4

# The Curious Complexity between Confidence and Accuracy in Reports from Memory

HENRY L. ROEDIGER III, JOHN H. WIXTED, AND K. ANDREW DESOTO

The relation between the probability of remembering an event and one's confidence in it seems obvious: The more confident a person is in remembering an event, the more accurate he or she will be (and vice versa). Imagine giving people a series of events to remember every day for a week, say 10 per day. The events could be sentences such as "The hippie touched the debutante in the park" or "The policeman arrested the homeless woman near the movie theater." Then, on the seventh day, people could be asked to recall (or recognize) all the sentences that had been presented that seventh day, and those from the third day of the experiment, and to rate the confidence of each reported memory. It would surprise no one to learn that people would correctly remember more sentences from the seventh day than from the third day; surely they would be more accurate for the recent memories. In addition, there is no doubt that their confidence would track their accuracy if confidence were measured on, say, a 7-point rating scale (from 7 = sure the event happened to 1 = sure the event did not happen). People would be much more confident for the recently presented sentences than for those heard 4 days previously. The reason people can intuit the result of this experiment so accurately is that we essentially live this experiment every day of our lives. We can tell an inquirer what events happened to us today with reasonable accuracy and certainty, but if we are asked to retrieve events from a particular day even a few days ago, they would be much hazier to us—we would be less accurate and less confident.

If the conclusion from this first paragraph were correct—that accuracy and confidence of retrieval were always strongly linked—then this could be a short chapter. In fact, the editors would not have asked for a chapter about this topic. However, as we shall see, the situation is much more complex—even, as the title has it, curiously so. This chapter is about that complexity and how to understand it.

The chapter is divided into several sections. In the next section, we note why the relation between accuracy and confidence in memory retrieval matters for the legal system and how the simple assumption usually made—that confidence and accuracy are always tightly linked—is wrong. In the following section, we outline a simple theory of memory that seems implicit in lay (and judicial) assumptions about remembering, but one that is at best incomplete and at worst wrong. In the next section we consider the widely varying opinions that psychologists have offered about the relation between confidence and accuracy. We also sketch out how those making strong claims about confidence and accuracy of retrieval are both partly right and partly wrong. As in most issues concerning remembering, the correct answer is "it depends" (Roediger, 2008); in this case, the relation between confidence and accuracy depends on the method of analysis, on the target material being remembered, on who is doing the remembering, and (in situations where memory is tested by recognition) on the nature of the lures and distractors. In addition, there is more than one way to measure the relationship between confidence and accuracy, and not every way is equally relevant to what courts of law would like to know about the issue.

The main part of our chapter is oriented around five different ways of analyzing the relation between confidence and accuracy of retrieval, which can lead to different conclusions depending on a host of factors. We will also consider other factors, such as individual differences among rememberers that might affect the confidence-accuracy relation. The final part of the chapter provides recommendations about confidence and accuracy that might be considered guidelines. To presage our conclusions, confidence and accuracy can be positively related, they can be unrelated, and they can even be negatively related (that is, in certain situations, factors that lead to greater numbers of errors also lead to greater confidence in those errors). Even so, it would be a mistake to conclude that confidence ratings are uninformative. But we are getting ahead of the game. First, why is this topic important?

## CONVICTION OF THE INNOCENT

In criminal courts of law, eyewitness testimony is critical in many cases. Often there is little physical evidence and the jury and judge must base their

decision about guilt or innocence on the testimony of an eyewitness to the crime. Even if we assume the best intentions of all parties involved to seek the truth—the eyewitness, the police, the prosecutors, the judge and jury—mistakes can occur. Innocent people can be convicted of crimes they did not commit (and guilty people can walk free) because of memory errors made by an eyewitness. Psychologists have argued this point for 100 years, since the pioneering work of Hugo Münsterberg in his book about the psychology of the witness (*On the Witness Stand*, 1908). In the last 40 years, since publication of the groundbreaking work of Elizabeth Loftus (e.g., 1975; Loftus & Palmer, 1974) and Robert Buckhout (1974), a huge volume of research has arisen on factors that affect eyewitness testimony and memory errors in general (see Roediger & Gallo, 2002, for a brief overview and Brainerd & Reyna, 2005, for a fuller treatment).

One critical factor that can taint testimony is information that the witness is exposed to after he or she witnesses a crime, although the same principle is true for any event and not just crimes. Information occurring after an event can supply retroactive interference (McGeoch, 1932) and disrupt retention of the original event. If erroneous information about details of the crime (or its perpetrator) is provided by other witnesses, by police, or even by erroneous recollections of the witness himself or herself, this can serve as a potent force to shape the recollection of the crime scene (or the perpetrator). This erroneous information will often be incorporated in the witness's recollections, leading the witness to confidently remember events differently from the way they happened. Loftus (e.g., 1975, 1992) has extensively studied the process by which misinformation delivered by others can be incorporated into a witness's recollection (see the chapter by Davis & Loftus in this volume, too). Further, the witness's act of recalling wrong information makes it even more likely to be misremembered in the future (Roediger, Jacoby, & McDermott, 1996). This is not the place to review the types of errors that eyewitnesses can make, but much laboratory research as well as forensic experience shows the validity of the claims. Loftus's (1996) book, *Eyewitness Testimony*, still provides a fine introduction to the basic issues involved in understanding eyewitness testimony and how it can go awry.

The point to take away for present purposes is that eyewitness testimony can be wrong—even in the absence of misleading postevent information—and yet the witness can be highly confident in her or his recollections. Thus, high confidence does not always mean that the witness is accurate, and this is a point the legal system has not adequately appreciated in years gone by. Instead, high-confidence eyewitness testimony, by virtue of the fact that it constitutes *direct* evidence (as opposed to *circumstantial* evidence), is often considered to be essentially infallible. As a result, people have been convicted

and sentenced to long jail terms based solely on high-confidence eyewitness testimony.

Of course, in thousands of cases decided on the basis of eyewitness testimony there is no way to tell if an error has been made. A suspect who actually committed a crime but who has been freed because an eyewitness confidently reported that a lineup did not contain the perpetrator is unlikely to later reveal the truth of the matter, so this kind of error (a miss) is unlikely to ever be detected unless other evidence comes to light to implicate the person. Similarly, an individual who is convicted on the basis of high-confidence eyewitness testimony may naturally protest his or her innocence, but how can the legal system know what the truth is? Occasionally, some other person may eventually be caught (usually for an unrelated crime) and then confess to the crime for which another person had been wrongfully convicted. Such events are rare. However, since the late 1980s, another source of information has come into play: DNA evidence. Scientists in the United Kingdom perfected technologies that make *DNA fingerprinting* (as it is often called) highly reliable (except for identical twins, since they share 100% of their DNA). Because DNA is associated with at least some crimes for which people have been convicted (e.g., rapes, some murders), if a court permits a test of the convicted person's DNA, it can be matched against the DNA left by the perpetrator at the crime scene. If the evidence shows a mismatch, then the person who has been convicted of the crime almost certainly did not commit it. However, even the legal process leading to testing of DNA is often fraught with difficulty (that is, often there is a legal battle over retrospective testing of a convicted person's DNA). Another difficulty is that often DNA evidence is disposed of after a conviction. Nonetheless, sometimes DNA testing is permitted and sometimes the conviction of a person for a crime is shown to have been in error.

The Innocence Project (affiliated with the Benjamin N. Cardozo School of Law at Yeshiva University) was founded by Barry Scheck and Peter Neufeld to help convicted prisoners seeking to establish their innocence from crimes through DNA testing and other means. As of this writing (April 2012), 289 people have been exonerated by the Innocence Project, many by DNA testing, including 17 who had spent time on death row. These innocent people served an average of 13 years in prison before their convictions were overturned. Brandon Garrett (2011) examined the first 250 DNA exonerations in his book, *Convicting the Innocent*, and concluded that 190 of the convictions (76%) were the result of eyewitness misidentification. (Other causes include improper forensics, false or coerced confessions, and use of informants who gave wrong testimony.)

Most cases of eyewitness identification come from people who are highly confident and believe they are correctly identifying the right person.

Identifications made with low confidence generally never make it to a court of law and are given little weight if they do. (If a witness said "that might be the man who robbed me, but it might not be. I'm just not sure" the case would never go to court.) Thus, if there was ever any doubt that high-confidence eyewitness errors can occur, such DNA exonerations establish beyond any reasonable doubt that they do. This fact, perhaps more than any other, has contributed to the impression that the relationship between confidence and accuracy is hopelessly weak. As we shall see, however, despite the occurrence of an uncomfortably large number of high-confidence errors in eyewitness testimony, the relationship between confidence and accuracy is not always so poor. The issue of confidence and memory is complex, and in some situations the relation can be quite high.

As indicated above, eyewitness errors made with high confidence have led to the conviction of over 200 innocent people. Of course, 200 erroneous convictions over many years out of the thousands of people convicted each year may not seem great. However, the danger is that this number represents the tip of the iceberg, a small proportion of the people who have been wrongly convicted. Wrongful convictions are obviously of paramount concern even if it could also be true that a much larger number of criminals have been rightly convicted (and innocent suspects rightly exonerated) on the basis of eyewitness testimony made with high confidence.

