19

APPLYING CONFIDENCE-ACCURACY CHARACTERISTIC PLOTS TO RECOGNITION MEMORY

Henry L. Roediger III, Eylul Tekin, and Wenbo Lin

The study of recognition memory, either introspectively (Allin, 1896) or more objectively (Strong & Strong, 1916), is over 100 years old. Hundreds, maybe thousands, of experiments have been published on this topic. Besides this great body of empirical work, certain developments in other fields applied to recognition have provided conceptual leaps in our understanding. Perhaps the greatest such leap is the application of signal detection theory to recognition memory (see Wixted, 2020, for a historical review). Other, less dramatic influences have come in the form of new analytic techniques. The purpose of this chapter is to review one such technique,—the confidence-accuracy characteristic plot,—and its application to recognition memory.

Confidence-accuracy characteristic (CAC) plots were developed from calibration measures used originally in the study of perception and metamemory. We review these developments briefly before we discuss how CAC plots have transformed the critical question of confidence-accuracy relations in eyewitness memory. We address several attendant questions, such as whether the "length" of confidence scales (a 1–4 scale vs. a 1–100 scale) matters. Do fine-grained scales permit better calibration? We also address the issue of whether numeric scales provide more accurate judgments of confidence than verbal scales. We then turn our attention to applying CAC plots to laboratory tasks involving recognition memory, where their use raises many interesting questions. Such applications are just beginning. We also discuss two fundamentally different ways of constructing CAC plots and how they illuminate different questions. We end with some speculations about future applications of CAC plots to understand remembering more generally.

DOI: 10.4324/9781003158134-20

Development of CAC plots

The study of how confidence is related to accuracy has a long history in psychology, and the question can be asked in several different ways (see Roediger et al., 2012, for a review). In an interesting early study, Dallenbach (1913) conducted experiments with rich results, which led him to conclude that: "The degree of certainty of the observer's replies bears a direct relation to the fidelity of the answer" (p. 335). In short, confidence and accuracy go together.

Skipping ahead about 70 years, we can see that the situation has become muddled with regard to confidence and accuracy. Two quotes help illustrate the conundrum in this area of research in the 20-year period from roughly 1985 until 2015. One tradition of research is that of the psychological laboratory in which subjects typically study a list of words or pictures and then are asked to recognize them later among lures. Reviewing this research, Dunlosky and Metcalfe (2009) concluded that, "The relative accuracy of people's confidence is high. Higher confidence ratings almost inevitably mean that the item had been previously presented" (p. 176). This statement confirms Dallenbach's (1913) early finding. On the other hand, in the study of eyewitness memory, numerous researchers reached the opposite conclusion from lab-based studies conducted to simulate the experience of an eyewitness. Reviewing this literature in 1989, Smith et al. argued that "confidence is neither a useful predictor of the accuracy of a particular witness nor of the accuracy of particular statements made by the same witness" (p. 358). Other later research up until around 2015 confirmed this conclusion.

How could research in these different traditions lead to such opposite conclusions? Is eyewitness memory really so different from studies using words and pictures? No, it's not. Other measurement factors explain the difference in conclusions. Juslin et al. (1996) pointed out that eyewitness identification research used the point-biserial correlation to assess confidence-accuracy relations. They argued that this measure is flawed (see Wixted & Wells, 2017, for the rationale) and that a different approach using calibration answers the critical question of whether confidence predicts accuracy much more directly. In calibration measures, confidence of a response is placed in bins on the abscissa of a graph, with accuracy plotted on the ordinate. Juslin et al. (1996) conducted eyewitness identification experiments using lineups provided by the Swedish police, and they analyzed the results by providing a calibration function. Confidence was measured on a 100-point scale in this experiment. The results are shown in Figure 19.1, and it is clear that confidence is a strong predictor of accuracy. In this crime scenario, both a central and a peripheral confederate appeared, but the confidence-accuracy (hereafter, CA relationship) was strong for both.

Calibration approaches are often used in human experimental psychology in psychophysics, metamemory, and judgment and decision making, among other areas. However, true calibration approaches require a 100-point scale so that one can determine if, say, 80% confidence corresponds to 80% accuracy. In the eyewitness identification literature, researchers do not solely use 100-point scales for the good reason that police departments do not use them. Researchers typically use much simpler scales in this area of research. For example, after a "witness" selects

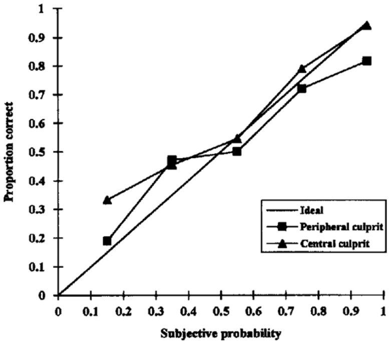


Figure 19.1 A calibration plot depicting the CA relationship for the central and peripheral culprits (from Juslin et al., 1996). High confidence indicates high accuracy.

someone from a lineup in an experiment, they might be asked to provide their confidence on a 4-point scale of "I am absolutely sure I picked the right person" to "I feel sure I picked the right person," to "I think I picked the right person, but I'm not sure," to "I'm just guessing that I picked the right person." This amounts to a 4-point rating scale, with decreasing confidence from 4 to 1.

