22

UNDERSTANDING THE RELATION BETWEEN CONFIDENCE AND ACCURACY IN REPORTS FROM MEMORY

Henry L. Roediger, III and K. Andrew DeSoto

The issue we address is central to understanding our memories and when we can trust them: How confident can we be in our memories for past events? Might we be confident that our recollections are accurate but be totally wrong? This issue of confidence in our memories is critical to understanding who we are—we define ourselves in many ways by memories of our past experiences—and is also central to legal issues such as when eyewitness testimony should be used in court. Only confident witnesses make it to court. If a witness were to say, while viewing a line-up of suspects, "I'm not really sure this is the guy who robbed the store but it might be," the case would never go to trial if the witmess did not become more confident over time. Unfortunately, however, some witnesses who are unconfident on an initial assessment do become more confident over time and with repeated testing. This is especially so when witnesses receive confirming feedback on an erroneous choice, such as "Good. You got the right guy" (Wells & Bradfield, 1998). This post-identification feedback effect has been replicated under various conditions and may be one reason low-confidence identifications rise in confidence over time (see Wells, Memon, & Penrod, 2006, pp. 66–67). This rise in confidence over time is one reason innocent people are sometimes convicted; a witness with wobbly confidence about his or her memory in an initial interview may become certain by the time the case reaches trial (especially when the interview is suggestive; Wells & Quinlivan, 2009; Wixted, Mickes, Clark, Gronlund, & Roediger, 2014).

Even without the problem of inflated confidence with repeated testing, other issues çan arise. The relation between confidence and accuracy is never perfect and varies widely in experimental situations depending on a variety of factors. Every experiment examining confidence and accuracy has shown that even the highest confidence judgments, ones given in laboratory situations under ideal conditions, are often wrong. With simple materials like word lists

(and with short retention intervals between study and test), one often finds that when people are 100% confident in their response, they may be only 80–90% accurate on average (e.g., Mickes, Hwe, Wais, & Wixted, 2011; Odinot, Wolters, & van Koppen, 2009).

A perfect positive relation between confidence and accuracy may not be possible, but what is the general relation between the two? Psychologists have been asking this question for many years and it seems straightforward. However, this framing of the question is misleading, because it assumes that there is one answer. As we shall see, this straightforward question can lead to many answers depending on the materials being tested, the conditions of testing, the type of test used, and the type of analysis performed on the data. As in most issues in memory research, the answer is "it depends" on a host of other factors (Roediger, 2008). This chapter is devoted to understanding some of these factors.

A Bit of History

Karl Dallenbach (1913) initiated the study of the confidence–accuracy relationship and the general design of his experiment would still be worthy of study today. He specifically framed his investigation as informing eyewitness testimony and reported two experiments with similar results. We will describe the first experiment. Subjects (Dallenbach called them "observers" following the visual psychophysics tradition of the time) were shown complex pictures for one minute each and were told to remember them. They were tested in two ways at various intervals, ranging from an immediate test to other tests given five, 15, and 45 days later. All subjects were tested at all successive intervals, as is often what happens to actual witnesses. The two types of test were an open narration (with subjects remembering everything they could about the picture) and then a "deposition" in which subjects were asked 60 questions about the pictures. Subjects could choose to answer a question or respond "don't know." When subjects did provide an answer, they were told to give a confidence rating on a three-point scale: *slightly sure*, *fairly certain*, or *absolutely certain*. Dallenbach used what today would be called output-bound scoring (Koriat & Goldsmith, 1996), because subjects' answers were counted correct as a percentage of the questions attempted; that is, if subjects answered "don't know" their answer was not counted, so the proportion correct was the number questions answered correctly divided by the number of questions attempted (answered correctly or incorrectly). The case was similar for error analyses, except that the numerator was the number of items answered incorrectly (and the denominator was the same).

Dallenbach's (1913) results were rich and interesting. We provide only a few details here. First, errors were directly related to retention interval: More errors (as a proportion of total responses) occurred at 45 days than at the earlier intervals and they increased in a regular fashion, with errors growing quickly at first and

then stabilizing. This point essentially confirms Ebbinghaus's forgetting curve (1885/1913), albeit using errors. More importantly for present purposes, Dallenbach concluded that "The degree of certainty of the observer's replies bears a direct relation to the fidelity of the answer" (p. 335). For his observers, confidence and accuracy of reports from memory were positively correlated—accurate memories were confident memories and vice versa. That claim is the issue of concern in our chapter. As we shall see, the issue turns out to be a bit more complex than suggested by Dallenbach's research, pioneering though it was. Still, his broad conclusion is correct in many situations. We mostly consider exceptions here.

Confidence and Accuracy: Differing Conclusions

In the century since Dallenbach studied confidence and accuracy, research in various traditions has reached wildly different conclusions on the issue. We consider some examples and have written about this issue much more extensively elsewhere (Roediger, Wixted, & DeSoto, 2012).

One type of confidence–accuracy research—certainly the dominant one in terms of number of papers published—comes from the tradition of research using simple materials like word lists and straightforward procedures such as recognition memory with confidence judgments. For example, subjects might study 100 unrelated words sampled from a pool of 200 such words. On a test given at a later point, they would be given the 200 words one at a time (half studied, half not studied) and asked to decide whether each was old (studied) or new (nonstudied). After each recognition decision they would be asked to judge the confidence of their response (e.g., on a 1–7 scale where 1 is *guessing* and 7 is *absolutely confident*). The outcome of such experiments is not in doubt. Summarizing this literature in their book, 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). We can wonder about the "almost" in that quote qualifying a powerful adverb like "inevitably," but even in list-learning experiments it is good to hedge one's bets. After all, as noted above, in virtually all experiments even the highest confidence ratings fail to produce 100% accuracy.