DNA evidence does not exist in most criminal cases, and exonerations are extremely difficult without such evidence (although some do occur). According to The Innocence Project website (http://www.innocenceproject.org), "Those exonerated by DNA testing aren't the only people who have been wrongly convicted in recent decades. For every case that involves DNA, there are thousands that do not." Although it is hard to defend the last claim rigorously, the point here is that every precaution should be taken beforehand to limit erroneous eyewitness testimony and to understand situations in which it may arise. As the English jurist William Blackstone argued, it is "better that ten guilty persons escape than that one innocent suffer." In order to satisfy this maxim, and due to the fact that so many have been wrongly convicted, we must ask: Under what conditions are eyewitnesses likely to make high-confidence errors? The aim of the remainder of this chapter is to provide answers to this question.

## TRACE THEORIES OF REMEMBERING

Both Plato and Aristotle used an analogy of traces left on memory to impressions created in a wax tablet. It is worthwhile to quote a few lines here from

Plato's dialog *Theaetetus* because, we will argue, the same assumptions used by Socrates (or Plato) in the dialog are still surprisingly common today:

> Imagine, then, for the sake of argument, that our minds contain a brick of wax, which in this or that individual may be larger or smaller, and composed of wax that is…harder in some, softer in others, and sometimes of just the right consistency.…Let us call it…Memory, and say that whenever we wish to remember something we hear or conceive in our own minds, we hold this wax under the perceptions or ideas and imprint on it as we might stamp the impression of a seal ring. What is so imprinted we remember and know so long as the image remains; what is rubbed out or has not succeeded in leaving an impression we have forgotten and do not know. (Translated by Hamilton, 1961, p. 897)

The metaphor was continued in *Theaetetus* in other interesting ways (Roediger, 1980).

No one today believes that memory works like the imprint of a seal on wax, but one dominant class of theories, called *trace-dependent theories* (Tulving, 1974) or *trace access theory* (King, Zechmeister, & Shaughnessy, 1980), is very much like Aristotle's and Plato's conception. These trace theory ideas are still found today in some accounts (perhaps especially in neurobiological theories) of memory. The basic ideas are straightforward: First, events and experiences change the nervous system, and these changes are referred to as the creation of memory traces or *engrams*. (Every theory of memory has some version of this assumption.) Second, memory traces vary in strength from weak to strong. In terms of Plato's metaphor, some impressions are deep and some are shallow. In terms of one popular modern theory, the *levels of processing framework* (Craik & Lockhart, 1972), processing of information can be either shallow (or superficial) or deep (involving meaning), leaving traces that, like Plato's, are more or less robust. If traces vary in strength, such variations determine accuracy. Memories with stronger traces are more likely to support later memory performance (recall, recognition, transfer) than are weaker traces. The theory can further assume that trace strength determines confidence—a person will be more confident in a memory underlain by a strong trace than one that arises from a weak trace. Thus, the trace strength account of the relation between accuracy and confidence is neat and tidy; according to this theory, both accuracy and confidence are supported by the same underlying entity, the strength of memory traces (and nothing more).

The trace theory of memory is straightforward and intuitive and accords well with common sense. Insofar as laypeople think about memory at all, it

is probably their theory, because we use the language of strength theory in speaking of memory. If I say I have a strong memory of the basketball game last week, everyone knows what I mean. One reason people generally believe that accuracy and confidence are tightly linked is probably that they subscribe to some version of trace theory.

The problem is that trace theory is wrong—not completely wrong, but still wrong. At the very least, the theory is incomplete in postulating that memory performance is totally determined by the strength of memory traces. The problem is that trace theory—the idea that remembering involves a direct readout from memory traces—ignores retrieval processes, and much evidence indicates that remembering is a cue-dependent process (Tulving, 1974). Toward the end of his great book *Remembering*, published 80 years ago, Sir Frederic Bartlett (1932) wrote:

> If there be one thing upon which I have insisted more than another throughout all the discussions of this book, it is that the description of memories as 'fixed and lifeless' [traces] is merely an unpleasant fiction. That views implying this are still very common is evidence of the astonishing way in which many psychologists, even the most deservedly eminent, often appear to decide what are the characteristic marks of the process they set out to study, before they ever begin to actually study it. (pp. 311–312)

What is missing from trace theory? Quite a bit, actually. For now, let us be content to fill out the point above about the critical omission of retrieval processes. The weakness or strength of memory traces is just one factor determining the memorability of an event. Another important set of processes concern retrieval, because the same trace may or may not eventuate in successful remembering, depending on many other factors occurring during access of stored information. The nature of cues the rememberer has (or can generate) when trying to remember is also crucial, as are the processes through which these cues are used. The *mental set* or instructions about retrieval with particular cues and traces also matter. For example, you may see the word *lamp* and no particular memory may come to mind. However, if you are told, "Recall an experience from your past involving a lamp," a specific episode may come rushing back. One key issue concerns how strong or vivid the experienced sense of remembering is once a memory is retrieved. The experience is partly determined by the cues in the retrieval environment, partly by the trace, partly by the interaction of these two factors, and partly by other considerations such as the instructions given for retrieval (the mental set or the retrieval mode; Tulving, 1974, 1983; Wixted & Mickes, 2010). Such retrieval experience probably gives rise to the sense of confidence that people have in

their specific memories. The strength of the trace or engram (which, of course, can never be measured—it is a hypothetical concept for psychologists) is just one component in the process of remembering. That, in a nutshell, is why trace theories are incomplete or wrong—they leave out other factors that determine remembering, especially ones arising during retrieval.

Whether or not an event is retrieved from memory is determined powerfully by the nature of retrieval cues used to prompt the memory. Suppose I want people to retrieve a specific word from the English language, one they usually do not use much, like *ghost*. The strength of the word's representation (trace) may not help. In an experiment, Rubin and Wallace (1989) gave one group the cue "a mythical being" and no one retrieved *ghost*. They gave another group the cue "the word ends in *ost*," and again, no one recalled the word. However, when they gave a third group the cue "a word naming a mythical being that ends in *ost*," 100% of the people were able to produce *ghost*. Two cues that were individually ineffective produced perfect retrieval when used together, all with the same trace.

The power of cues also matters in remembering events from one's life. Traces of experience that are inaccessible with one type of cue may be easily retrieved with another type of cue (e.g., Barclay, Bransford, Franks, McCarrell, & Nitsch, 1974; Roediger & Payne, 1982; Tulving & Pearlstone, 1966). As Tulving (1974) has written:

> Memory for an event is always a product of information from two sources. The first is conceptualized as the memory trace—information laid down and retained in a person's memory store as a result of the original perception of the event. Its postulation is necessary to account for the residual effects of the event. The other source is the retrieval cue—information that is present in the individual's cognitive environment at the time retrieval occurs. (p. 74)

Thus, to the extent that the concept of memory strength is applicable to the understanding of the relationship between confidence and accuracy, it is the strength of memories *as retrieved*, not the strength of memories as encoded (i.e., not the strength of the memory trace), that is critical. Because remembering is reconstructive (Bartlett, 1932), retrieval processes are crucial: We usually take the traces of experience and weave them together into a more or less coherent description of a remembered event, a description that depends heavily on the cues used during retrieval. We shall have more to say on this topic later in the chapter, but now we turn to a survey of opinions psychologists have provided about the relation between confidence and accuracy of memories.

## CONFLICTING CLAIMS OF PSYCHOLOGISTS

Psychologists have issued a variety of pronouncements about the relation between confidence and accuracy of memory reports. Cognitive psychologists tend to perform experiments in which a list of unrelated words is presented for study and then twice as many are presented during the test (say, 100 studied and 200 tested). Subjects in the experiments are asked to judge whether each tested item is old (studied) or new (nonstudied) and to rate their confidence on a straightforward scale (say, 1–7, with 7 being most confident). In considering research mostly of this kind, 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. Low ratings correlate very well with the item being new" (p. 176). In commenting on others' research, Wixted and Mickes (2010, p. 1030) remarked that confidence "is a useful proxy for memory strength" (with "memory strength" construed as the strength of a retrieved memory, not the strength of an encoded trace). The authors cited here have a powerful ally: The U.S. Supreme Court ruled in the case of *Neil v. Biggers* (1972) that highly confident eyewitness identifications (ones that meet certain criteria) are likely to be accurate, although not all outside observers were convinced of the Court's argument (see Wells & Murray, 1983).

In striking contrast, researchers from a different tradition of research (mostly investigating memory for faces in eyewitness situations) have sometimes reached a quite different conclusion. Surveying the evidence in 1989, Smith, Kassin, and Ellsworth concluded that "confidence is neither a useful predictor of the accuracy of a particular witness or of the accuracy of particular statements made by the same witness" (p. 358). Kassin, Ellsworth, and Smith (1989) surveyed forensic psychologists and reported that 80% of respondents believed that confidence and accuracy were actually unrelated. Similarly, a 1995 article in *The New York Times* that covered research on eyewitness identification arrived at the same conclusion: "there is little or no relationship between the accuracy of the witness identification and his or her confidence in it" (January 17, 1995, cited by Juslin, Olsson, & Winman, 1996, p. 1304). More recently, Odinot, Wolters, and van Koppen (2009) argued that the relationship between confidence and accuracy is so weak that confidence ratings "should never be allowed as evidence for memory accuracy in the courtroom" (p. 513).

