MEMORY
2021, VOL. 29, NO. 4, 427–443
https://doi.org/10.1080/09658211.2021.1901937

Routledge
Taylor & Francis Group

Applying confidence accuracy characteristic plots to old/new recognition memory experiments

Eylul Tekin[a], K. Andrew DeSoto[b], John H. Wixted[c] and Henry L. Roediger III[a]

[a]Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA; [b]Association for Psychological Science, Washington, DC, USA; [c]Department of Psychology, University of California San Diego, San Diego, CA, USA

ABSTRACT

Confidence-accuracy characteristic (CAC) plots were developed for use in eyewitness identification experiments, and previous findings show that high confidence indicates high accuracy in all studies of adults with an unbiased lineup. We apply CAC plots to standard *old/new* recognition memory data by calculating response-based and item-based accuracy, one using false alarms and the other using misses. We use both methods to examine the confidence-accuracy relationship for both correct old responses (hits) and new responses (correct rejections). We reanalysed three sets of published data using these methods and show that the method chosen, as well as the relation of lures to targets, determines the confidence-accuracy relation. Using response-based accuracy for hits, high confidence yields quite high accuracy, and this is generally true with the other methods, especially when lures are unrelated to targets. However, when analyzing correct rejections, the relationship between confidence and accuracy is less pronounced. When lures are semantically related to targets, the various CAC plots show different confidence-accuracy relations. The different methods of calculating CAC plots provide a useful tool in analyzing standard *old/new* recognition experiments. The results generally accord with unequal-variance signal detection models of recognition memory.

ARTICLE HISTORY
Received 30 April 2020
Accepted 5 March 2021

KEYWORDS
Confidence-accuracy characteristic plot; confidence; accuracy; recognition memory

In everyday life, when an objective assessment of accuracy is not available, people tend to believe accuracy of a high confidence statement more (e.g., "I am sure I have met this person before") than accuracy of a low confidence statement. In these situations, people intuitively assume that confidence and accuracy have a positive relationship. That is, the more confident a person is about a memory episode, the more accurate she will be. Not surprisingly, the confidence-accuracy relationship has been of interest to scientists who study recognition memory. In a typical recognition experiment, subjects study multiple items (e.g., words, faces, pictures), and take a binary *yes/no* or *old/new* recognition test that consists of presenting both studied items (targets) and non-studied items (lures), one at a time, with instructions to distinguish the studied/old items from the non-studied/new items (e.g., Strong & Strong, 1916). After each recognition decision, subjects may also rate how confident they were in their decision on a confidence scale.

Traditionally, the confidence-accuracy relationship has been examined either by using some type of correlation (Busey et al., 2000; DeSoto & Roediger, 2014; for a review, see Roediger et al., 2012), or by plotting calibration curves (Gigerenzer et al., 1991; Keren, 1991; Lichtenstein et al., 1982; Weber & Brewer, 2003, 2004). The calibration approach consists of plotting proportion correct as a function of confidence measured using a 100-point scale. Perfect calibration exists when decisions made with 100% confidence are 100% correct, decisions made with 90% confidence are 90% correct, and so on. In *old/new* recognition tests, proportion correct corresponds to the proportion of accurate *old* responses across different levels of confidence.

Typically, confidence and accuracy are positively related in recognition memory in laboratory tasks (Mickes et al., 2011; Murdock & Dufty, 1972; Robinson & Johnson, 1996; Weber & Brewer, 2003, 2004). Although calibration studies have found that people show poor calibration (or absolute accuracy) through overconfidence, they have also demonstrated a positive relationship between confidence and accuracy in recognition memory (Gigerenzer et al., 1991; Lichtenstein et al., 1982). That is, although people postdict that they are more accurate than they actually were for higher confidence judgments (i.e., poor calibration), overall, their accuracy increase as their confidence increase (i.e., a positive relationship).

Critically, most of these calibration studies in recognition memory have used n-alternative-forced choice (n-AFC) recognition test (i.e., a multiple-choice test) rather

CONTACT Eylul Tekin elifeylultekin@wustl.edu 1 Brookings Dr, St. Louis, MO 63130

than an *old/new* recognition test, with a few exceptions in the face recognition literature (Weber & Brewer, 2003, 2004). In an n-AFC recognition test, subjects see two or more items simultaneously and select the one they studied. In this test, missing a studied item means selecting a non-studied item, and thus making a false alarm. Similarly, selecting a studied item (or a hit) equates to correctly rejecting a non-studied item. In an *old/new* recognition test, however, these recognition decisions are separate because studied and non-studied items are presented one by one rather than simultaneously as in n-AFC recognition tests. As an example, in an *old/new* recognition test, a subject can miss an old item by selecting *new* and can make a false alarm by stating a new item is *old*. This separation allows researchers to examine recognition decisions about old and new items or *old* and *new* responses separately.