In a rather different tradition inspired by issues of eyewitness testimony, researchers often examine memory for faces and assess correlations between confidence and accuracy by one of several measures. Without going into details, results from these studies have been mixed. In reviewing the literature years ago, Smith, Kassin, and Ellsworth (1989) 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). Note that this conclusion refers to two different kinds of correlations, one across people (i.e., are accurate people more confident?) and the other across various items for the same person (i.e., are

individual people more accurate when they are more confident?). The latter question is answered by a within-subjects correlation known as *resolution* in the metacognition literature (Dunlosky & Metcalfe, 2009). Researchers in the eyewitness literature have offered conclusions like this one more recently (e.g., Odinot et al., 2009), although other researchers disagree to some extent (Brewer & Wells, 2006; Juslin, Olsson, & Winman, 1996; Lindsay, Nilsen, & Read, 2000; see Wixted et al., 2014, for a review). Still, in 2007 Krug remarked that the lack of a correlation between confidence and accuracy is "one of the most consistent findings in the memory research literature" (p. 31). Obviously, such a statement ignores a huge body of work that documents strong positive correlations in many experiments (Dunlosky & Metcalfe, 2009).

To compound matters, it is also possible to find negative correlations between confidence and accuracy: In certain situations, the more confident a person is, the more likely he or she is to be in error. To illustrate this point, we reanalyzed data reported by Roediger and McDermott (1995) in the paper that established what is now called the DRM (Deese–Roediger–McDermott) paradigm. The prototype of the paradigm was Experiment 2 of that paper, but for our purposes we consider Experiment 1. Although it was really more of a pilot experiment, the procedure used confidence ratings and hence is relevant here. Undergraduate students heard six lists composed of 12 words each, and the words were the 12 most common associates to nonpresented words (like *chair*, *needle*, *mountain*, and *sleep*) that were not in the list. Thus subjects heard words like *table*, *sit*, *legs*, *seat*, *couch*, *desk*, and others. After each list they were asked to recall the words, and after all six lists they were given a recognition test in which items from the list (targets) were intermixed with others that had not been studied (lures). Some of these lure items were those that had generated the lists (*chair* and *needle*) and others were unrelated to the presented lists. Subjects provided judgments on a four-point scale from *sure old* (4) to *sure new* (1), with the intervening values being *probably old* (3) and *probably new* (2).

Subjects were both highly likely to falsely recall the critical items like *chair* from which the lists were derived (.40) and also to falsely recognize them during the recognition test. The false alarm rate (FAR) was .84 for critical lures relative to .02 for completely unrelated lures. In addition, subjects assigned related lures high confidence ratings (3.3 on the 4-point scale, compared to 3.6 for target items and 1.5 for the unrelated lures). Roediger and McDermott (1995) did not report the correlation between confidence and accuracy across their six lists, but we provide them here in Figure 22.1, which plots confidence rating against accuracy (correct rejections, or 1 – FAR) for the six lists. The overall correlation is –.68, indicating that the less accurate the subjects were for a critical lure, the more confident they were. Stated more intuitively, confidence was positively correlated with the false alarm rate. Items for which subjects were most likely to false alarm produced errors with higher confidence. Of course, in this experiment recognition followed (and hence was confounded

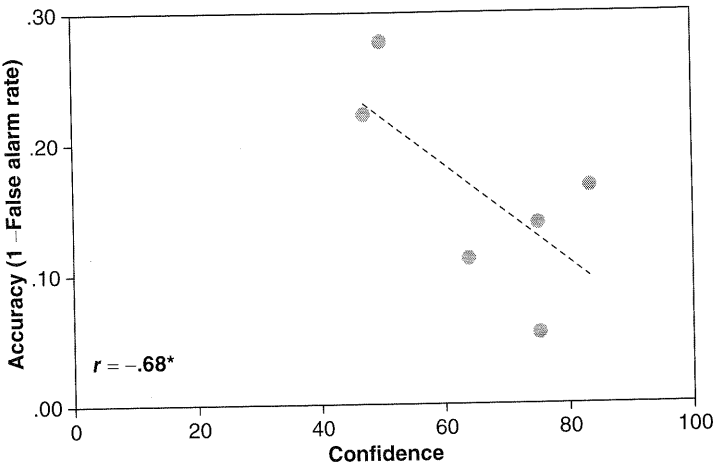


FIGURE 22.1 Scatterplot depicting the relationship between accuracy on the six critical lures in Experiment 1 of Roediger and McDermott (1995) and the mean confidence rating for those lures (i.e., the between-events correlation for critical lures). Average confidence was highest for lures with the lowest accuracy. The correlation is positive (r = .83★) when calculated for the targets in this experiment.

by) recall, but we report similar negative correlations later in the chapter that do not suffer from this problem.

Interestingly, few researchers have used confidence ratings in the DRM paradigm (but see Read, 1996), because the version adopted by most employed *remember/know* judgments (from Experiment 2 of Roediger & McDermott, 1995). Nonetheless, other groups of researchers have also reported negative correlations between confidence and accuracy in different situations, starting with Tulving (1981). For example, Brewer and colleagues (Brewer, Sampaio, & Barlow, 2005; Sampaio & Brewer, 2009) gave subjects both deceptive and nondeceptive sentences to study. The former were ones that usually led subjects to make an inference that was implied by the sentence but not necessarily correct. For example, *The baby stayed awake all night* or *The karate champion struck the cinder block* were used as deceptive sentences. Earlier research (Brewer, 1977) had shown that people were likely to recall these sentences using a different verb, showing that an inference had been made (e.g., *The baby cried all night* or *The karate champion broke the cinder block*). When such lure items were tested in a recognition memory procedure, Sampaio and Brewer (2009) showed a negative correlation between confidence and accuracy. On the other hand, nondeceptive sentences (ones that were less ambiguous) showed a positive correlation. Koriat (2008, 2012) has reported similar results with general knowledge materials. In this case, trivia

items for which most students give incorrect answers (e.g., *Sydney* as *the capital of Australia* for Americans and Europeans) also lead to a negative correlation between confidence and accuracy: The more likely people are to get an item wrong, the more confident they are that their answer is correct.