The statements in the first paragraph of this section arguing that confidence and accuracy are highly related came from cognitive psychologists surveying their type of research (as well as from Supreme Court justices). The statements in the second paragraph were derived from social and applied (forensic) psychologists examining research in a different tradition. What are we to make

of these conflicting statements? We will argue that both groups have a legitimate point to make with respect to how expressions of confidence should be interpreted in the legal system. However, we shall argue that the conclusions cited above by social and forensic psychologists about confidence never being related to accuracy are far too strong. Often, as cognitive psychologists maintain, confidence and accuracy are positively correlated. Nevertheless, this does not change the fact that confidence is clearly malleable or the fact that high-confidence errors occur considerably more often than was once believed (as long argued by eyewitness memory researchers). The aim of this chapter is to lead toward a more nuanced view of the relation between confidence and accuracy.

Complicating any inquiry into this issue is the fact that the relation between confidence and accuracy can be measured in very different ways. Indeed, different analyses answer different questions about that relationship. Here are some questions that have been asked: Are experimental conditions that are associated with high accuracy also associated with higher confidence compared to conditions that are associated with low accuracy? Are people who are more confident also more accurate than people who are less confident? When individuals express high confidence, are they usually more accurate than when they express low confidence? Because these are different questions, they need not have the same answers. For that reason alone, one cannot make blanket statements about the way in which confidence and accuracy are related in memory reports.

Actually, the situation is even more complex than the preceding paragraphs indicate. Even when the question asked is held constant (e.g., Are experimental conditions that are associated with high accuracy also associated with higher confidence compared to conditions that are associated with low accuracy?), there are different ways of computing the statistic of interest. The different computational methods can yield wildly different answers, and this has contributed to the impression that there is widespread disagreement about the relation between confidence and accuracy. However, it turns out that some computational methods that have been influential in the debate over the confidence-accuracy relationship are not as relevant as was once thought, and those methods have been largely replaced by newer and more useful methods that yield a different answer.

Finally, even when the question that is asked and the computational method of analyzing the relationship between confidence and accuracy are both held constant, it is possible to find data showing both positive correlations and zero correlations between confidence and accuracy, and several experiments have even shown that negative correlations can exist. That is, one can find conditions in which the more errors people make, on average, the more confident they are, on average, in those errors (Brewer & Sampaio, 2005; DeSoto &

Roediger, 2011; Roediger & DeSoto, 2011; Sampaio & Brewer, 2009). By the end of the chapter, we will have explained how all these relationships are possible and try to make sense of them.

## ANALYZING THE RELATION BETWEEN ACCURACY AND CONFIDENCE IN MEMORY: FIVE ANALYSES

Psychologists often ask for confidence judgments in studies of perceiving, remembering, decision making, and social behavior. Subjects in experiments readily supply such judgments, but their basis is not well understood. Still, it seems a natural judgment to make. For the purposes of this chapter, we consider only what are called *retrospective confidence judgments*; that is, after events have happened and a person is given some form of test, what is the confidence that the answer provided is correct? Confidence can be measured on various sorts of scales: 1–4, 1–7, 1–20, or even 1–100. In calibration studies, the 1–100 confidence scale has a specific meaning: Subjects are instructed to give a confidence rating of X when they believe that their chances of being correct are X%. Although the measurement scale may have subtle effects on judgments, we think this factor does not play a big role for our points below.

In an excellent paper, Busey, Tunnicliff, Loftus, and Loftus (2000) outlined three different ways of analyzing the confidence-accuracy relation in recognition memory. We partly use their framework here, although we extend it to recall and we describe one type of analysis that Busey et al. did not mention (and we discuss a variant of two methods that they did describe). As we shall see, this analysis is critical to understanding zero and negative correlations between confidence and accuracy. The five methods address different questions about the relationship between confidence and accuracy. All of these methods are relevant to psychological theory, but some methods have more direct implications for the legal setting than other methods. Here are the five methods in brief; we consider each one at length in succeeding sections.

1. *Manipulating an independent variable*: An independent variable is manipulated in an experiment (e.g., the retention interval is 1 day or 1 week), and measures of both confidence and accuracy of memory reports are obtained in the various conditions. The question of interest is whether average confidence and average accuracy are correlated across conditions (e.g., are confidence and accuracy both higher, on average, in the 1-day condition compared to the 1-week condition?).

2. *Between-events correlations*: A second type of analysis one can perform (one not described by Busey et al., 2000) is to make the events or items to be remembered the unit of analysis. That is, if people study 100 pictures or

100 words or 100 faces, then a researcher can average across people to determine the accuracy and the confidence with which the events are recollected. Are there some sets of events for which average confidence and average accuracy are highly correlated (say, words) and other sets for which the correlation is zero or even negative (say, faces)?

3. *Between-subjects correlations*: In this case, for each participant, confidence and accuracy are averaged over the events studied. The question then asked of the data is: Are people who are more accurate in their recollections, on average, also more confident in their recollections, on average? Using this method, we can also compare groups: Are confident children no more accurate in their memories than other children, on average, whereas average confidence is a reliable predictor of average accuracy in young adults?

4. *Within-subjects correlations*: In this method, individual subjects are exposed to materials for later recognition or recall. On the test, they are asked to recall or recognize the target items and to give a confidence judgment for each one. The question of interest here is whether items on which subjects are more confident are also items on which they are more accurate (compared to items on which they express low confidence). That is, this analysis asks whether an individual who expresses high confidence in a decision is more likely to be correct than when that same individual expresses low confidence in a decision. This type of analysis can be performed when participants recall or recognize multiple items (e.g., from a list) or multiple details (when eyewitnesses are asked about many different aspects of a crime).

5. *Within/between hybrid analysis*: The last method is the one that is most commonly used in applied research investigating the relationship between confidence and accuracy. It involves elements of both the between-subjects analysis and the within-subjects analysis. In this method, each subject watches a single event (e.g., a video of an individual committing a crime). Later, memory for the culprit is tested (e.g., using a photo lineup), and a confidence rating is taken. The data for each subject consist of one particular accuracy score (with "incorrect" and "correct" coded as 0 or 1) and one particular confidence rating. Under these conditions, the confidence rating could be a function of both (a) the subject's general inclination to express high confidence (an individual difference, between-subjects variable) and (b) the clarity of the subject's memory for that particular episode (a within-subjects variable because the subject would have made a lower confidence rating if, for example, he or she had paid less attention to the video and had a less clear memory of it). Both play a role because neither source of variance has been averaged out.

We orient the remainder of our review of the literature around these five types of analyses. All five are perfectly legitimate ways of asking about the

relation between confidence and accuracy, and they need not (and, in fact, do not) lead to the same answer in all cases. However, the way in which the different answers inform the legal system is not entirely straightforward. As a first step, computing a correlation coefficient between confidence and accuracy is a useful and informative technique, though it can also be misleading (as we shall see later in the chapter). Moreover, regardless of what a correlation coefficient suggests, a particularly informative way of analyzing the relationship for its implications with respect to the legal system does not involve computing a correlation coefficient at all. Instead, it involves plotting the accuracy associated with each level of confidence. Courts of law mainly need to know if accuracy is higher for eyewitnesses who express high confidence compared to eyewitnesses who express low confidence (or, when a witness provides multiple details, whether the high-confidence recollections of the witness should be given more weight than the low-confidence recollections of that same witness).

The most straightforward way to address this question about confidence and accuracy in a research study is to simply compute the probability of making a correct decision for each level of confidence. For example, imagine that confidence ratings in recognition decisions were taken using a 5-point scale (1 = guessing, 5 = certain). For each subject, an accuracy score would be computed for each level of confidence. For example, for all ratings of 5—the highest possible confidence—the percent correct score would equal [correct 5s/(correct 5s + incorrect 5s)] × 100% to answer the question "What percentage of items given the highest confidence rating are actually correct?" The same approach would be used to compute accuracy scores for the remaining levels of confidence. The confidence-accuracy scores for the five levels of confidence would then be averaged across subjects and plotted. A weak relation between confidence and accuracy would be indicated if accuracy for ratings of 1 were similar to accuracy for ratings of 5. A strong relationship would be implied if the accuracy associated with ratings of 5 far exceeds the accuracy associated with ratings of 1. The strength of the relation can, of course, be captured in a correlation coefficient, such as the gamma statistic, and it is important to consider which approach is more informative (e.g., for the legal system). For example, one question that is quite informative is the percentage correct of judgments given the highest level of confidence. If this is, say, 75%, that would mean that 25% of the time the subject gave the highest confidence rating, he or she was wrong. So, even if the overall correlation between confidence and accuracy is high (say, +.80), the most highly confident cases could still often be in error. This fact would still be quite troubling and, in the case of eyewitness confidence in legal settings, could still lead to many wrongful convictions.