A true calibration approach, as used in constructing Figure 19.1, requires a 100-point scale, to see if subjects' estimates of confidence match their accuracy, also on a 100-point scale. Therefore, for smaller scales, a different technique called confidence accuracy characteristic (CAC) plots have been developed (Mickes, 2015). In eyewitness experiments, after viewing a crime scenario, "witnesses" may see a target-present lineup (the suspect randomly placed among five filler faces who generally match characteristics of the suspect in terms of race, hair color and style, eye color, etc.) or a target-absent lineup with six filler faces. The hit rate is the percentage of times people pick the suspect in the target-present lineup. In the target-absent lineup often one person is selected as the "innocent suspect," so the false alarm rate is the percentage of people picking that person. If there is no designated innocent suspect, then the total number of picks from the target-absent lineup is divided by six to obtain a corrected false alarm rate.

A confidence rating is given after each selection. To calculate the CAC accuracy for identifications made with a particular level of confidence (e.g., a confidence rating of 3), the number of target-present suspect IDs given a confidence rating of 3 is divided by the number of target-present suspect IDs plus the number of target-absent suspect IDs, all of whom received ratings of 3. Thus, the formula for accuracy for each confidence bin is:

Suspect IDs_TP/(Suspect IDs_TP + Suspect IDs_TA)

The general logic is that accuracy in CAC plots is the hit rate divided by the hit rate plus the false alarm rate. The exception is that selection of fillers by the witness in the target-present lineup is ignored in this analysis, because they are known to be innocent and thus would never be convicted of a crime.

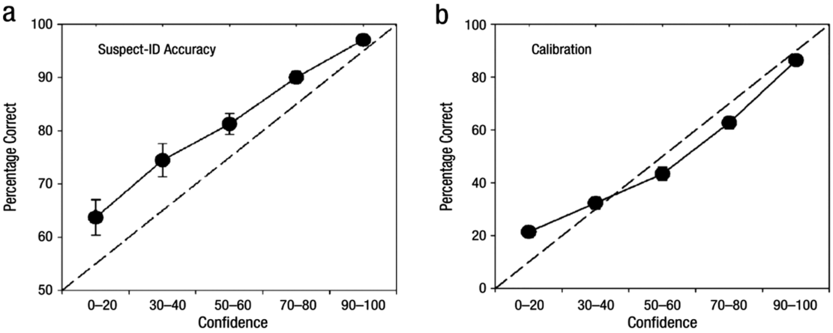


Figure 19.2 (A) shows a CAC plot and (B) shows a calibration plot (from Wixted & Wells, 2017). High confidence indicates high accuracy.

In many experiments, the CA relationship as measured by a CAC plot looks about the same as the calibration plot in Figure 19.1. Figure 19.2 here is from an analysis by Wixted and Wells (2017) in which they aggregated over 15 sets of data. Higher confidence clearly indicates higher accuracy. However, points fall above the diagonal line here, indicating that the witnesses are actually underconfident in their identification responses. That is, for the lower confidence bins, accuracy is greater than the level of confidence expressed in the report of accuracy.

The use of CAC plots in eyewitness identification experiments has helped change the conclusion about confidence in eyewitness situations: High confidence indicates high accuracy. We should add that this conclusion is true only on a first test with a fair lineup in adults (see Wixted & Wells, 2017; Wixted, 2021 for discussion of these issues). If the lineup is biased by having a filler stand out or a close look alike to the actual suspect appear in the lineup, then the confidence-accuracy relationship may break down or even reverse (e.g., DeSoto & Roediger, 2014).

Applying CAC plots to problems in eyewitness memory

In this next section, we report on the use of CAC plots in investigations of repeated lineups. Repeated presentation of faces from mugshots to lineups or from a photo lineup to a live lineup may be problematic, because only one face—i.e., the suspect's—appears twice, and so he or she might be selected in the second lineup based on mere familiarity from the first mugshot or lineup. We tried to get around this issue by seeing if performance would improve if the same suspect and fillers were used in two lineups. The other issue we report on here is whether the grain size of lineups (coarse, as in a 1–4 confidence scale, or fine-grained as in a 1–20 or 1–100 confidence scale) makes a difference in eyewitness identification studies. Do finer grained scales lead to better judgments?

Effects of repeated lineups

Numerous studies have shown that eyewitness confidence reliably predicts accuracy in the initial identification by the witness, but what about eyewitness confidence expressed in a subsequent identification? In some criminal investigations, witnesses' memories are tested on more than one occasion (e.g., a six-person photo lineup followed by a six-person live lineup). Law enforcement officers may use repeated identification procedures as a means to confirm an earlier identification, especially when a prior identification is disputed (Steblay & Dysart, 2016). Some examples of repeated identification procedures are mugshots followed by a lineup, a prior show-up (one-person identification procedure) followed by a lineup, and consecutive lineups such as a photo lineup followed by a live lineup or another photo lineup. Regardless of the presentation format, repeated identification procedures typically involve a single repeated suspect (i.e., only the suspect appears in both the initial and subsequent identifications).