This section may leave the reader in a state of hopeless perplexity: How is one to make sense of the conflicting claims of positive, null, and negative correlations between confidence and accuracy? One tempting—but wrong—conclusion would be to attribute the differences to materials. That is, perhaps research with unrelated word lists leads to positive correlations between confidence and accuracy, research with faces and line-ups can lead to null correlations (e.g., Wells & Murray, 1984), and research with deceptive materials (DRM lists, misleading sentences, or tricky general knowledge questions) can lead to negative correlations. This glib conclusion seems unlikely to be true. We unpack the reasons in the rest of the chapter.

Measuring Confidence and Accuracy

At the beginning of the chapter we alluded to the fact that there is more than one way to measure confidence–accuracy correlations, but we have generally ignored that point in the descriptions above. Yet it is this fact that leads to some of the puzzles we have mentioned. Roediger et al. (2012) identified five different ways of assessing the confidence–accuracy relation and there may be more. For example, researchers can manipulate independent variables known to have a powerful effect on memory measures and see if the variables affect confidence in the same manner (see Busey, Tunnicliff, Loftus, & Loftus, 2000). The correlation in this type of case is, as far as we can tell, always positive: Any independent variable that affects explicit memory performance also affects confidence in the same way. We (Roediger et al., 2012) could find no exceptions to this generalization. However, let us consider three other methods of assessing confidence–accuracy correlations that are perfectly plausible ways to consider the issue:

- *Between-subjects correlations.* In this type of analysis, the researcher performs a confidence–accuracy memory experiment with many subjects and many items and analyzes the data by combining across items to get an average accuracy and average confidence score for each person. This analysis assesses whether more confident people are more accurate and vice versa.
- *Between-events correlations.* In this analysis the same sort of experiment is conducted (many subjects, many items), but now the data are combined across subjects and events are the units of analysis. Average confidence and accuracy ratings are obtained for each event and these measures are correlated. This analysis answers the question of whether confidence and accuracy are related across events of one type or another. For example, this type of analysis was used in our reanalysis of the Roediger and McDermott (1995) data in Figure 22.1, albeit only for the critical lures.

- *Within-subjects correlations (resolution).* In this type of measure, many subjects are again tested with (ideally) many items and the interest centers on how well a subject can assess his or her likely accuracy, via confidence ratings, as memory decisions are made. This measure is referred to as resolution and it is one of the most common ones used.

All three of these measures are valid methods of assessing confidence and accuracy, and others exist, too (e.g., calibration curves and within-between hybrid correlations; see Roediger et al., 2012). For present purposes, the point is that researchers often use only one of these measures and write as if it were the only way of assessing confidence–accuracy in experiments, reaching a general conclusion about confidence and accuracy relationships from just one of several measures. Outside the metacognition literature, it is relatively rare to find a report in which the authors used even two methods of measuring confidence and accuracy, much less a wider set of measures. These thoughts led us to the research project described in the final part of our chapter.

Simultaneous Assessment of Confidence–Accuracy Correlations

The variety of claims made about confidence and accuracy, reviewed earlier in this chapter, are often based on different methods of assessing the relation. We began a program of research asking if it would be possible to find different patterns of correlation in the same subjects with the same (or similar) materials only by changing the method of analysis. If so, then such experiments would help to clear up the mystery of why different groups of researchers working in various traditions of research could arrive at such different answers about confidence–accuracy correlations. We have two papers on this issue (DeSoto & Roediger, 2014; Roediger & DeSoto, 2014), and we describe one experiment from each here. Briefly, the method in these experiments was to present subjects with lists of items from common categories such as *birds* or *articles of furniture* and then test them both for studied members of the category and for other lure words taken from the same categories. This paradigm produces reasonable levels of false recognition for the nonstudied category members (e.g., Cho & Neely, 2013; Dewhurst, 2001; Meade & Roediger, 2006, 2009; Smith, Ward, Tindell, Sinfonis, & Wilkenfeld, 2000), which was essential for our purposes.

Roediger and DeSoto (2014) selected ten sets of categorized items from the Van Overschelde, Rawson, and Dunlosky (2004) norms and then took the first 25 items from each list as materials for the experiment. These lists are ordered by the frequency with which subjects produce words when they are given a category name and asked to generate items belonging to the category. So, for example, in the category of vegetables, *carrot* and *lettuce* are the first two in the norms, *cabbage* and *radish* are the 14th and 15th, and *rutabaga* and *artichoke* are 24th

and 25th. This dimension is referred to as output dominance of items in the category (or sometimes as response frequency rank).

Roediger and DeSoto (2014) performed two quite similar experiments and we describe the second one here. Following the procedure developed by Meade and Roediger (2006, 2009), subjects heard 150 words under intentional learning instructions, including 15 items from each of the 10 categories. The items were blocked by category (e.g., all vegetables occurred together); the category name was presented and then followed by the items from the category in a random order. Importantly, the presented words were items 6–20 in the norms, thus omitting the five most frequently given items (e.g., *carrot, lettuce*) and the least frequently given among the 25 (e.g., *rutabaga, artichoke*). After presentation of the 150 words, subjects were given a recognition test that contained 300 words: 150 studied words, 50 strongly related lures (the first five items from the categories), 50 weakly related lures (items 21–25 from the norms), and 50 unrelated lures (items taken from categories not used in the study materials). Subjects examined each word and provided a judgment of old (studied) or new (nonstudied). Afterwards, they gave a confidence rating on a 100-point scale from 0 (*not at all confident*) to 100 (*entirely confident*; see DeSoto, 2014).