In many studies, called *calibration experiments*, the confidence rating scale that is used for each recognition decision is meant to provide more than just an ordinal scale (which is all that the 1–5 scale mentioned above provides). In these studies, subjects are asked to give a confidence rating of X when they believe that their chances of being correct are X%. Thus, a confidence rating of 80% means that the subject believes that across all items given this rating, 80% of the decisions will be correct. As when a 1–5 scale is used, accuracy can be plotted for each level of confidence, but in this new analysis the confidence and accuracy scores are meaningfully related. This kind of plot is known as a *calibration plot*, and it allows for further considerations (e.g., one can see from such a plot whether subjects are generally overconfident or underconfident). But the key point for present purposes is that this kind of analysis (whether on a 1–5 scale or a 1–100 scale) is probably most pertinent to the legal system. Sauer, Brewer, Zweck, and Weber (2010) put it this way:

> The forensic utility of the calibration approach, when compared to correlation, lies in its indication of probable accuracy for each level of confidence. As Juslin et al. (1996) note, the knowledge that the CA [confidence-accuracy] correlation is, for example, .28 does not help assess the accuracy of an individual identification made with 80% confidence. On the other hand, knowing that 80% (or 70, or 90%) of identifications made with 80% confidence are correct provides a guide for assessing the likely reliability of an individual identification decision. (p. 338)

As noted above, the correlation could be very high between confidence and accuracy, and the correctness of the highest confidence accuracy scores might still not be that great if the task is difficult (e.g., 75% correct for the highest level of confidence).

In what follows, we refer to any plot showing the accuracy associated with each level of confidence as a *calibration plot* whether the confidence scale is an ordinal scale (e.g., 1–5) or a probability scale (though, technically, only the latter is a true calibration plot).

## Manipulation of Independent Variables

This tactic is the grist for the experimental cognitive psychologist's mill: Select an independent variable that is known (or strongly suspected) to increase accuracy on some measure of memory, manipulate it, and see (1) if the variable does affect the memory measure and (2) whether confidence is affected in a similar manner as the accuracy measure. Many experiments have tried this tactic. A related strategy is to manipulate a variable that is thought to

affect confidence in memory judgments and see if accuracy is also affected. We consider each in turn.

Busey et al. (2000) reported three experiments using the first strategy. They had subjects study 30 faces (pictures of bald men, some with facial hair and some without), and they manipulated several variables. In Experiment 1 the faces were presented three times for varying amounts of time (ranging from 230 to 930 milliseconds), and after each picture was presented, 15 seconds were permitted for possible rehearsal or, in a different condition, for math problems (to block rehearsal of those items). Experiments 2 and 3 were similar, except that instead of manipulating presentation duration of the faces, Busey et al. manipulated their luminance (from low to high, or relatively dark to relatively bright). Considering analogs to these variables outside the lab, as in witnessing the perpetrator of a crime, the variables correspond to how long a look the witness had, how bright the scene was, and whether the witness could reflect on (rehearse) the face or was distracted by something else (math problems in the experiment). After viewing the 30 faces, subjects took a test on 60 faces, the 30 previously viewed ones and 30 similar distractors, presented in a random order. They judged each face to be old or new (studied or not studied in the previous phase), and then they gave a confidence judgment on a 5-point scale (from "not at all certain" to "100% certain"). We are considering only part of their experiments here, the part concerned with retrospective confidence judgments (the focus of our chapter). The full experiments are more complex than the portions of interest for present purposes.

The results across the three experiments were highly consistent. Duration of presentation of the faces, luminance, and whether or not time was given for rehearsal of the faces all affected both recognition accuracy and confidence in the same way. We need not go through the details of the results, because the retrospective confidence judgments always followed accuracy. The two measures were tightly bound. The bottom line of the story from Busey et al. is that when an independent variable is manipulated that affects accuracy of memory reports, confidence comes along for the ride. More generally, confidence and accuracy seem well correlated in this kind of experiment in which independent variables are manipulated. In fact, the exceptions are sufficiently few that we can safely conclude that when an independent variable affects accuracy of memory reports, subjects' confidence in those reports will virtually always be affected the same way (however, see Tulving, 1981, for a somewhat different case).

But what about the other type of manipulation described above? If a variable is manipulated that is expected to affect confidence, will it always affect accuracy? The answer here is "no," or at least "not always," because it is possible to manipulate a person's retrospective confidence without influencing his

or her accuracy. In the cognitive psychology lab, this is commonly done in the context of Receiver Operating Characteristic (ROC) analysis, in which the hit rate is plotted on the ordinate against the false alarm rate on the abscissa. For example, in a standard list-learning paradigm, Mickes, Hwe, Wais, and Wixted (2011) provided error feedback to subjects who supplied confidence ratings for their old/new recognition decisions using a 20-point confidence scale. In response, the subjects became more cautious about supplying ratings of high confidence on both ends of the confidence scale (i.e., they produced fewer high-confidence old judgments at or near 20 and fewer high-confidence new judgments around 1). The data are shown in Figure 4.1 for the subjects who did not receive feedback and for those who did (and became less likely to give extreme ratings). Each point on the ROC represents the hit rate and the false alarm rate associated with a particular level of confidence. The lower left point represents the hit and false alarm rates associated with the highest confidence rating of 20—the false alarm rate is quite low. Moving up and to the right, the next point represents the hit and false alarm rates associated with confidence ratings of 19 or 20, and so on. As shown in Figure 4.1, overall accuracy remained unchanged between the two conditions; that is, the ROC curve did not move further from the diagonal when subjects became more cautious

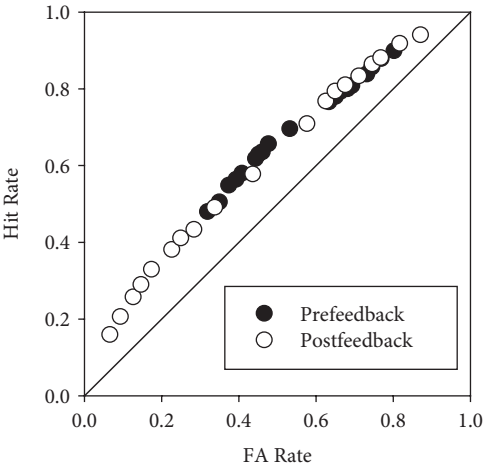



**Figure 4.1** In an experiment, subjects made recognition memory judgments and gave confidence ratings. They either did or did not receive feedback. Subjects tended to use the entire range of the confidence scale (open circles) when they did not get feedback. However, with feedback, they were much less likely to use the whole range of the scale. Nonetheless, all the points fall on the same ROC curve, showing that accuracy did not differ between the two conditions despite the variations in confidence. Data are from Mickes, Hwe, Wais, and Wixted (2011).

about using the extreme ratings (e.g., for confidence ratings of 20, both the hit rate and the false alarm rate are lower postfeedback compared to prefeedback). Instead, the points shifted along the same ROC curve after error feedback was provided. Thus, confidence was affected but accuracy (measured by d′ in signal detection theory or the distance of the ROC curve from the diagonal) was not in this experiment.

Using an eyewitness memory paradigm in which subjects saw a crime scene, Shaw (1996) had subjects give answers to forced-choice questions during a test of their memories for target items from the scene. That means they had to respond, even if they knew they were guessing. Later, they were exposed to some of their forced-choice answers via questioning. Shaw found that such exposure greatly inflated subjects' confidence in their answers but had no effect on their accuracy. In a related study, Shaw and McClure (1996) showed that repeated questioning influenced witnesses' confidence without increasing their accuracy. This effect was also replicated by Odinot et al. (2009; but see Ebbesen & Rienick, 1998, and a review by Knutsson, Allwood, & Johansson, 2011, that shows that the issue is still in some doubt).

Similarly, an experiment by Wells and Bradfield (1999) revealed that merely introspecting on one's own confidence affects confidence without affecting accuracy. In this study, subjects watched actual security camera footage of a gunman who shot a security guard while off camera. After watching the footage, subjects were asked to identify the gunman from a spread of five photos. Unknown to the subjects, the gunman never appeared in the lineup, yet all 156 subjects in the experiment made a (false) identification.

Two of five experimental conditions are relevant to the present discussion. In one condition, subjects waited 6 minutes after making their identification before they rated their confidence in their identification. In a second condition, an experimenter gave the subjects written instructions prompting them to consider how sure they were that they identified the right person in the photo spread. These subjects waited 6 minutes and were then asked to rate their confidence (on a 100-point scale) in the same fashion as in the first condition.

This study revealed a striking finding: Subjects in the first condition were roughly 50% confident that their identifications were accurate, but subjects who were asked to consider their own confidence were, on average, 70% confident in their false identifications. Wells and Bradfield (1999) dubbed this finding the "thought-alone effect," concluding that "merely thinking about one's confidence, view, and so on, itself seems to produce confidence inflation" (p. 142).