Laboratory studies have suggested a number of ways in which the use of a single-repeated suspect inadvertently biases witnesses towards choosing that suspect (Brigham & Cairns, 1988; Pezdek & Blandon-Gitlin, 2005; Steblay et al., 2013; Valentine et al., 2012). In the subsequent identification, some witnesses may misattribute their familiarity of having recently seen the suspect in the initial identification to having seen the person committing the crime. Other witnesses may be biased towards selecting the same suspect in the subsequent identification because they misinterpret the police's intention (i.e., believing that the police have identified that particular member as the perpetrator), or they feel compelled to commit to their initial decision to appear reliable and consistent to the police (a commitment effect).

To address the aforementioned issues in repeated identification procedures, Lin et al. (2019) proposed repeating both the suspect and fillers across lineups (repeated lineups). Issues with misplaced familiarity and misinterpretation of the police's intention should be eliminated because the same lineup members would appear in both identification occasions and all members would be familiar to the witness. This procedure would not necessarily eliminate the commitment effect, but it would provide a measure of the commitment effect unconfounded by the other issues. Of course, the critical question is whether the eyewitness confidence-accuracy relationship would be impaired in the subsequent identification.

The short answer is no. The CAC curves did not significantly differ across both the initial and subsequent identifications, as seen in Figure 19.3. Eyewitness confidence did not inflate across repeated lineups, but initial identification decisions made with high confidence were more likely to be carried over to the subsequent identification; however, eyewitness confidence remained highly predictive of accuracy. Furthermore, Lin et al. (2019) also varied the length of two types of retention intervals: Event to Lineup1 and Lineup1 to Lineup2. That is, either the Event to Lineup1 and/or Lineup1 to Lineup2 intervals could be long or short, or both. Regardless of retention interval manipulations, CAC analyses revealed that high confidence was still associated with high accuracy in both the initial and subsequent lineups.

Confidence scales: Does granularity matter?

Laboratory studies on eyewitness memory might use numeric and granular confidence scales to assess participants' confidence concerning their eyewitness identifi-

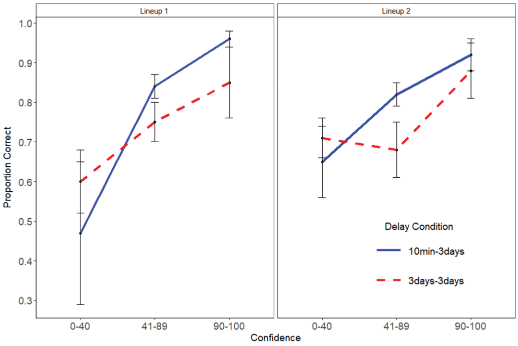


Figure 19.3 CAC plots for the initial and subsequent identifications from Experiment 1 of Lin et al. (2019). The two numbers indicate the time between the crime scenario and the first lineup and the time between the first lineup and the second lineup. Error bars indicate standard errors of the mean.

cations (e.g., 10-, 20-, or 100-point scales). Although these studies report a positive and strong relationship between confidence and accuracy on initial eyewitness identifications using calibration or CAC analysis (Mickes, 2015; Palmer et al., 2013; Sauer et al., 2010; Wixted et al., 2015), the confidence scales in the laboratory studies are not representative of eyewitness identification procedures conducted by police departments. In contrast to granular (or more fine-grained) confidence scales, police departments in the US assess eyewitness confidence either through their verbal expressions without using an official confidence scale (e.g., "I am sure this is the guy") or through verbal and narrow confidence scales with few levels of confidence (Behrman & Richards, 2005; Wells, 2014). As an example, the Houston Police Department used a verbal confidence scale with three levels, i.e., positive, strong tentative, weak tentative (Wells, 2014). Critically, the positive confidence-accuracy relationship observed in laboratory eyewitness identification experiments might weaken or break down when wider confidence scales are employed because witnesses might not have much room to differentiate across confidence levels. Furthermore, using verbal confidence scales without numbers might also affect how people map their confidence on the scale and thus change the confidence-accuracy relationship.