Recently, a related method of analysis, known as confidence-accuracy characteristic (CAC) analysis has been used in the eyewitness identification literature (Mickes, 2015), where a strong relationship between confidence and accuracy is typically observed (Wixted & Wells, 2017). Eyewitness identification experiments typically involve lineups consisting of a suspect (who is old/guilty or new/innocent) and five or more fillers who generally resemble the suspect. For lineups, calibration analysis (computed from responses to both suspects and fillers) differs fundamentally from CAC analysis (computed from responses to suspects only), because calibration plots treat fillers as relevant errors in eyewitness memory, whereas CAC plots do not. However, a CAC plot is essentially the same as a calibration plot when applied to a standard *old/new* recognition memory test because the relevant error for both types of plots is the same (i.e., false alarms). The only difference is that any monotonic rating scale can be used in CAC plots (e.g., plotting accuracy as a function of confidence using a verbal low-, medium-, or high-confidence scale instead of a 100-point scale), whereas only a 100-point scale is used in calibration plots. Although CAC analysis has rarely been used for *old/new* recognition tasks, in a study testing recognition memory for lists of words or faces, Tekin and Roediger (2017) used CAC analysis with a wide range of confidence rating scales, including 4-, 5-, 20-, and 100-point scales and found that the scales were essentially convertible from one to the other. Moreover, the relationship between confidence and accuracy was strong, and high confidence was indicative of high accuracy even with hundreds of items.

The main purpose of the current study was to further examine the confidence-accuracy relationship in *old/new* recognition experiments using CAC plots. We employed CAC analysis in another domain, namely, when the relatedness of lures differed across experiments. We examined how the similarity of lures to target items influenced the strong relation between confidence and accuracy observed in CAC plots. The similarity of lures is known to affect recognition memory performance as well as the confidence-accuracy relationship (e.g., Benjamin & Bawa, 2004; DeSoto & Roediger, 2014; Tulving, 1981). For instance, by employing lures related to target scenes, Tulving (1981) demonstrated an inverse relationship between average confidence and accuracy in a 2-AFC recognition test (also see, Chandler, 1994; Dobbins et al., 1998). DeSoto and Roediger (2014) also reported a negative confidence-accuracy relationship in recognition memory in a study involving similar lures. Subjects studied 10 words from 12 semantic categories (i.e., 120 words total) and took an *old/new* recognition test on 120 targets, 120 unrelated lures, and 120 related lures, which consisted of 10 unstudied words from the same 12 categories. For related lures, the confidence-accuracy relationship was *negative* when plotted across items, meaning that the more confident subjects were in a saying old to a lure item, the less accurate they were in their recognition decision. These findings suggested that the similarity of lures affected the typical positive confidence-accuracy relationship in recognition memory.

Findings like these have been interpreted to weigh against signal detection theory (SDT), which inherently predicts a strong—and positive—confidence-accuracy relationship (Wixted, 2020). But are they as inconsistent with SDT as they appear to be at first glance? In part to also address this question, we reanalysed data from three previous *old/new* recognition experiments that had varying levels of lure relatedness using CAC plots. The experiments consisted of the following recognition tests: (1) a recognition test with unrelated lures (Tekin & Roediger, 2017); (2) a recognition test with related lures (Tekin & Roediger, 2017); and (3) a recognition test with both related and unrelated lures (DeSoto & Roediger, 2014).

In conducting these reanalyses, we describe two different accuracy measures, a typical and a novel one, that can be used in CAC plots in any *old/new* recognition experiments, which provide answers to different questions about the confidence-accuracy relationship. These accuracy measures can be calculated separately for correct responses for old items (hits) and correct responses for new items (correct rejections). Typically, in an *old/new* recognition experiment, at a given level of confidence (e.g., 60–80%), proportion correct in a CAC plot is calculated using the formula *# hits / (# hits + # false alarms)* with only hits and false alarms that receive confidence ratings of 60–80% (in this example). This measure represents the proportion of correct *old* responses when subjects identify items as *old* with that level of confidence (i.e., response-based accuracy). As noted above, the confidence-accuracy relationship is typically strong for *old* responses using response-based accuracy (Tekin & Roediger, 2017).

CAC analysis is less commonly applied to *new* responses, but those decisions are theoretically interesting as well. Correct rejections represent correct *new* responses in *old/new* recognition experiments, although they are rarely of primary interest. Critically, confidence tends to be less predictive at explaining accuracy for correct

rejections (Kantner & Dobbins, 2019). Misses (incorrect *new* decisions) are also important errors in *old/new* recognition experiments because the failure to recognise old items (especially with high confidence) seem odd: How can it be that subjects judge an item studied just ten minutes or so before and say with high confidence that they have not seen it in the list (see Roediger & Tekin, 2020 for more on this issue). Some theories hold that high-confidence *new* responses are not actually meaningful. For example, the low-threshold theory assumes that *new* responses are made to targets and lures that fail to exceed a threshold of conscious awareness (Luce, 1963). Such below-threshold items are theoretically associated with no memory signal whatsoever. If *new* responses are untethered to an underlying memory signal, then confidence in a *new* recognition decision would not be meaningful. Thus, the natural prediction would be that the CAC plot using *new* responses (i.e., correct rejections and misses) should be flat.