Assuming that Shaw and McClure (1996) and Wells and Bradfield (1999) are correct, then unlike manipulations that affect accuracy and then induce a correlation with confidence, one can manipulate confidence without accuracy

showing a corresponding increase. Obviously, to the extent that this latter finding occurs, problems can be created for the criminal justice system. Witnesses who are frequently questioned about a point may become increasingly confident in their answer, even if they are wrong (Shaw, 1996). On the positive side, it might be possible to train witnesses to be more cautious about making high-confidence judgments, thereby increasing their high-confidence accuracy and decreasing their high-confidence false alarm rate (as Mickes et al., 2011, found in the cognitive laboratory). However, in most legal settings, such a suggestion is impractical. People do not practice being witnesses to a crime; by definition, they become witnesses unexpectedly.

To further inform the legal system, a useful issue to consider is what effect postevent questioning and other experimental manipulations (such as retention interval) have on the accuracy associated with different levels of confidence (e.g., using a calibration approach). Counterintuitively, it is possible that the accuracy associated with each level of confidence could decrease even if, as Shaw (1996) observed, the average level of accuracy remained unchanged and average confidence increased with postevent questioning. This could happen if correct and incorrect decisions that were made with low confidence in one condition were made with high confidence in a different condition. Overall accuracy would remain the same if the number of correct and incorrect decisions did not change, but average confidence would increase because more decisions were made with high confidence. At the same time, accuracy for high-confidence decisions could selectively decrease. For example, using a 2-point confidence scale (1 = low, 2 = high), imagine that a subject in condition A made 20 high-confidence decisions, all of which were correct (high-confidence accuracy = 1.0) and 20 low-confidence decisions, 10 of which were correct and 10 of which were incorrect (low-confidence accuracy = 0.50). In this condition, overall accuracy would be 0.75 because 30 out of 40 decisions were correct, and average confidence would be 1.5 because 20 decisions were made with high confidence (2) and 20 decisions were made with low confidence (1). In condition B, suppose this same subject now made 30 high-confidence decisions, 25 of which were correct and 5 of which were incorrect (high-confidence accuracy = 0.83) and also made 10 low-confidence decisions, 5 of which were correct and 5 of which were incorrect (low-confidence accuracy = 0.50). Compared to performance in condition A, high-confidence accuracy selectively decreased in condition B (from 1.0 in A to .83 in B). Even so, overall accuracy remained unchanged at 0.75 (because, again, 30 out of 40 decisions were correct), and average confidence increased to 1.75 (because 30 decisions were made with a high-confidence rating of 2 and 10 were made with a low-confidence rating of 1). Admittedly, this is a hypothetical example, although it is perfectly plausible.

Alternatively, if the overall level of confidence and the overall level of accuracy decreased in response to an experimental manipulation (e.g., as the retention interval increased), the accuracy associated with each individual level of confidence could remain unchanged. In fact, this is essentially the pattern of data reported by Sauer et al. (2010). When subjects identified a suspect as being in the lineup with greater than 50% confidence, the accuracy associated with ratings of 50% to 90% ranged from approximately 50% (not very accurate) to approximately 80% (reasonably accurate). As shown in Figure 4.2, this outcome held true for both short and long retention intervals even though, in the long retention interval condition, average confidence and average accuracy were both lower than in the short retention interval condition. Thus, retention interval, per se, did not have a dramatic effect on calibration. In either condition, if a subject expressed high confidence, accuracy was approximately 80% correct. Ratings of high confidence occurred more often in the short retention interval condition than in the long retention interval condition, but accuracy for high-confidence responses was approximately the same in both cases (though accuracy for low-confidence responses was higher in the short retention interval condition).

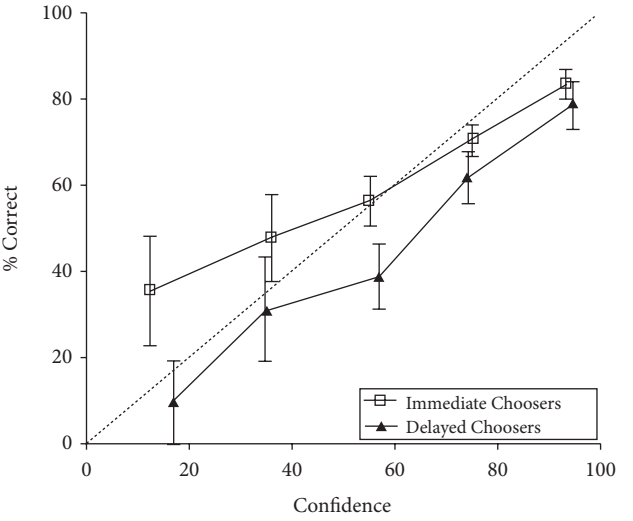



**Figure 4.2** Subjects chose a face from a lineup either shortly after seeing the person in a simulated crime or after a delay. Both confidence and accuracy were poorer after the delay, but the calibration plot showed that the relation between confidence and accuracy did not much change. Subjects were somewhat underconfident (i.e., their data are below the diagonal line) after the delay. Data are from Sauer, Brewer, Zweck, and Weber (2010). With kind permission from Springer Science + Business Media: *Law and Human Behavior*, "The Effect of Retention Interval on the Confidence-Accuracy Relationship for Eyewitness Identification," 34, 2009, 343, James Sauer, Fig. 1.

## Between-Events Correlations

A surprisingly overlooked type of analysis is that in which *memory events* or *items to be remembered* is the unit of analysis. The question can be posed in various ways. Is the confidence-accuracy relation (say, in a calibration plot) different across faces than across, say, sentences? For Caucasian viewers of faces, is the confidence-accuracy correlation different when they viewed a crime committed by Caucasians relative to Asians relative to African Americans? We cannot know the answers to these questions, for the good reason that no one seems to have asked them. Researchers have shown that there is an effect of race on accuracy, with people better able to recognize and differentiate people of their own race (e.g., Meissner & Brigham, 2001), but surprisingly, no one seems to have addressed the confidence-accuracy relationship in these cases.

Some evidence exists on the between-events correlations with verbal materials, primarily sentences and categorized word lists, and the results are unlike those we have observed to this point. Sampaio and Brewer (2009) used a sentence recognition task and looked at normal (nondeceptive) sentences (ones that generally were meaningful and straightforward) and what they called "deceptive" sentences. The latter consisted of sentences such as "The baby stayed awake all night" or "The karate champion hit the cinder block." In a later sentence recognition test, subjects could be given either the sentence they studied or a changed sentence. Subjects rated the sentence as old or new (studied or not studied), and they rated the confidence in their judgments. For the deceptive sentences the lures were "The baby cried all night" or "The karate champion broke the cinder block." These deceptive variations employ what are called *pragmatic implications*, because the studied sentences do not logically require the implication (the baby could have watched TV all night; the karate champion could have broken her hand), but nonetheless, most people draw the inference that the baby cried and the karate champion broke the cinder block. Sampaio and Brewer compared the correlation of accuracy and confidence for the two types of sentences. They found a modest positive correlation (+.30) between the two for the normal (nondeceptive) sentences, but they obtained a strongly negative correlation (–.61) for the deceptive sentences. That is, the more likely subjects were to false alarm to the sentence, the more confident they were in making the error (across sentences). Sampaio and Brewer concluded:

> With a list of nondeceptive items, one can have a strong positive relationship between confidence and accuracy. With a list including a mixture of deceptive and nondeceptive items, one can have no relationship between confidence and accuracy. With a list of only deceptive items, one can have a strong negative relation between confidence and accuracy. (p. 162)

DeSoto and Roediger (2011) showed a similar pattern in recognition of categorized word lists (that is, words belonging to common categories like birds or articles of clothing). Norms exist for categories that list members of the category in terms of their output dominance, or how likely people are to produce the word when asked to produce members of the category. For birds, *eagle* and *robin* are high-dominance members (likely to be produced by most people), whereas *kiwi* and *penguin* are lower dominance. DeSoto and Roediger had subjects study 10 items from a category like birds (some high dominance and some low dominance among the first 20 birds in the category) and then later gave them a recognition test. The subjects had studied 120 items (10 words from each of 12 categories), and the test provided them with 240 items (the 120 studied items and 120 distractors or lures—nonstudied items from the same categories). Subjects judged each word as old or new and then gave a confidence rating on a 100-point scale so that calibration curves could be plotted. DeSoto and Roediger were especially interested in false recognition to the lures, words from the same categories as studied words but ones that had not been studied. They found that false recognition was much greater for items of high dominance (like *eagle* and *robin*) than for those of low dominance (like *kiwi* and *penguin*). Further, confidence followed the same trend. The items on which subjects were most likely to produce a false alarm also led to false alarms with highest confidence. Thus, they also obtained a negative correlation between confidence and accuracy with related lures, much like Sampaio and Brewer (2009) with their deceptive sentences. Roediger and McDermott (1995) also showed high-confidence false alarms using somewhat different types of word lists.