To address these issues, Tekin et al. (2018) examined whether confidence assessed by narrow verbal confidence scales, which are more likely to be used by police departments, predicted eyewitness accuracy. In their experiment, participants watched two 30-s silent videos in which a suspect stole a laptop and showed his face for approximately 4 s. After each video, participants completed a short filler task and then received lineup instructions that stated that the suspect may or may not be present in the lineup. They were instructed to select the suspect if he was present and to reject the lineup if the suspect was not present. Participants then saw either the corresponding target-present or target-absent lineup for the video they had just watched. If participants saw the target-present lineup for the first video, they saw the target-absent lineup for the second video and vice versa. The lineups had six headshots that were presented in a 2 by 3 matrix. The target-present lineups consisted of the corresponding suspect and five fillers, whereas the target-absent lineups consisted of six fillers (i.e., the same five fillers and an additional one). The fillers were chosen due to their general resemblance to the suspect (e.g., age, race, hair color). In addition, two lineups were used, and they differed greatly in their difficulty. For the Set A lineup, the suspect identification rate in target-present lineups was 27%, whereas for the Set B lineup it was 73%. Thus, we could examine performance for a relatively easy and a relatively hard lineup.

After participants made an identification decision (i.e., choosing someone or rejecting the lineup), they rated their confidence on one of the following scales: (1) a verbal-only 2-point scale; (2) a verbal and numeric 2-point scale; (3) a verbal-only 4-point scale; or (4) a verbal and numeric 4-point scale. The verbal labels on the 4-point scales were "not sure at all," "somewhat sure," "very sure," and "absolutely sure," respectively, whereas the labels on the 2-point scales were "not sure at all" and "absolutely sure." For verbal and numeric scales the corresponding number was presented along with the verbal statement (e.g., "3—very sure").

Using the CAC analysis, Tekin et al. (2018) calculated the accuracy of 2- and 4-point scales across different confidence levels. They binned the lowest two points (i.e., "not sure at all" and "somewhat sure") and the highest two points (i.e., "very sure" and "absolutely sure") of the 4-point scales and compared them to the low and high points of the 2-point scales, respectively. Figure 19.4 demonstrates this comparison. First, both scales showed a positive confidence-accuracy relationship for both easy and difficult eyewitness videos. Second, the scale granularity did not matter for high confidence responses. That is, high confidence identifications on 2- and 4-point scales yielded similar levels of accuracy for both eyewitness videos. However, for the easy video, low confidence identifications on the 2-point scale led to lower accuracy relative to low confidence responses on the 4-point scale.

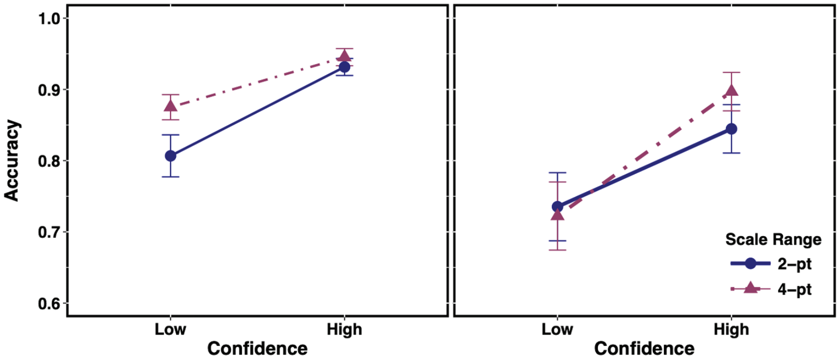


Figure 19.4 CAC plots for the 2- and 4-point scales for easy (left) and difficult (right) lineups from Experiment 1 of Tekin et al. (2018). Error bars indicate standard errors. Data are combined over verbal and verbal + numeric lineups because no difference was observed between them (see Tekin et al. 2018, Figure 3).

Lastly, for both videos, CAC analysis showed no accuracy difference between the verbal-only and verbal and numeric scales across confidence levels (see Figure 19.3 in Tekin et al., 2018). Dodson and Dobolyi (2015) also compared more granular verbal and numeric scales (e.g., 6-points and 11-points) in an eyewitness identification experiment. Using calibration plots (a variation of Juslin et al., 1996) rather than CAC plots, they found that the scale granularity did not matter.

These findings suggest that CAC plots (or calibration plots) are useful for addressing methodological questions, such as confidence granularity, in the eyewitness literature. Critically, in CAC analysis, accuracy can be plotted as a function of any confidence granularity (e.g., 2-point versus 100-point), not just 100-points. Using CAC plots, Tekin et al. (2018) demonstrated a strong confidence-accuracy relationship for narrow and verbal confidence scales, which are more likely to be administered by police departments.

Applying CAC plots to recognition memory experiments in the lab

As discussed previously, CAC plots have changed important conclusions about eyewitness identification. However, until recently this same analytic technique has not been applied to recognition memory in laboratory settings. Might CAC plots also provide interesting new information in this area of inquiry?

Granularity, revisited

CAC plots are most commonly employed in eyewitness identification experiments to examine the confidence-accuracy relationship. In these experiments, participants are usually presented with a single perpetrator (i.e., a one-item recognition experiment) and asked to identify the perpetrator from a lineup (Brewer et al., 2002; Wetmore et al., 2015). Even in studies using multiple perpetrators, as in Dodson and Dobolyi (2015), the number of perpetrators (or items) does not exceed 12. Therefore, eyewitness identification experiments are different from face recognition experiments that use large numbers of faces. It is plausible that the strong confidence-accuracy relationship obtained in eyewitness identification experiments using CAC plots might weaken or break down with large numbers of faces due to interference among targets and lures. Furthermore, although Tekin et al. (2018) showed that 2- and 4-point scales yielded similar CAC plots, it is plausible that much more granular scales might permit better judgments of confidence (e.g., 20-point, 100-point scales), especially in recognition experiments where there are many items to discriminate from one another.