We structured the materials in the manner described because we knew from prior work that (as in the DRM procedure) the first five items would be especially likely to lead to false alarms relative to the other type of lures (Meade & Roediger, 2006, 2009). This did indeed occur: The hit rate for studied items was .70, whereas the false alarm rate for the three types of lures was .43 for the strongly related lures (items 1–5), .28 for weakly related lures (21–25), and .08 for unrelated lures.

We also predicted that correlations between accuracy and confidence would be low (or possibly even negative) in the case of strongly related lures, because they would be "deceptive," to use Sampaio and Brewer's (2009) term. We performed three types of correlational analysis on the results: between-subjects analysis, between-events analysis, and resolution (within-subjects analysis). Pearson product–moment correlations were used for the first two types of analysis and gamma was used for resolution. The results are shown in Table 22.1 for all 300 items (as is the typical case in experimental reports) as well as for the various classes of items: targets (studied items), strongly related lures, weakly related lures, and unrelated lures. We predicted that the various measures would lead us to see positive, null, and negative correlations between confidence and accuracy across the various types of material. Even a glance at Table 22.1 shows that our expectations were upheld, with all three types of correlations appearing. We consider the various analyses in turn.

For all items, the correlation was significantly positive regardless of method of analysis, albeit rather modestly so in the case of the between-events analysis. However, examination of the rest of the columns showed that the analysis of all items

TABLE 22.1 Correlations between confidence and accuracy as a function of item type in Experiment 2 of Roediger and DeSoto (2014)

Item type	*Between subjects*	*Between events*	*Within subjects*
All items	.48★	.12★	.38★
Targets	.62★	.69★	.73★
Strongly related lures	−.18	−.34★	−.23★
Weakly related lures	.17	.14	−.02
Unrelated lures	.44★	.29★	.08

Note: Between events and between subjects correlations are computed with Pearson correlations (r). Within subjects correlations are computed with Goodman–Kruskal gamma (γ). Statistically significant differences from zero ($p < .05$) are indicated with asterisks (★).

obscured different patterns of results that appeared for the different types of items. For studied words, there was a strong positive correlation by all methods of analysis. For the strongly related lures, there was actually a negative correlation between confidence and accuracy in all three measures, although it was only statistically significant for two of them. The two significant correlations show that when examined between events (across the 50 items) subjects were fooled; the less likely they were to correctly reject the item, the more confident they were that they made a correct response (see Figure 22.2). Stated more intuitively, there was a positive correlation between the false alarm rate and the confidence in making that false alarm. The

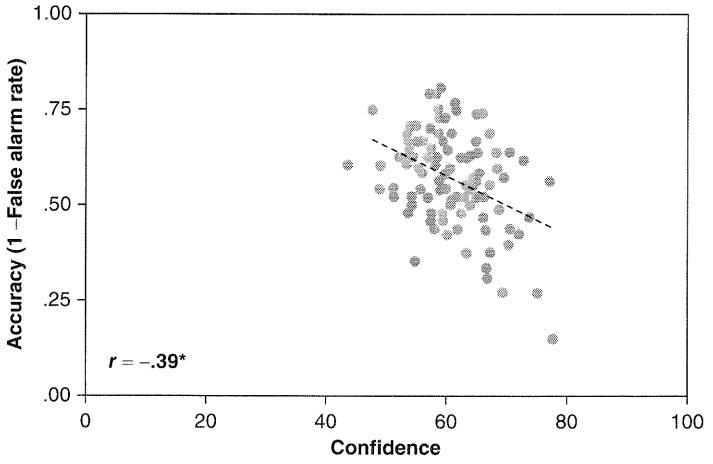


FIGURE 22.2 The between-events scatterplot for strongly related lures in Experiments 1 and 2 of Roediger and DeSoto (2014). This plot shows that those strong lure items to which subjects are more likely to correctly respond "new" (i.e., correctly reject) are also more likely to be assigned *lower* confidence ratings. If the bottom right point is removed, the correlation remains $r = -.33$★. The correlation is positive ($r = .67$★) when calculated for the targets in these experiments.

significantly negative gamma correlation for strongly related lures shows that subjects were fooled on an individual basis too. It is rare to see negative resolution in the metacognition literature. The weakly related lures showed essentially no (significant) correlation between confidence and accuracy, whereas the unrelated lures showed a positive correlation in two methods of analysis. The more likely subjects were to correctly reject them, the higher the confidence they had in doing so.

In sum, using four different types of items and three different methods of analysis, we obtained all three possible patterns of correlation in one experiment: For studied items, there was a strong positive correlation between confidence and accuracy; for strongly related lures, negative correlations between confidence and accuracy emerged; for weakly related lures, there was no significant correlation; and for unrelated lures the correlation between confidence and accuracy was positive. Given these results, it is clear that the apparently bewildering conclusions about confidence–accuracy correlations in reports from memory from prior work may not be so mysterious. Depending on the nature of materials and the nature of analysis, one may find any pattern of correlation. In the following section of this chapter, we will turn to how such a state of affairs might be understood.

In the experiment just described, Roediger and DeSoto (2014) obtained negative correlations between confidence and accuracy by omitting the five most common items in category norms and using those as lures on the recognition test. This tactic was the same as used by Brewer and his colleagues (e.g., Brewer & Sampaio, 2006), by Koriat (2008), and by others (Meade & Roediger, 2006; Roediger & McDermott, 1995). In all these cases, some items are omitted that are "deceptive" or "consensually wrong" relative to other material or are semantically related to the target material. This situation implies that it is only special materials or situations that may create confidence–accuracy inversions and perhaps these cases may be rare. If so, the correct conclusion might be that positive correlations between confidence and accuracy are the rule, and the negative correlations simply represent a weird exception with certain types of material.

