

Perfect, T. J. (2002). When does eyewitness confidence predict performance? In T. Perfect & B. Schwartz (Eds.), *Applied metacognition* (pp. 95–120). Oxford, UK: Oxford University Press.

Perfect, T. J., Hollins, T. S., & Hunt, A. L. (2000). Practice and feedback effects on the confidence–accuracy relation in eyewitness memory. *Memory*, *8*, 235–244.

Perfect, T. J., Watson, E. L., & Wagstaff, G. F. (1993). Accuracy of confidence ratings associated with general knowledge and eyewitness memory. *Journal of Applied Psychology*, *78*, 144–147.

Pohl, R. F. (2004). *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Hove, UK: Psychology Press.

Robinson, M. D., & Johnson, J. T. (1996). Recall memory, recognition memory, and the eyewitness confidence–accuracy correlation. *Journal of Applied Psychology*, *81*, 587–594.

Roediger, H. L. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology*, *59*, 225–254.

Roediger, H. L., & DeSoto, K. A. (in press). The psychology of reconstructive memory. In J. Wright (Ed.), *International encyclopaedia of the social and behavioural sciences* (2nd ed.). Oxford, UK: Elsevier.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803–814.

Roediger, H. L., Wixted, J. H., & DeSoto, K. A. (2012). The curious complexity between confidence and accuracy in reports from memory. In L. Nadel & W. Sinnott-Armstrong (Eds.), *Memory and law* (pp. 84–118). Oxford, UK: Oxford University Press.

Sampaio, C., & Brewer, W. F. (2009). The role of unconscious memory errors in judgements of confidence for sentence recognition. *Memory & Cognition*, *37*, 158–163.

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.

Smith, V. L., Kassin, S. M., & Ellsworth, P. C. (1989). Eyewitness accuracy and confidence: Within-versus between-subjects correlations. *The Journal of Applied Psychology*, *74*, 356–359.

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior*, *20*, 479–496.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded

version of the Battig and Montague [1969] norms. *Journal of Memory and Language*, *50*, 289–335.
Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgements. *Psychological Review*, *117*, 1025–1054.
Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review*, *107*, 368–376.

# APPENDIX A

## An example word list

| | |
|---|---|
| **a vegetable** | |
| 1. carrot | 1–5: strongly related lures |
| 2. lettuce | |
| 3. broccoli | |
| 4. tomato | |
| 5. cucumber | |
| 6. pea | 6–20: studied items |
| 7. corn | |
| 8. potato | |
| 9. celery | |
| 10. onion | |
| 11. bean | |
| 12. spinach | |
| 13. cauliflower | |
| 14. cabbage | |
| 15. radish | |
| 16. asparagus | |
| 17. pepper | |
| 18. beet | |
| 19. pumpkin | |
| 20. turnip | |
| 21. zucchini | 21–25: weakly related lures |
| 22. yam | |
| 23. leek | |
| 24. rutabaga | |
| 25. artichoke | |
| centimeter | unrelated lures |
| amethyst | |
| dresser | |
| canyon | |
| whisk | |