Tekin and Roediger (2017) examined these issues in a face recognition experiment. They manipulated the granularity of confidence scales between-subjects and examined whether 4-, 5-, 20-, and 100-point confidence scales produce similar confidence-accuracy relationships using CAC analysis. In their experiment, participants studied 50 neutral faces one-by-one for 2 s each, took a recognition test on 100 faces (50 targets, 50 lures) presented one-by-one, and made confidence ratings on one of the four scales after each recognition decision. They then repeated the same procedure with participants studying a new set of 50 neutral faces and being tested on 100 faces. Thus, altogether, participants studied 100 old and 100 new faces under the same conditions by the end of the experiment. Overall, participants correctly recognized .71 of old faces (hit rate) and falsely recognized .16 of new faces as old (false alarm rate).

For CAC analysis, Tekin and Roediger divided the more granular scales, 20- and 100-points, into 4 and 5 equal bins, and compared them to narrower scales, 4- and 5-points, respectively. As such, for 5 bins, accuracy of 5 on the 5-point scale was contrasted to accuracy of the 17–20 bin on the 20-point scale and the 81–100 bin on the 100-point scale. Similarly, for 4 bins, 3 on the 4-point scale was compared to 11–15 on the 20-point scale and 51–75 on the 100-point scale. Using two separate CAC plots for 4- and 5-point comparisons, Tekin and Roediger found that the 20- and 100-point confidence scales produced similar confidence-accuracy relationships to the narrower scales, especially for the middle to high confidence judgments (see Figures 19.5A and 19.5B). Furthermore, for all ranges of scales, high confidence indicated high accuracy on the recognition test even with 200 faces (.95 hit rate for both the 4- and 5-point comparisons). These findings were in line with previous face recognition studies that demonstrated a positive relationship between confidence (rated on a 100-point scale) and accuracy using calibration plots (Weber & Brewer, 2003, 2004) and extended them by showing the granularity of confidence scale did not affect this relationship.

In old/new recognition tests with equal numbers of targets and lures, calibration and CAC plots are essentially the same, except that the latter require a 100-point confidence scale. However, whereas calibration plots assess how subjective

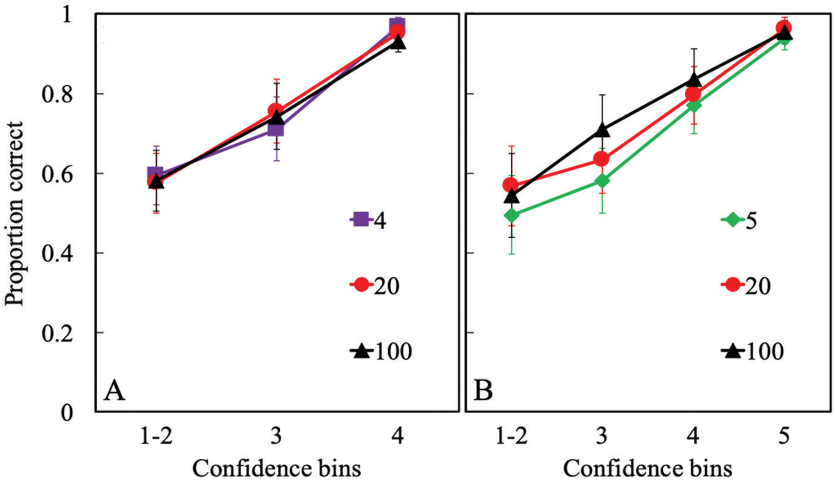


Figure 19.5 CAC plots of 4-point and 5-point comparisons for hits (Experiment 2, Tekin & Roediger, 2017). The first bin in both cases combines ratings of 1 and 2 to increase the number of observations at the low confidence part of the scale. Error bars indicate 95% CI.

probability of confidence on a 100-point scale maps on to objective accuracy, CAC plots assess accuracy as a function of any confidence scale. Therefore, CAC plots allow for comparisons across different confidence scales. As such, by using CAC analysis, Tekin and Roediger demonstrated that the granularity of confidence scales did not affect the confidence-accuracy relationship observed in face recognition. They also replicated their results using word lists, obtaining high CA relationships with 200 words studied and 200 lures. Although researchers rarely use CAC analysis to assess the confidence-accuracy relationship in old/new recognition tests, CAC plots can be beneficial tools for recognition memory.