In order to extend the generality of our results, we (DeSoto & Roediger, 2014) asked a different question: Could we find both positive and negative correlations with exactly the same materials, namely, a positive correlation when the materials were studied, and a negative correlation when they were not studied? We again used categorized lists but with a different twist—20 words from 12 categories (the first 20 in the norms), for 240 in all. Subjects studied lists composed of either the 10 odd items (in positions 1, 3, 5, 7, etc.) or the ten even items (2, 4, 6, 8, etc.) from the categories so that each subject studied 120 words. They then took a recognition test over the entire set of 360 words (120 studied, 120 lures, and 120 unrelated). The test set comprised the same words for all subjects, but the two groups of subjects had each studied a different half of the categorized words (odd or even words). Subjects

judged each test word to be old or new and then gave a confidence rating with a 100-point slider in the manner described above. DeSoto and Roediger reported two experiments, with the only difference being that the unrelated words were omitted from the test in Experiment 2. We report the results for Experiment 1 here.

The overall recognition results were as expected: The hit rate for studied words was .73 and the false alarm rates for related and unrelated lures were .39 and .10,

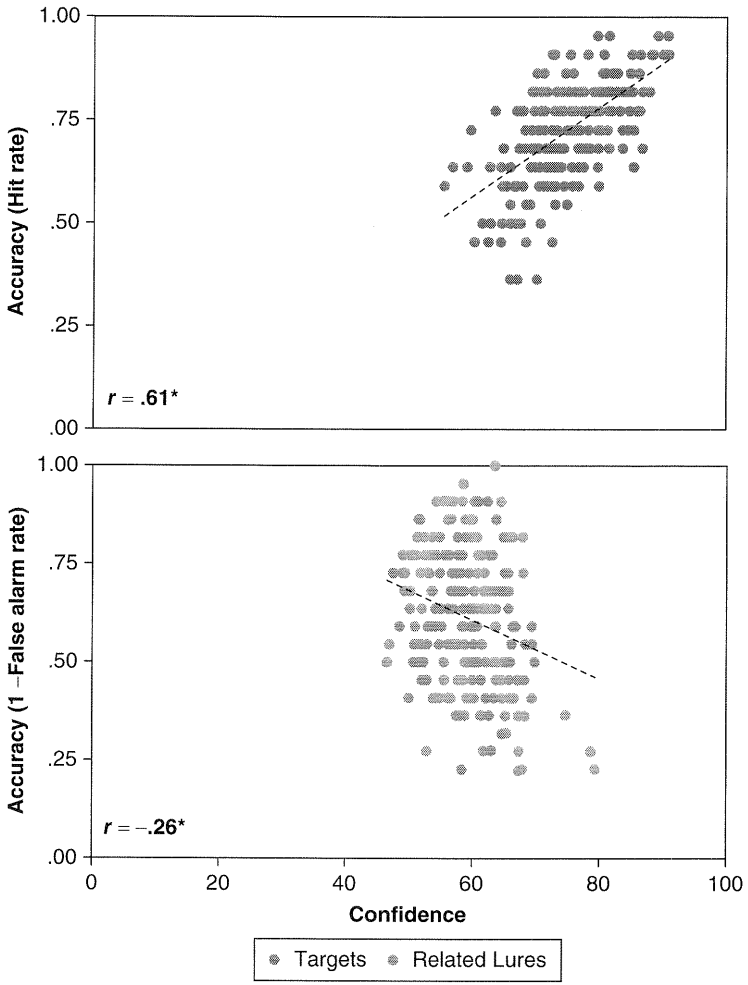


FIGURE 22.3 Between-events confidence-accuracy correlations for the same 240 category items when they were studied (targets) and unstudied (related lures) in Experiment 1 of DeSoto and Roediger (2014). Each point is an individual item. If the two bottom right points of the bottom panel are removed, the correlation remains $r = -.20^{\star}$.

TABLE 22.2 Correlations between confidence and accuracy as a function of item type in Experiment 1 of DeSoto and Roediger (2014)

Item type	*Between subjects*	*Between events*	*Within subjects*
All items	.63★	.21★	.26★
Targets	.68★	.61★	.73★
Related lures	.22	−.26★	−.21★
Unrelated lures	.54★	.38★	.16

Note: Between events and between subjects correlations are computed with Pearson correlations (r). Within subjects correlations are computed with Goodman–Kruskal gamma (γ). Statistically significant differences from zero ($p < .05$) are indicated with asterisks (★).

respectively. Of more interest are the correlations between confidence and accuracy, and these data are shown in Table 22.2 in a similar manner as used in Table 22.1. Unlike the results in Table 22.1, though, the results in Table 22.2 use exactly the same 240 items as targets and as lures (albeit across subjects). There is no special set of "deceptive" items. Nonetheless, we see the essential results from Roediger and DeSoto (2014) replicated here: Strong positive correlations appeared for all methods of analysis for studied words, whereas for the related lures two methods of analysis revealed a negative correlation (between-events and within-subjects) and the third method (between-subjects) led to a correlation not significantly different from zero. The between-events analysis reveals the different outcomes best. As shown in Figure 22.3, the same 240 words showed a strong positive correlation (.61) between confidence and accuracy when they were studied but a moderately negative correlation (−.26) when they were not studied. These between-events correlations were replicated in a second experiment in which the 120 unrelated lures were removed from the test.

The experiments reported in this section, along with the literature cited above, indicate that the relation between confidence and accuracy in reports from memory is complex. The factors we have examined here include the nature of items on the test (studied items, semantically related lures of varying strength, unrelated lures) as well as three different methods of analysis (between subjects, between events, resolution). Depending on the type of material and the method of analysis, a researcher may find positive, negative, or null correlations between confidence and accuracy. Although many prior papers and chapters have been written seeking "the" relation between confidence and accuracy, our work shows that no such single answer can be given (see Roediger et al., 2012).

We hasten to add that we have only begun to explore the issue of how confidence and accuracy vary across situations. In our research, we have only used categorized lists of words and three of the possible measures of the relation between confidence and accuracy. Thus, if anything, our work is likely to underestimate the complexities in relating confidence to accuracy. We turn next to the beginnings of a possible theoretical account of our research and related work by others.