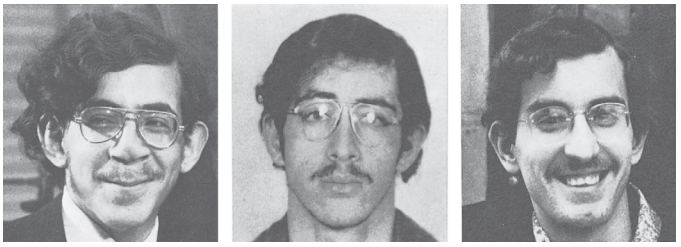
This type of negative relation between confidence and accuracy is especially troublesome for issues of courtroom testimony—what if eyewitnesses are more likely to be in error the more confident they are? Does this ever happen? Thus, we need to ask if the situation often faced by witnesses is at all like the situation in these experiments. After all, the experiments just described were conducted with verbal materials, whereas eyewitness cases often involve faces. Although evidence is sparse, certainly a case can be made that eyewitness situations may have this character of high similarity between the target person and the members of a lineup. Take an extreme case: A man commits a robbery and his identical twin is arrested for committing it. The witness views a lineup and decides that the man in the lineup is the perpetrator; in addition, she is sure that she is right. It is easy to understand how this situation could occur; the two people look very much alike (unless their hairstyles or other surface features are dissimilar), so a highly confident judgment is understandable, even if it turns out to be false.

The case of the lineup can be a less extreme version of the situation just described: A person witnesses a crime and sees the perpetrator. She gives

the police a general description. The police catch a suspect who fits the general description and construct a lineup. If the lineup is constructed properly according to most protocols, the people in the lineup will usually fit the perpetrator's general description. For example, if the perpetrator is described as an Asian American male, no African Americans or Caucasians would be in the lineup. Thus, by design, the members of the lineup will be somewhat similar to the perpetrator, and perhaps the suspect might be most similar. If so, this situation might lead to an erroneous identification. It seems plausible that this situation can arise in true lineups.

Consider the three men whose pictures appear in Figure 4.3. The man on the left was arrested in New York City and accused of committing a rape. He was picked out in a lineup by the victim. The man on the right was arrested for a robbery and also picked out in a lineup by a different witness. Both men spent time in jail before being exonerated. Eventually, the police captured the man in the middle, and he was convicted of both crimes. Obviously, the similarity among the three men is great. Most of us looking at Figure 4.3 could see how the first and third men might be mistakenly identified as the one in the middle. This case (a true case, from New York City in the early 1970s) shows how similarity relations can often affect recognition (the case comes from Buckhout, 1974).

Similarity relations in recognition are, in a way, obvious (but see Tulving, 1981, for a principled exception). Every student knows that a multiple-choice test with highly similar response alternatives is tricky. Lineups with "filler" people who are highly similar to the perpetrator can be viewed as tricky multiple-choice tests where the task is to pick a person from among similar alternatives



Lawrence Berson Richard Carbone George Morales

**Figure 4.3** 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. This example illustrates the problem of similarity in recognition memory. The suspect in the lineup may be judged to be the perpetrator of the crime just because he looks like the perpetrator.

or to say "none of the above." Because of high similarity, sometimes people can make high-confidence false identifications. The results described above by Sampaio and Brewer (2009) and DeSoto and Roediger (2011) show similar results, albeit with verbal materials.

One recent study to show similarity relations in recognition for faces is found in experiments by Carlson, Gronlund, and Clark (2008). In Experiment 2, 619 college students watched a video of a carjacking and then were asked to pick the perpetrator's face out of either a simultaneous or sequential lineup. In a simultaneous lineup, all the faces (both target and lures) are shown to the chooser at the same time; in a sequential lineup, however, only one face is shown at a time. Whether the perpetrator was present or absent in the lineup varied between conditions. The similarity of the different faces in the lineup also varied such that some subjects were presented with extremely similar faces and others were presented with dissimilar faces. Carlson et al. found that false identification in simultaneous lineups was dramatically affected by similarity, such that witnesses were much more likely to falsely identify an innocent face when all the faces presented were similar. In contrast, false identification was low when the faces were not overly similar. Similarity also appeared to have a reduced effect on false identification in sequential lineups. Although Carlson et al. did not collect confidence ratings in their experiment, confidence would probably have increased as a result of visual similarity as well.

## Between-Subjects Correlations

A different approach to understanding the relation between confidence and accuracy has been to ask whether there are individual differences in that relationship. That is, are people who generally express high confidence in their decisions also more accurate than those who express low confidence? Conversely, are those who are more accurate also more confident (relative to those who are less accurate)? To answer these questions, subjects might respond to a series of general knowledge questions, expressing confidence in each answer. At the end of the test, an average accuracy score and an average confidence score can be computed for each subject, and then a correlation across subjects can be computed. Of course, the same type of experiment can be done with any sort of material, so long as each subject is assessed on both accuracy and confidence for many events.

Robinson and Johnson (1996) provide a relevant example of the between-subjects approach. In this research, subjects watched a 3-minute video depicting a thief taking money from a woman's purse. After performing a neutral distractor task, subjects answered a questionnaire containing 32 multiple-choice or short-answer questions (e.g., "What was the thief doing before the

female teacher entered the snack bar?"). Subjects rated their confidence on a scale from 1 to 9 (1 = not at all confident, 9 = extremely confident). The researchers correlated mean confidence and percent correct for each subject over the 32 questions. Across a variety of recall and recognition conditions, a strong Pearson correlation was found, leading the researchers to conclude that eyewitnesses who tend to be more accurate also tend to be more confident. However, the strength of this between-subjects correlation varied depending on the experimental condition; for instance, the confidence-accuracy relation for subjects making confidence ratings immediately after they had made recognition judgments was nonsignificant ($r = .24$), but the confidence-accuracy relation when confidence ratings were made after all recall responses had been offered was quite strong ($r = .63$). Obviously, this consideration tempers general conclusions about the correlation between confidence and accuracy across people.

A second example of this type of analysis comes from a paper by Perfect, Watson, and Wagstaff (1993). In this experiment, subjects either answered 35 general knowledge questions or watched a forensically relevant 30-minute clip (from the film *Midnight Express*) and answered 35 questions about the clip. After each response, subjects indicated their confidence on a scale from 1 to 5 (1 = very confident, 5 = no idea). For each subject, mean confidence was calculated across all 35 responses, and the proportion correct of the 35 responses was obtained. Then, Pearson product-moment correlation coefficients were calculated between confidence and proportion correct for each individual subject.

Perfect et al. (1993) used the between-subjects approach to compare the confidence-accuracy relation for general knowledge responses and for the forensically relevant responses. They found that for general knowledge questions, there was a strong relation between confidence and proportion correct—subjects who were more confident on average were also more accurate. No relationship, however, was found between confidence and proportion correct for responses to the forensically relevant responses. This effect has been replicated in additional research (e.g., Perfect & Hollins, 1996). These researchers explained this effect by suggesting that significant between-subjects confidence-accuracy correlations can only result when the use of confidence is consistent across people—which is more likely for general knowledge questions than in eyewitness memory (Perfect, Hollins, & Hunt, 2000). Obviously, this research raises doubt about whether courts of law should give more weight to witnesses who are known to express (on average) high confidence in their decisions compared to witnesses who are known to express (on average) lower confidence in their decisions. It may be that highly accurate people are sometimes often lacking in confidence of their knowledge and/or that highly confident people may not be as accurate as they think.

In short, the low correlations between confidence and accuracy when measured between subjects in forensically relevant situations seem to argue that one need not necessarily believe in the accuracy of a witness who displays great confidence. It may well be that other individuals, less confident in their own abilities, may exhibit accuracy just as great as that of the highly confident person. This point essentially says that there are general individual differences in response bias and confidence that may inflate or deflate correlations, depending on their nature.

Although not strictly relevant to the issue at hand, bringing the qualities of subjects into the issue of eyewitness accountability leads to the issue of individual differences in eyewitness memory and testimony. Are children less reliable witnesses than adults? Are older adults less reliable than younger adults? We could add many more questions like this for other groups of people (e.g., people suffering from severe depression who also have poorer memories than matched controls). Reviewing this voluminous literature is beyond the bounds of this chapter, although readers could get a start on the question of children as witnesses from Ceci and Bruck's (1995) book and about older adults from a chapter by Roediger and McDaniel (2007).

## Within-Subjects Correlations

A within-subjects analysis asks whether different levels of confidence expressed by an individual are associated with different levels of accuracy. In the metacognition literature, this sort of analysis is referred to as *resolution* (Dunlosky & Metcalfe, 2009). This kind of analysis has been performed in studies in which witnesses provide answers to a variety of questions about the incident they witnessed. For example, Odinot et al. (2009) interviewed 14 witnesses to an actual armed robbery at a supermarket 3 months after the event. The crime was recorded on multiple security cameras, so the accuracy of witness recollections could be assessed. The witnesses were asked a variety of questions (e.g., to provide a full description of the robbers, the guns, the bag used, the position and acts of the robbers, the position and acts of the witness and his or her colleagues). For each answer, a confidence rating was taken using a 7-point scale (1 = very uncertain, 7 = absolutely certain). For each subject, a gamma correlation was computed between confidence and accuracy, and the obtained values ranged from .09 to .96. The mean value was only .38, which seems low. On that basis, the authors concluded that the relationship between confidence and accuracy is so weak that confidence ratings should never be taken into consideration in a court of law.