CAC plots for correct rejections

Thus far, the aforementioned recognition studies have used CAC plots to analyze the confidence-accuracy relationship for *old* responses (hits and false alarms). However, participants also make confidence judgments when they identify a face as *new* during an old/new recognition test. In these situations, participants can either correctly reject an unstudied face (a lure) or incorrectly miss a studied one by claiming it is new (a target). Therefore, it is possible to examine whether high confidence indicates high accuracy for *new* responses, as it does for *old* responses. To answer this question, CAC analysis can be conducted for correct rejections by using the formula, *correct rejections/(correct rejections + misses)* for a given confidence level.

Tekin and Roediger (2017) examined the confidence-accuracy relationship for *new* responses in CAC analysis. They calculated CAC plots for correct rejections and compared narrower 4- and 5-point confidence scales to the more granular 20- and 100-point confidence scales across confidence bins using the binning methodology previously discussed (e.g., 5 on the 5-point scale corresponds to the 17–20 bin on the 20-point scale and the 81–100 bin on the 100-point scale, respectively). Similar to hits, the granularity of confidence scales did not matter in CAC analysis for correct rejections (see Figures 19.6A and 19.6B). In other words, all scale ranges produced similar confidence-accuracy relationships for both *old* and *new* responses. CAC plots demonstrated a critical difference between hits and correct rejections, however. The slope of the confidence-accuracy function was much steeper for hits in comparison to correction rejections, which only showed a slightly upward slope. That is, although accuracy of *new* responses increased as their confidence ratings increased (i.e., positive relationship), this increase was flatter relative to the increase observed for *old* responses. As such, collapsed across 5-, 20-, and 100-point scales, accuracy of the lowest and highest confidence bins were .53 and .95 in the CAC plot for hits, respectively, whereas the same proportions were .62 and .84 in the CAC plot for correct rejections, respectively. These findings suggest that, as opposed to the strong confidence-accuracy relationship obtained for hits, for correction rejections confidence is more weakly related to accuracy (for similar findings using calibration plots, also see Weber & Brewer, 2003, 2004).

Similarly, eyewitness studies have also examined the confidence-accuracy relationship for choosers (those who selected someone from a lineup) and non-

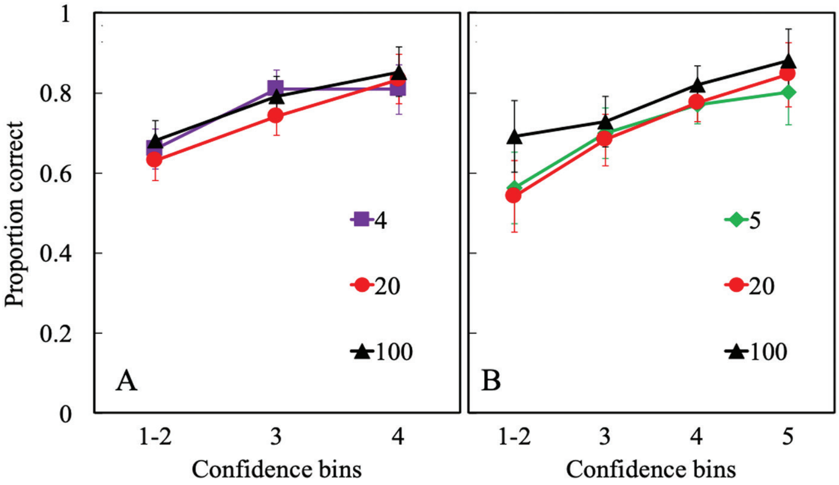
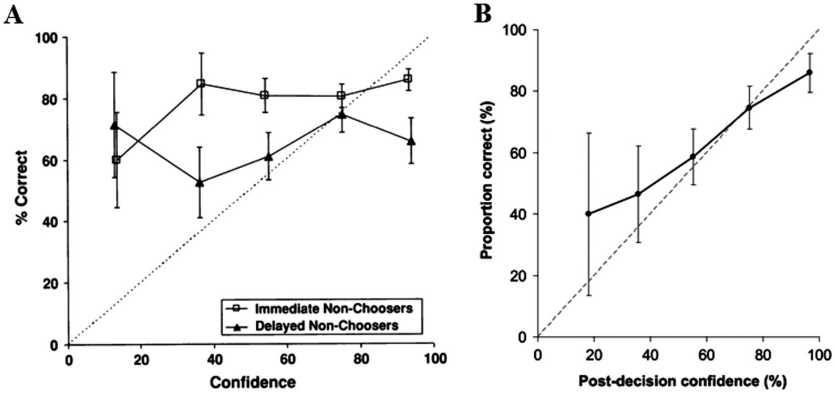


Figure 19.6 CAC plots of 4-point and 5-point comparisons for correct rejections (Experiment 2, Tekin & Roediger, 2017). The first bin in both cases combines ratings of 1 and 2 to increase the number of observations at the low confidence part of the scale. Error bars indicate 95% CI.