A Framework for Understanding Confidence and Accuracy in Recognition Memory

The general expectation from theories of memory retrieval is that confidence and accuracy should be strongly correlated. This idea arises from trace strength theories of memory postulating that the experience of events leaves memory traces that vary in strength, such that stronger traces are more likely to be recalled and recognized. By this simple view, confidence is also a matter of trace strength, with stronger traces giving rise to greater confidence. Hence accuracy and confidence should be positively correlated because both are reflections of the strength of the memory traces. This idea of pure memory trace theories of memory is surprisingly resistant to both empirical and theoretical analyses showing that it is, if not wrong, woefully incomplete (see Roediger et al., 2012, pp. 88–94).

The results we have reviewed should provide a few more nails in the coffin of strength theory, as we have shown that even with the same materials (as in the DeSoto & Roediger [2014] experiments), positive, null, and negative correlations can be obtained between accuracy and confidence. Of course, for studied items, the confidence–accuracy correlation is almost always positive and relatively strong, no matter what the method of analysis. This type of outcome can be handled by most any theory of recognition memory. We consider here the other types of relation, especially those found with lures strongly related to target items. In these cases, the relation between confidence and accuracy is zero or negative, depending on the type of analysis. The case of negative correlations seems especially puzzling, but below we provide a general framework that may help in understanding how this situation arises.

Lures on a recognition test that are strongly related to studied items are often falsely recognized (e.g., Roediger & McDermott, 1995) and in some cases are given high confidence ratings and are judged to be *remembered* (using Tulving's [1985] *remember/know* paradigm). Thus to understand inversions between confidence and accuracy one must provide a general framework for understanding false memories. Tulving's (1974) analysis of remembering is helpful in this regard; he argued that remembering is always a product of information in the traces of experience left in the nervous system (the memory trace) and information provided in the retrieval environment during the act of remembering (the cues). The overlap or match of the cue/trace complex determines what is remembered (the encoding specificity principle; Tulving & Thomson, 1973). We may assume that study of a categorized list (say, of articles of furniture) leaves a combination of traces of the 10 or 15 studied words (in the two experimental paradigms we described in the prior section). Thus, following study of a group of words representing furniture, even if a cue word like *chair* given on a test had not been studied, the features extracted from *chair* when it is presented as a cue on a recognition test would greatly overlap with features in the stored traces. If so, *chair*

may be falsely judged as old, and the more features that overlap, the greater the level of false recognition and the greater the judged confidence.

This version of the encoding specificity principle can help us understand our results and also accords with much other research on recognition memory (see Tulving, 1983, Ch. 13–14). We do not mean to imply that other accounts of negative correlations between confidence and accuracy are not possible. Roediger and DeSoto (2014) suggested that their results could be understood in terms of signal detection theory (e.g., Wixted, 2007; Wixted & Stretch, 2000, among many papers) and we see nothing inconsistent between Tulving's encoding specificity principle and signal detection theory at a broad level. In fact, Wixted and his colleagues refer to distributions of items on a "strength of evidence" dimension where "strength" is conceived not as simple trace strength but more like the cue/target match that Tulving and Thomson (1973) described in their theory (see Wixted & Mickes, 2010, for an explicit formulation of just these ideas). That is, the cue/target match represents the signal of "strength of evidence" in signal detection theory.

With this translation between theories, let us reconsider the paradigm and results of Roediger and DeSoto (2014) described above. We had four types of test items in our experiments after subjects studied 15-word categorized lists (items 6–20 in the norms): Studied words, strongly related lures (items 1–5 in the norms), weakly related lures (items 21–25 in the norms) and completely unrelated lures. The probability of calling an item old varied directly with these

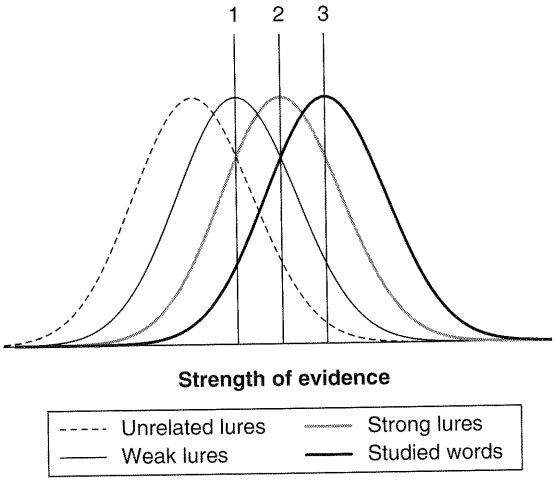


FIGURE 22.4 An illustration of how the different item types used by Roediger and DeSoto (2014) fall on a hypothetical "strength of evidence" axis. Subjects may respond with a liberal criterion (like Line 1) or more strict (conservative) criteria like 2 or 3, but the argument advanced in the text holds to a first approximation for any of the criteria.

four classes of items (albeit hits for the first category and false alarms for the other three types of items).

The most natural way to understand this pattern within signal detection theory (e.g., Wixted & Mickes, 2010; Wixted & Stretch, 2000) is to conceive of the cue/trace match providing a "strength of evidence" axis on which four distributions of items would exist (see Figure 22.4 for a depiction). Studied items are furthest to the right—having the greatest strength of evidence—followed by distributions for strongly related lures, then weakly related lures, and finally a distribution for unrelated lures. That is, rather than the case in standard signal detection models in which there are two distributions of items (for studied items and lures), to explain our results we would assume that the different types of lures represent three different distributions. The criterion for responding old would thus (wherever placed by an individual subject) provide the correct ordering of old responses due to their differences in strength (studied items > strongly related lures > weakly related lures > unrelated lures). Figure 22.4 shows three criteria that vary on a dimension of conservative (high-confidence) to liberal (low-confidence) responding. Alternatively, the criteria may refer to high-confidence (3), medium-confidence (2) and low-confidence (1) ratings.