In the current study, we also explored novel variants of CAC analysis by calculating proportion correct based on *item* accuracy both for targets and lures. For targets, the measure corresponds to *# hits / (# hits + # misses)* for a given level of confidence. This calculation translates to the proportion of old items that are correctly identified as *old* for a given level of confidence and would correspond to the overall hit rate if collapsed across confidence levels. The only difference between response-based and item-based accuracy for targets is whether false alarms or misses are used with hits in the denominator. Critically, these two CAC analyses answer different questions. Using response-based accuracy for hits, the question is how likely the response is to be correct given that the response was *old*, whereas using item-based accuracy for hits, the question is how likely the response is to be correct given that the test item is old.

For lures, CAC plots with correct rejections can also be created using the number of correct rejections in the numerator for each confidence bin divided by the number of correct rejections plus either the number of misses (response-based) or the number of false alarms (item-based) in that confidence bin. The latter calculation translates to the proportion of new items that are correctly identified as *new* for a given level of confidence and would correspond to the overall correct rejection rate if we collapse across confidence levels. Using response-based accuracy for correct rejections, the question is how likely the response is to be correct given that the response was *new*, whereas using item-based accuracy for correct rejections, the question is how likely the response is to be correct given that the test item is new.

This new item-based approach to CAC analysis does not provide information that is independent of the information provided by standard CAC analysis, but it does highlight certain theoretical notions that are worthy of consideration. For example, intuition suggests that because unrelated lures were not presented on a recent list, and because they are (by definition) unrelated to the items that were presented on the list, they will fail to generate a memory signal. After all, where would the memory signal come from? Indeed, this was an explicit assumption of the now outdated high-threshold model. If unrelated lures do not generate a memory signal, then it follows that the item-based CAC plot for correct rejections should be flat. Table 1 shows four different formulas we used to calculate CAC plots in an *old/new* recognition test for hits and correct rejections, one based on accuracy of the *old/new* response (response-accuracy) and the other based on the nature of items correctly responded to (item-accuracy).

Table 1. Recognition Decisions and CAC Plot Calculations in Recognition Memory.

		Response		
		"Old"	"New"	Item-Based Accuracy
Item Type	Old	Hit (H)	Miss (M)	Hs/(Hs + Ms) *(b)*
	New	False Alarm (FA)	Correct Rejection (CR)	CRs/(CRs + FAs) *(d)*
Response-Based Accuracy		Hs/(Hs + FAs) *(a)*	CRs/(CRs + Ms) *(c)*	

Current research

To reiterate, the main aim of the current study was to examine the confidence-accuracy relationship in *old/new* recognition tests with hundreds of items using CAC plots. In three *old/new* recognition experiments, we examined the confidence-accuracy relationship for hits and correct rejections by comparing response-based and item-based accuracy calculations of CAC plots across different levels of lure relatedness. The item-based analysis is new for examining the confidence-accuracy relationship in *old/new* recognition experiments, because it uses misses as errors for hits and false alarms as errors for correct rejections. Therefore, we also explored whether similar confidence-accuracy relationships were obtained for hits and correct rejections across CAC plots when using the two accuracy measures. Experiment 1 served as a baseline experiment to demonstrate the confidence-accuracy relationship in recognition memory where lures were unrelated, whereas Experiment 2 used related lures during the test, and Experiment 3 consisted of both related and unrelated lures. Using CAC plots to analyze these experiments, we asked: (1) Is confidence strongly related to accuracy, and does high confidence indicate high accuracy in recognition memory?; (2) Does the confidence-accuracy relationship remain strong regardless of lure relatedness?; (3) Do correct rejections yield a similar confidence-accuracy relationship as hits?; (4) Do the two methods of calculating accuracy in CAC plots affect the obtained confidence-accuracy relationship?