choosers (those who rejejcted a lineup) separately. These studies typically show a positive relationship between confidence and accuracy for choosers (e.g., Brewer & Wells, 2006; Sauer et al., 2010; Sporer et al., 1995) but not for nonchoosers (see Figure 19.7A for an example of the confidence-accuracy relationship for nonchoosers). Once again, nonchoosers are the people who rejected the lineup or who said "none of these six people is the suspect." The poor confidence-accuracy relationship for nonchoosers is due, at least in part, to the nature of the lineup identification task itself. Nonchoosers may reject a lineup because (1) they believe the perpetrator is not in the lineup, (2) they believe the perpetrator may be in the lineup but they are not confident enough to choose, or (3) they simply do not know whether or not the perpetrator is in the lineup. Regardless of their reason, nonchoosers are making one identification decision for multiple lineup members (i.e., rejecting all members of a lineup with a single decision). Thus, choosers only have to indicate their confidence for one lineup member, whereas nonchoosers have to provide a single confidence rating for their rejection of multiple lineup members.

A positive confidence-accuracy relationship can be observed for nonchoosers when people are asked to provide a single identification decision and confidence rating for a single target in a show-up procedure (i.e., one-person lineup) or to provide a separate decision and confidence rating for each member from a photo array (Lindsay et al., 2013; Sauerland, et al., 2012). Figure 19.7B shows an example of a positive CA relationship for nonchoosers using the show-up procedure. Thus, the poor confidence and accuracy relation for nonchoosers is partially due to the design of the identification task.



Figures 19.7 (A) shows the nonchooser calibration curves for an eight-person simultaneous lineup (Sauer et al., 2010). (B) shows the nonchooser calibration curve for a show-up procedure (Sauerland et al., 2012).

Plotting CAC functions with misses

In all aforementioned CAC functions, response-based calculations have been used to calculate accuracy. That is, for hits, accuracy at a given confidence level corresponds to the proportion of correct *old* responses at the same confidence level (i.e., using hits and false alarms). Nonetheless, overall accuracy in recognition memory has been frequently calculated item-based. For instance, overall hit rate corresponds to # *hits*/(# *hits* + # *misses*), or the hit rate. Adopting a similar approach, CAC functions can also be plotted using item-based accuracy. In this case, for hits, accuracy at a given confidence level corresponds to the proportion of correct old items that are correctly identified as *old* at the same confidence level (i.e., using hits and misses). Unlike response-based accuracy, where the numbers of hits and false alarms are independent, the numbers of hits and misses are dependent on one another in item-based accuracy (i.e., as the number of hits increases, the number of misses decreases and vice versa). Furthermore, item-based accuracy does not take false alarms into account. Instead, it treats misses as relevant errors and examines whether confidence can predict how accurate the response is, given that the item is old.

Tekin et al. (2021) introduced this novel item-based accuracy calculation, # *hits*/(# *hits* + # *misses*), for CAC plots and compared it to response-based accuracy calculation, # *hits*/(# *hits* + # *false alarms*). They examined whether the two calculations yielded similar confidence-accuracy relationships across various levels of lure relatedness. Here, we only focus on their results regarding face recognition. To address this question, Tekin et al. reanalyzed the face recognition data from Tekin and Roediger (2017), where lures were unrelated (not similar) to targets. That is, although lures were matched to targets in their general characteristics (e.g., if there were 20 young female targets, there were 20 young female lures), the lures were not selected to resemble targets based on any facial similarity index. Because

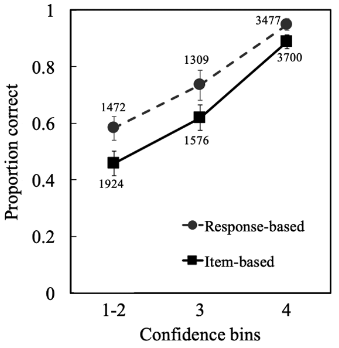


Figure 19.8 Comparison of CAC plots for hits across four confidence bins using either response-based or item-based accuracy. The first bin in both cases combines ratings of 1 and 2 to increase the number of observations at the low confidence part of the scale. Error bars indicate 95% CI.

Tekin and Roediger (2017) demonstrated that granularity of the confidence scales did not influence the confidence-accuracy relationship, Tekin et al. combined the data from 4-, 20-, and 100-point scales and plotted CAC plots using item-based and response-based accuracy. They found that although response-based accuracy yielded a higher CAC function than item-based accuracy, both CAC functions showed similar confidence-accuracy patterns: Confidence was a strong predictor of either type of accuracy (see Figure 19.8). The differences between the two accuracy calculations stemmed from the larger number of misses across all confidence levels relative to the number of false alarms. These findings suggest that regardless of the dependency between hits and misses when the lures are not similar to targets, confidence and accuracy were highly related. It is important to note that the two accuracy calculations can produce different confidence-accuracy relationships when highly related lures are employed, and thus false alarm rates are high (e.g., with categorized lists of words, see Tekin et al., 2021 for more details). Nonetheless, the CAC plots with item-based accuracy can be used in distraction-free recognition experiments (i.e., no lures) to examine the confidence-accuracy relationship. Furthermore, they provide important theoretical insights about recognition memory (for theoretical implications, see Tekin et al., 2021).