If we assume that the strength of evidence continuum (based on cue/target overlap) gives rise to both calling an item old and to the confidence of a response, then the negative correlation for strongly related items arises naturally. For these items, the distribution of "strengths of evidence" would overlap considerably with that for studied items and many of these lures would lie to the right of the criterion. For studied items, the stronger the item, the more likely the subject is to call it old and the higher confidence the subject will ascribe to the judgment. This situation would give rise to the high positive correlation between confidence and accuracy. However, for the related lures, the same principle operates: The greater the strength of evidence (the greater the match between the recognition cue and the trace), the more likely subjects will be to provide a false alarm to the item and the greater the confidence they will exhibit in providing the false alarm. Thus, there will be a positive correlation between the tendency to false alarm and the confidence with which the false alarm is made. When one considers the same situation in terms of accuracy, however, the correlation between accuracy (correct rejections) of the items and confidence will be negative. This is exactly the case we reported in the between-events analysis, which is the most relevant one in this case. With weakly related lures, the overlap in distributions is less. The confidence–accuracy relation drops to near zero. Fewer items are falsely recognized than for the strongly related lures and the confidence expressed for these false alarms is also much less. For the unrelated items, subjects mostly correctly reject them and thus confidence tracks the correct rejections. Thus, the correlation between confidence and accuracy for these items is positive—the greater the tendency to correctly reject an item, the more confidence subjects display in their rejection.

We admit that our account here is speculative and awaits testing. Moreover, we have really only dealt with the between-events correlational analyses and not the between-subjects or within-subjects correlational analyses. Much work remains to be done.

Implications for Applied Situations

The issue of confidence and accuracy in reports from memory has played out most dramatically in the case of eyewitness testimony in legal cases. In particular, individuals have been convicted of heinous crimes largely on the basis of eyewitness testimony, but in some instances their convictions have been overturned. Since the introduction of DNA evidence in 1995, many people convicted of crimes have been exonerated based on DNA saved at the scene of the crime. In his book *Convicting the Innocent*, Brandon Garrett (2011) reviewed 161 cases in which people were exonerated after having been convicted in a court of law by eyewitness testimony. Garrett noted that the eyewitnesses were confident in their courtroom testimony that they had identified the right person as the perpetrator of the crime. This did not surprise him. However, he was surprised by another fact he uncovered: "I did not expect, however, to read testimony by witnesses at trials indicating that they earlier had trouble identifying the defendants ... Yet in 57% of these trials transcripts (92 of 161 cases) the witness reported that they had *not* been certain at the time of their earlier identification" (p. 49, emphasis in the original statement). The situation in the other 43% of the cases was indeterminate, with no information available about confidence or for some other reason, so the actual number of cases in which confidence was low in identifying a suspect might be even higher.

Garrett's (2011) finding leads to the possibility that in legal cases the witness's level of confidence in selecting a suspect when viewing a line-up for the first time shortly after a crime may provide good evidence about the culpability of the suspect. In fact, after reviewing the relevant evidence, Wixted et al. (2014) have made just this recommendation and suggested that it be given to jurors in court cases:

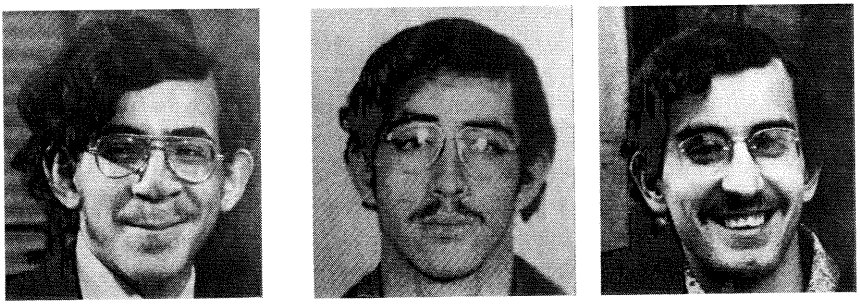
> Jurors should consider the level of certainty expressed by an eyewitness during the initial identification (at which time confidence is likely to be a reliable indicator of accuracy) while disregarding the level of certainty expressed at trial (because, by then, confidence may no longer be a reliable indicator of accuracy).
>
> (p. 5)

Wixted et al. (2014) base this advice on studies using forensically relevant materials showing a strong relationship between confidence and accuracy in calibration plots (e.g., Brewer & Wells, 2006; Juslin et al., 1996; Mickes, Flowe, & Wixted,

2012; Palmer, Brewer, Weber, & Nagesh, 2013), but Wixted and colleagues do offer some points of caution. First, all studies (whether with word lists or with forensically relevant materials like faces viewed in crime scenes) show that even subjects/witnesses who are 100% confident are never 100% accurate. As noted earlier in the chapter, the correct identification rate for subjects who are completely confident may be more like 80–90% (depending on such factors as retention interval, similarity of distractors, etc.). The results of some studies are more glum, with high-confidence responses associated with even lower accuracy (Wells et al., 2006, p. 66). Second, we believe our results reported in this chapter may reveal an important exception to the general relation between confidence and accuracy advocated by Wixted et al., viz., when the lures used in a recognition test are highly similar to (and hence confusable with) the targets. Of course, our experiments used lists of words belonging to common categories, and it would be foolhardy to generalize to applied settings such as eyewitness identification without more forensically relevant evidence. However, we will venture the hypothesis that our results may well be relevant in at least a small percentage of eyewitness cases. Consider that "the general recommendation for selecting fillers for lineups has been to use the eyewitness's description of the target and to take any additional measures needed to make sure the suspect does not stand out in the lineup" (Wells, et al., 2006, p. 62). Of course, the suspect will also fit this description, and, on occasion an innocent suspect may inadvertently resemble the perpetrator so closely in terms of other facial features as to be a virtual twin. When that happens, our findings (admittedly based on studies using lists of words) suggest that witnesses will be inclined to erroneously identify the innocent suspect with high confidence.