Does an intrasubject correlation coefficient provide the information that is needed to make such a recommendation? Possibly not. Odinot et al. (2009)

also provided information that allowed one to compute a calibration-like plot of the relationship between confidence and accuracy. Of the 14 witnesses in their study, 9 were centrally involved and were interviewed by the police (the other 5 were not). For the 9 central witnesses, the accuracy of recall associated with ratings of low confidence (1 through 3) was 61% correct. For intermediate confidence ratings (4 through 6), accuracy was 71% correct. For ratings of high confidence (7, which is the confidence rating that was most frequently supplied), accuracy was 85% correct. Thus, for low-confidence ratings, the witness's recollections were only 1.6 times more likely to be correct than incorrect. For high-confidence ratings, they were 5.7 times more likely to be correct than incorrect. This indicates that the relationship between confidence and accuracy is meaningful. At the same time, expressions of high confidence were associated with a 15% error rate. Thus, as we observed in an earlier analysis, a meaningful relationship between confidence and accuracy does not by any means imply that high-confidence memories are even close to being error free. Still, contrary to Odinot et al.'s (2009) conclusion, confidence does seem useful in forensic cases, including the real-life case they investigated.

## Within/Between Hybrid Correlations

In the eyewitness domain, it is often the case that the correlation between confidence and accuracy is based on a single accuracy score for a test item and a single confidence rating associated with that item for each subject. For example, subjects might watch a video of a staged crime scene. Later, the subjects would be asked to try to identify the suspect from a photo lineup. Each subject would provide a confidence rating using a numeric scale (e.g., 1 = low confidence, 5 = high confidence), and his or her accuracy would be scored as either being correct (1) or incorrect (0). A point-biserial correlation would then be computed between the confidence and accuracy scores across subjects. This approach is actually a hybrid within/between-correlational approach because the point-biserial correlation computed in this manner is sensitive to both within-subjects variation in confidence and accuracy (if subjects are more accurate when they are more confident) and between-subjects variation in confidence and accuracy (if subjects who are more confident, on average, are also more accurate, on average).

Most of the claims in the eyewitness literature to the effect that the relationship between confidence and accuracy is weak or nonexistent used the hybrid within/between correlational method. The hybrid design of these experiments seems reasonable if the goal is to generalize to courts of law, which are often faced with different individuals who have each been exposed to a single event and who are then asked to make an eyewitness identification.

In studies that have used the hybrid approach, the obtained point-biserial correlation between confidence and accuracy is often very weak, and this created the once-widespread impression that "Witnesses who are confident in their testimony are not substantially more accurate than those who are not" (Smith et al., 1989, p. 358). However, it has since become clear that this conclusion may have been premature because it was based on the point-biserial correlation coefficient, which is not well suited to answering the question of whether witnesses who are more confident in their testimony are also more accurate in one-shot situations. As indicated above, a much more straightforward way to answer this question is to simply compute the likelihood of being correct as a function of confidence expressed (Juslin et al., 1996). Theoretical considerations based on signal-detection theory suggest that, when computed in this more direct way, a substantial relationship between confidence and accuracy should be evident (even if the point-biserial correlation often suggests otherwise). Juslin et al. first showed that for data that exhibit a very strong relationship between confidence and accuracy (in that expressions of high confidence are associated with high accuracy and expressions of low confidence are associated with low accuracy), the point-biserial correlation is nevertheless rather low. This indicates that the point-biserial correlation is problematic (not that the relationship between confidence and accuracy is weak).

## Summary of the Various Analyses of Confidence and Accuracy

What is the relationship between confidence and accuracy in reports from memory? The preceding analyses show why this question, as stated, is not really meaningful because it could be asking about the relationship between confidence and accuracy across different conditions or across different people or across different responses made by an individual or across different items within a condition. The answers need not be the same in each domain, so no single answer is possible. Thus, for example, there is no contradiction between the statement that the relationship between confidence and accuracy across different individuals is weak and the statement that the relationship between confidence and accuracy across different responses made by individuals is strong. Consider, for example, a confident person (one who averages 4 on a 5-point confidence scale) whose average accuracy in memory decisions is relatively low (e.g., 59% correct) versus a less confident person (one who averages 3 on a 5-point scale) whose average accuracy is somewhat higher (e.g., 65% correct). In this case, we would be rather unimpressed by the relationship between confidence and accuracy (across people). However, for those same two people, it might also be true that when they express low confidence, their accuracy

tends to be low, whereas when they express high confidence, their accuracy tends to be much higher.

Table 4.1 illustrates this situation with hypothetical data for subjects with varying levels of accuracy (percent correct) and confidence (measured on a 1–5 scale). For both subjects in this example, accuracy is very low when confidence is low (1) but accuracy is much higher when confidence is high (5). Even so, the average accuracy for the relatively confident Subject 1 (89 out of 150 correct) is only 59%, whereas the average accuracy of the less confident Subject 2 (97 out of 150 correct) is 65%. The average confidence for Subject 1 is higher because most decisions were made with high confidence (5), whereas most of Subject 2's were made with medium confidence (3).

Every approach to assessing the relationship between confidence and accuracy is relevant to psychological theory, whether that relationship is measured by a correlation coefficient or by a calibration curve. For example, why is the relationship between confidence and accuracy higher for some stimulus materials than for others? Such questions are important for the experimental psychologist to answer to inform theory development, but the answers are potentially relevant to the legal system as well. For example, if the events of interest in a particular crime happen to be stimuli for which ratings of confidence in the laboratory have been found to be especially poor indicators of accuracy, such information would be important to know. However, at the present time, we know precious little about how to characterize the stimuli that are encountered in real-world crime situations in terms of how they might influence the relationship between confidence and accuracy in later reports

*Table 4–1* Hypothetical Illustration of Correct and Incorrect Responses as a Function of Confidence Made Using a 5-Point Scale for Two Subjects

| | Confidence | Total Responses | Correct | Incorrect | % Correct |
|---|---|---|---|---|---|
| Subject 1 | 1 | 10 | 1 | 9 | 10% |
| | 2 | 15 | 3 | 12 | 20% |
| | 3 | 20 | 6 | 14 | 30% |
| | 4 | 25 | 15 | 10 | 60% |
| | 5 | 80 | 64 | 16 | 80% |
| | **Σ** | **150** | **89** | **61** | |
| Subject 2 | 1 | 30 | 9 | 21 | 30% |
| | 2 | 20 | 10 | 10 | 50% |
| | 3 | 50 | 35 | 15 | 70% |
| | 4 | 20 | 16 | 4 | 80% |
| | 5 | 30 | 27 | 3 | 90% |
| | **Σ** | **150** | **97** | **53** | |

from memory. Thus, for the time being, the best we can do in order to generalize conclusions about the relationship between confidence and accuracy to the legal setting is to arrange ecologically valid experiments that seem to be as closely modeled on real-world situations as possible. Many experiments in the applied literature have followed that approach, and some general conclusions for the legal system can be drawn from them.

## PRACTICAL IMPLICATIONS FOR THE LEGAL SYSTEM

Of the various questions that have been asked about the relationship between confidence and accuracy, the two that seem most relevant to the legal setting are the within-person measures (e.g., when a witness recounts multiple crime-scene details, each accompanied by a confidence rating) and the within/between hybrid measures (e.g., when different witnesses make a single eyewitness judgment accompanied by a confidence rating). Ideally, studies of these questions would be conducted in the real world and would involve events associated with real crimes, known perpetrators, and real police lineups. Obviously, this ideal is rather hard to achieve, but in studies that have attempted to match reality as closely as possible, two points seem to emerge (from both within-person and within/between hybrid studies):

1. Confidence is a reliable indicator of accuracy in the sense that reports from memory made with low confidence are generally associated with low accuracy, whereas reports made with high confidence are generally associated with higher accuracy. This is true even though point-biserial and gamma correlation coefficients are often low. Whereas the correlation coefficient offers little useful information for courts of law, the descriptive relationship between levels of confidence and associated levels of accuracy provides more intelligible and (therefore) useful information (Juslin et al., 1996). Although it is contrary to the legal testimony of many expert witnesses in recent years, the evidence suggests that, in the absence of known contamination (e.g., exposure to misleading postevent information), it is reasonable to regard the confidence expressed by eyewitnesses as a useful indicator of the reliability of the memory decision. This is especially so for more immediate tests of memory and not necessarily for courtroom testimony that often occurs much later. The reason is that we base our conclusions on experiments in which confidence and accuracy are assessed shortly after the witnessed event and without repeated testing. Thus, our recommendation that confidence should be taken into account applies most strongly to police interrogations. Repeated questioning of witnesses has sometimes been found to increase confidence without increasing accuracy (Shaw & McClure, 1996). By the time a witness arrives in court (often months or even years after

the occurrence of the crime), confidence may be relatively fixed by prior tests. If the witness has been confident in his or her judgment ever since the first examination by police, then the high confidence may be warranted. However, if confidence was low on the initial examination (say, a photo lineup) but then grew over time and repeated testing (more photo lineups, a real lineup, identification in court), then the confidence may be less trustworthy. Certainly, even on an immediate test, confidence is not infallible, which leads to the next point.