Summary and conclusion

Calibration plots and CAC plots have proved useful in the study of eyewitness identification in examining the relation between confidence and accuracy. In situations with fair lineups and adult witnesses, high confidence indicates high accuracy. We have also reported use of CAC plots in experiments with large numbers of items rather than the usual one-item (one suspect) condition of lineup experiments. Even with 100 faces, reports made with high confidence are highly accurate. We also showed how results can differ when one looks at correct rejections in list experiments or in lineup experiments (when a witness rejects the lineup). In this case, CAC plots are relatively flat, except when witnesses judge lineup members one at a time. We finally showed how CAC plots can be created using false alarms as the contrasting measure to hits or correct rejection (the standard way) or with the use of misses to create other CAC plots. The application of CAC plots to issues of recognition memory is just beginning, and we can anticipate exciting developments in the future.

References

Allin, A. (1896). Recognition. *Psychological Review*, *3*(5), 542–545.

Behrman, B. W., & Richards, R. E. (2005). Suspect/foil identification in actual crimes and in the laboratory: A reality monitoring analysis. *Law and Human Behavior*, *29*, 279–301.

Brewer, N., Keast, A., & Rishworth, A. (2002). The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. *Journal of Experimental Psychology: Applied*, *8*(1), 44–56.

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*(1), 11–30.

Brigham, J. C., & Cairns, D. L. (1988). The effect of mugshot inspections on eyewitness identification accuracy 1. *Journal of Applied Social Psychology*, *18*(16), 1394–1410.

Dallenbach, K. M. (1913). The relation of memory error to time interval. *Psychological Review*, *20*(4), 323–337.

DeSoto, K. A., & Roediger, H. L. (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. *Psychological Science*, *25*(3), 781–788.

Dodson, C. S., & Dobolyi, D. G. (2015). Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology*, *30*(1), 113–125.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Sage Publications.

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*(5), 1304–1316.

Lin, W., Strube, M. J., & Roediger, H. L. (2019). The effects of repeated lineups and delay on eyewitness identification. *Cognitive Research: Principles and Implications*, *4*(1), 1–19.

Lindsay, R. C. L., Kalmet, N., Leung, J., Bertrand, M. I., Sauer, J. D., & Sauerland, M. (2013). Confidence and accuracy of lineup selections and rejections: Postdicting rejection accuracy with confidence. *Journal of Applied Research in Memory and Cognition*, *2*(3), 179–184.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, *4*(2), 93–102.

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, *19*(1), 55–71.

Pezdek, K., & Blandon-Gitlin, I. (2005). When is an intervening line-up most likely to affect eyewitness identification accuracy?. *Legal and Criminological Psychology*, *10*(2), 247–263.

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

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*(4), 337–347.

Sauerland, M., Sagana, A., & Sporer, S. L. (2012). Assessing nonchoosers' eyewitness identification accuracy from photographic showups by using confidence and response times. *Law and Human Behavior*, *36*(5), 394–403.

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.

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, *118*(3), 315–327.

Steblay, N. K., & Dysart, J. E. (2016). Repeated eyewitness identification procedures with the same suspect. *Journal of Applied Research in Memory and Cognition*, *5*(3), 284–289.

Steblay, N. K., Tix, R. W., & Benson, S. L. (2013). Double exposure: The effects of repeated identification lineups on eyewitness accuracy. *Applied Cognitive Psychology*, *27*(5), 644–654.

Strong, M. H., & Strong, E. K. (1916). The nature of recognition memory and of the localization of recognitions. *The American Journal of Psychology*, *27*(3), 341–362.

Tekin, E., DeSoto, K. A., Wixted, J. H., & Roediger III, H. L. (2021). Applying confidence accuracy characteristic plots to old/new recognition memory experiments. *Memory*, *29(*4), 427–443.

Tekin, E., Lin, W., & Roediger, H. L. (2018). The relationship between confidence and accuracy with verbal and verbal+ numeric confidence scales. *Cognitive Research: Principles and Implications*, *3*(1), 1–8.

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications*, *2*(1), 1–13.

Valentine, T., Davis, J. P., Memon, A., & Roberts, A. (2012). Live showups and their influence on a subsequent video line-up. *Applied Cognitive Psychology*, *26*(1), 1–23.

Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *Journal of Applied Psychology*, *88*(3), 490–499.

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied*, *10*(3), 156–172.

Wells, W. (2014). The Houston Police Department Eyewitness Identification Experiment: Analysis and results. http://www.lemitonline.org/research/projects.html

Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition*, *4*(1), 8–14.

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 201–233.

Wixted, J.T., Mickes. L., Brewin, C.R. & Andrews, B. (2021). Doing right by the eyewitness evidence: a response to Berkowitz et al. *Memory*, 1-2. DOI: 10.1080/09658211.2021.1940206

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger III, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*, *70*(6), 515–526.

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*(1), 10–65.