Of course, line-ups are composed of people fitting the general verbal description of the perpetrator for a good reason—so that a suspect will not be selected as the perpetrator simply because he is the only one in the line-up who fits the general description (Buckhout, 1974). Nonetheless, the rules for constructing line-ups make it seem possible that similarity can be an issue. Does this problem occur very often? We cannot say, but we hazard the guess that similarity close enough to represent a danger would occur only in a small proportion of cases. Retrospective studies of real line-ups in which the perpetrator was different from the suspect and was eventually arrested and convicted would be informative on this point.

Does the problem of similarity *ever* exist in line-ups? We end the chapter with a case in point illustrated in Figure 22.5, which first appeared in an article on eyewitness testimony by Buckhout (1974). The case occurred in New York City in the early 1970s. Lawrence Berson, the man on the left, was picked out of a line-up by a rape victim as the man who committed the crime. The man on the right, George Morales, was picked out of another line-up as having committed a robbery. Both these men were jailed for their crimes. However, eventually the New York City police arrested Richard Carbone, the man in the center, and he



Lawrence Berson Richard Carbone George Morales

FIGURE 22.5 This example illustrates the problem of similarity in recognition memory. Lawrence Berson was arrested for a rape and picked out by the victim in a lineup. The same thing happened to George Morales for a robbery. Later, Richard Carbone was arrested for another crime and confessed to the first two. The suspect in the lineup may be judged to be the perpetrator of the crime just because he looks like the perpetrator. From Buckhout (1974).

was eventually convicted of both crimes. The erroneous identifications in this true case are clearly ones based on close similarity. Even viewing the witnesses side-by-side, as in the figure, anyone could understand how the similarity between Berson and Morales to the actual criminal, Carbone, could have led to false identifications. To most viewers, it is also probably apparent how this judgment could have been made with high confidence—the victims did not have the advantage of the reader of viewing the suspects in the context of the actual perpetrator. This case points to the problem that similarity may play in line-up identifications, although it is impossible to know the frequency with which this problem arises in actual legal cases.

Conclusion

We have reviewed the complicated issue of confidence and accuracy in reports from memory. The evidence is vast and somewhat conflicting. Nonetheless, we hope we have brought some clarity to the issue by pointing out several factors that must be borne in mind. First, there are several ways of measuring the relation between confidence and accuracy and they need not agree with one another. In fact, in results from our experiments, the findings sometimes do not agree across various measures even with the same materials. Second, in considering the confidence–accuracy correlation even for one measure, different outcomes can arise for studied items and for various types of lures. In particular, if lures are highly similar to the target materials, the correlation between confidence and accuracy for lures may be zero or even negative. Third, this outcome

emphasizes the problematic role that similarity may play in recognition judgments or even (by extension) in eyewitness identification. The more similar lures (or innocent foils) are to the target items, the more likely false recognitions (or false identifications) are to occur.

Acknowledgment

We thank Kit Cho, Steve Lindsay, Jim Neely, and John Wixted for their helpful comments on an earlier draft of this manuscript. In addition, we thank John Wixted for providing Figure 22.4.

References

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., & Barlow, M. R. (2005). Confidence and accuracy in the recall of deceptive and nondeceptive sentences. *Journal of Memory and Language, 52*, 618–627.

Buckhout, R. (1974). Eyewitness testimony. *Scientific American, 231*, 23–31.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*, 26–48.

Cho, K. W., & Neely, J. H. (2013). Null category-length and target-lure relatedness effects in episodic recognition: A constraint on item-noise interference models. *The Quarterly Journal of Experimental Psychology, 66*, 1331–1355.

Dallenbach, K. M. (1913). The relation of memory error to time interval. *Psychological Review, 20*, 323–337.

DeSoto, K. A. (2014). Confidence ratings in cognitive psychology experiments: Investigating the relationship between confidence and accuracy in memory. In P. Brindle (Ed.), *SAGE research methods cases*. Thousand Oaks, CA: Sage Publications. doi:10.4135/978144627305013507683

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

Dewhurst, S. A. (2001). Category repetition and false recognition: Effects of instance frequency and category size. *Journal of Memory and Language, 44*, 153–167.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. New York, NY: Sage Publications.

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology*. New York, NY: Teachers College, Columbia University.

Garrett, B. F. (2011). *Convicting the innocent*. Cambridge, MA: Harvard University Press.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1304–1316.

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.

Krug, K. (2007). The relationship between confidence and accuracy: Current thoughts of the literature and a new area of research. *Applied Psychology in Criminal Justice, 3*, 7–41.

Lindsay, D. S., Nilsen, E., & Read, J. D. (2000). Witnessing-condition heterogeneity and witnesses' versus investigators' confidence in the accuracy of witnesses' identification decisions. *Law and Human Behavior, 24*, 685–697.

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., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied, 18*, 361–376.

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.

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

Palmer, M., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*, 55–71.

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.

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. (2014). Confidence in memory: Assessing positive and negative correlations. *Memory, 22*, 76–91.

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. T., & 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 judgments 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. *Journal of Applied Psychology, 74*, 356–359.

Tulving, E. (1974). Cue-dependent forgetting. *American Scientist, 82*, 74–82.

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

Tulving, E. (1983). *Elements of episodic memory*. Oxford, UK: Oxford University Press.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26*, 1–12.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 352–373.

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.

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.

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest, 7*(2), 45–75.

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

Wells, G. L., & Quinlivan, D. S. (2009). Suggestive eyewitness identification procedures and the Supreme Court's reliability test in light of eyewitness science: 30 years later. *Law and Human Behavior, 33*, 1–24.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152–176.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117*, 1025–1054.

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D. & Roediger, H. L. (2014). Initial eyewitness confidence reliably predicts eyewitness identification accuracy, Manuscript submitted for publication.

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review, 107*, 368–376.