2. Ratings made with high confidence can be associated with an error rate that is far too high for someone to be considered guilty of a crime solely on the basis of high-confidence identification by a single eyewitness. Indeed, it is not uncommon to find accuracy rates associated with high confidence in the range of 80% to 90% correct (i.e., a 10% to 20% error rate) even in situations that do not involve misleading postevent information or potentially confidence-inflating activities such as repeated postevent questioning (Wells, Memon, & Penrod, 2006). Thus, in light of the evidence, convicting a defendant solely on the basis of the high-confidence memory-based testimony of one witness who is identifying a previous stranger should itself be a crime (or at least it is wrong). Reports from memory—including ones with high confidence—are simply not a reliable enough indicator of truth to unilaterally adjudicate guilt or innocence. The situation is different if one is identifying a well-known person rather than a stranger. Nonetheless, the point is that eyewitness testimony should be considered one piece of evidence in a complex web of information (direct or circumstantial) that would indicate a person's guilt or innocence in criminal situations. We believe it is a mistake to convict someone based on this single piece of evidence because of the many problems discussed in this chapter.

## CONCLUSION

The main point of our chapter is that eyewitness memory confidence is a useful but imperfect indicator of the truth. In that sense, it is much like all other forms of evidence that courts must deal with on a daily basis. With the possible exception of DNA evidence, which can approach the ideal of infallibility, evidence in the real world is almost invariably fallible, including fingerprint evidence, fiber evidence, microscopic hair analysis, bloodstain pattern analysis, handwriting analysis, and so on. Even DNA evidence is not as simple as it is often portrayed (see Lindsey, Hertwig, & Gigerenzer, 2003). In a recent report to Congress on the subject of forensic analysis, the National Research Council (2009) observed:

> A body of research is required to establish the limits and measures of performance and to address the impact of sources of variability and potential

> bias. Such research is sorely needed, but it seems to be lacking in most of the forensic disciplines that rely on subjective assessments of matching characteristics. These disciplines need to develop rigorous protocols to guide these subjective interpretations and pursue equally rigorous research and evaluation programs. (p. S-6)

Subjective interpretation is an inescapable component of many forms of evidence evaluation, not just when the evidence consists of eyewitness memory. Whether the evidence consists of eyewitness memory or expert fingerprint analysis, confidence ratings can help to provide an indication of its reliability. With regard to fingerprint evidence, for example, Mnookin (2008) argued that fingerprint examiners would do well to provide graded expressions of confidence in their analyses instead of always expressing certainty so that courts of law could better appreciate how reliable the analysis is:

> Given the general lack of validity testing for fingerprinting; the relative dearth of difficult proficiency tests; the lack of a statistically valid model of fingerprinting; and the lack of validated standards for declaring a match, such claims of absolute, certain confidence in identification are unjustified, the product of hubris more than established knowledge. (p. 139)

Fingerprint analysts presumably have the training and experience to judge the confidence of their analyses in a way that corresponds to the accuracy of their analyses. However, it seems reasonable to suppose that research would show that while their expressions of confidence are indicative of accuracy, expressions of certainty would not indicate 100% accuracy. Whether or not that is true of fingerprint analysts, it does appear to be true of eyewitnesses. By virtue of a lifetime of training and experience in the use of memory, eyewitnesses appear to have acquired some degree of expertise in judging the reliability of their own recollections. Even so, that expertise has its limits, with the most important one being that expressions of 100% certainty or confidence do not reflect 100% accuracy.

Because virtually all forms of evidence are fallible (even when accompanied by expressions of 100% certainty), multiple indicators of guilt should be combined to eliminate reasonable doubt, as might happen if all of the indicators point strongly in the direction of guilt. High-confidence eyewitness testimony alone can never do that, for the simple reason that virtually all studies show that ratings of high confidence occur with significant probabilities of error (often 10%–20% in experimental settings). At the same time, ignoring confidence ratings altogether is chucking the baby out with the bathwater. A witness who is highly confident about some details of an event almost certainly is

more accurate on these details than on those details provided with low confidence. Precisely because evidence is fallible, every reliable indicator should be made available to the court, and confidence in reports from memory should be no exception.

## ACKNOWLEDGMENTS

We thank Pooja Agarwal, Sean Kang, Kathleen McDermott, Laura Mickes, Lynn Nadel, John Nestojko, Adam Putnam, Walter Sinnott-Armstrong, and Yana Weinstein for comments on prior drafts of this chapter.

## REFERENCES

Barclay, J. R., Bransford, J. D., Franks, J. J., McCarrell, N. S., & Nitsch, K. (1974). Comprehension and semantic flexibility. *Journal of Verbal Learning and Verbal Behavior*, *13*, 471–481.

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

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. New York: Oxford University Press.

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.

Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied*, *14*, 118–128.

Ceci, S. J., & Bruck, M. (1995). *Jeopardy in the courtroom: A scientific analysis of children's testimony*. Washington, DC: American Psychological Association.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

DeSoto, K. A., & Roediger, H. L. (2011). *Often wrong but never in doubt: Typicality relates to confident false memories*. Manuscript in preparation, Psychology Department, Washington University, St. Louis, MO.

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

Ebbesen, E. B., & Rienick, C. B. (1998). Retention interval and eyewitness memory for events and personal identifying attributes. *Journal of Applied Psychology*, *83*(5), 745–762.

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

Hamilton, E. (1961). *Plato: The collected dialogues*. New York: Bollingen Foundation.

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.

Kassin, S. M., Ellsworth, P. C., & Smith, V. L. (1989). The "general acceptance" of psychological research on eyewitness testimony: A survey of the experts. *American Psychologist*, *44*, 1089–1098.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American Journal of Psychology*, *93*, 329–343.

Knutsson, J., Allwood, C. M., & Johansson, M. (2011). Child and adult witnesses: The effect of repetition and invitation-probes on free recall and metamemory realism. *Metacognition and Learning, 3*, 213–228.

Lindsey, S., Hertwig, R., & Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics, 43*, 147–163.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, *7*, 550–572.

Loftus, E. F. (1992). When a lie becomes memory's truth: Memory distortion after exposure to misinformation. *Current Directions in Psychological Science*, *1*, 121–123.

Loftus, E. F. (1996). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*, 585–589.

McGeoch, J. A. (1932). Forgetting and the law of disuse. *Psychological Review*, *39*, 352–370.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*, 3–35.

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.

Mnookin, J. L. (2008). The validity of latent fingerprint identification: Confessions of a fingerprinting moderate. *Law, Probability and Risk, 7*, 127–141.

Münsterberg, H. (1908). *On the witness stand*. New York: Doubleday.

National Research Council (2009). *Strengthening forensic science in the United States: A path forward.* Washington, DC: National Academies Press.

Neil v. Biggers, 409 U.S. 188 (1972).

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., & Hollins, T. S. (1996). Predictive feeling of knowing judgments and postdictive confidence judgments in eyewitness memory and general knowledge. *Applied Cognitive Psychology*, *10*, 371–382.

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.

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. (1980). Memory metaphors in cognitive psychology. *Memory & Cognition*, *8*, 231–246.

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. (2011). *Complexities in the relation between confidence and accuracy in recognition memory.* Manuscript in preparation, Psychology Department, Washington University, St. Louis, MO.

Roediger, H. L., & Gallo, D. A. (2002). Processes affecting accuracy and distortion in memory: An overview. In M. L. Eisen, G. S. Goodman, & J. A. Quas (Eds.), *Memory and suggestibility in the forensic interview* (pp. 3–28). Mahwah, NJ: Erlbaum.

Roediger, H. L., Jacoby, J. D., & McDermott, K. B. (1996). Misinformation effects in recall: Creating false memories through repeated retrieval. *Journal of Memory and Language*, *35*, 300–318.

Roediger, H. L. & McDaniel, M. A. (2007). Illusory recollections in older adults: Testing Mark Twain's conjecture. In M. Garry & H. Hayne (Eds.), *Do justice and let the sky fall: Elizabeth F. Loftus and her contributions to science, law, and academic freedom* (pp. 105–136). Hillsdale, NJ: Erlbaum.

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., & Payne, D. G. (1982). Hypermnesia: The role of repeated testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*, 66–72.

Rubin, D. C., & Wallace, W. T. (1989). Rhyme and reason: Analyses of dual retrieval cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 698–709.

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

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior*, *34*, 337–347.

Shaw, J. S. (1996). Increases in eyewitness confidence resulting from postevent questioning. *Journal of Experimental Psychology: Applied*, *2*, 126–146.

Shaw, J. S., & McClure, K. A. (1996). Repeated postevent questioning can lead to elevated levels of eyewitness confidence. *Law and Human Behavior*, *20*, 629–653.

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. (1974). Cue-dependent forgetting. *American Scientist*, *62*, 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*. New York: Oxford University Press.
Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, *5*, 381–391.
Wells, G. L., & Bradfield, A. L. (1999). Distortions in eyewitnesses' recollections: Can the postidentification-feedback effect be moderated? *Psychological Science*, *10*, 138–144.
Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest*, *7*, 45–75.
Wells, G. L., & Murray, D. M. (1983). What can psychology say about the *Neil v. Biggers* criteria for judging eyewitness accuracy? *Journal of Applied Psychology*, *68*, 347–362.
Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review*, *117*(4), 1025–1054.