Based on eyewitness studies (Wixted & Wells, 2017) and initial findings in recognition memory (Tekin & Roediger, 2017), we hypothesised high confidence hits to be highly accurate in *old/new* recognition memory, and thus strong confidence-accuracy relationships in all experiments using CAC plots. We also predicted lure relatedness to have a small effect on the confidence-accuracy relationships reported in CAC plots. Based on prior evidence (Weber & Brewer, 2003, 2004), we hypothesised the confidence-accuracy relationship observed for hits and correct rejections in CAC plots to differ from one another, with correct rejections showing a weaker but still positive confidence-accuracy relationship than hits. Lastly, we expected the two different accuracy calculations of CAC plots to show similar confidence-accuracy relationships when the lures were unrelated. However, we predicted the two methods to produce a different pattern when lures were related because the similarity of lures (related or unrelated) directly affects false alarm rates and confidence in false alarms in recognition memory. That is, people tend to false alarm more to related lures than unrelated ones and they also assign higher confidence ratings to related lures than unrelated lures (Benjamin & Bawa, 2004; DeSoto & Roediger, 2014). Thus, we hypothesised that the similarity of lures would especially affect CAC plots for hits when the plots are based on response-accuracy and yet to have little effect on CAC plots based on item-accuracy. The different CAC analyses to be reported in the bulk of our paper are, admittedly, empirical and exploratory. Therefore, we also applied SDT to our results to see how compatible SDT is with the reported findings, in an ex post facto analysis.

These experiments have been previously published, but our analyses are either mostly new (Experiments 1 and 2; Tekin & Roediger, 2017) or completely new (Experiment 3; DeSoto & Roediger, 2014). We are calling them Experiments 1, 2, and 3 here for purposes of exposition. Experiments 1 and 2 here from Tekin and Roediger (2017) have their order reversed from the original report, whereas Experiment 3 was reported as Experiment 1 in DeSoto and Roediger (2014). For all three experiments, we report key features of the method needed to understand the current analyses; fuller descriptions are available in the original publications.

Experiments 1 and 2

Experiments 1 and 2 were both from Tekin and Roediger (2017) in which the original interest was to compare effects of different ranges of confidence scale (4-, 5-, 20- and 100-points) on the confidence-accuracy relationship. The range of confidence scale was manipulated between-subjects. Both experiments consisted of two identical phases with different material sets in which subjects studied the material and then took a recognition test with item-by-item confidence judgments. Tekin and Roediger employed CAC plots for the comparisons in which the wider scales, 20- and 100-points, were divided into four and five equal confidence bins and were compared to the 4- and 5-point scales, respectively. For example, the confidence bin of 4 on the 4-point scale was compared to the confidence bins of 16–20 on the 20-point scale, and 76–100 on the 100-point scale. These CAC plots used response-based accuracy (Table 1(a) for hits and Table 1(c) correct rejections). The results revealed no accuracy difference amongst different confidence scale ranges for hits and correct rejections, especially for higher confidence bins (e.g., 4 on the 4-point scale, 16–20 on the 20-point scale, and 76–100 on the 100-point scale), suggesting that the confidence-accuracy relationship was similar regardless of the range of the confidence scale. We report those data again here, briefly, to make a direct comparison between different accuracy calculations. For the current report (and given that the scales did not differ), we increased power by combining data from 4-, 20- and 100-point scales, and we binned them into four separate confidence bins (e.g., 1 for the 4-point scale, 1–5 for the 20-point scale, and 1–25 for the 100-point scale). We further combined the lowest two confidence bins (i.e., 1 and 2 for the 4-point scale, 1–10 for the 20-point scale, and 1–50 for the 100-point scale) due to the small numbers of observations in these bins. The final three bins are labelled as 1–2, 3, and 4, corresponding to each bin from the 4-, 20- and 100-point scales.

Experiment 1

In Experiment 1, subjects studied two sets of 50 faces each and took two *old/new* recognition tests with 100 faces, making a confidence judgment after each decision. The lure faces were not related to the target faces (i.e., we did not deliberately choose lures that resemble targets on facial similarity). Thus, Experiment 1 served as a recognition experiment with unrelated lures.

Method

Subjects were 72 undergraduate students from Washington University who participated either for payment or course credit. Two hundred neutral faces were selected from Minear and Park's (2004) database, 100 females and 100 males, from 19 to 50 years of age. Faces were counterbalanced across study and test phases. Thus, all faces served as both targets and lures across subjects. The experiment consisted of two study-test phases, and in each phase, subjects saw 50 faces one by one for 2 s, performed a 10-min filler task, and then took an *old/new* recognition test. The test included 100 items, 50 old (studied) and 50 new (non-studied). The set of lures on the test matched the target set for general features of age and race, but no matching occurred for facial similarity (i.e., the lures were not related to the targets). After deciding *old* or *new*, subjects made a confidence rating on a 4-point, 20-point, or 100-point confidence scale. The highest point

on the scale indicated *totally confident* and the lowest point indicated *not confident at all.* The recognition test was self-paced and subjects typed in a number to indicate confidence. After testing on the first 100 faces, a short break ensued and the subjects studied another 50 faces, completed a filler task, and then were given a second 100-item recognition test (half old and half new) and made confidence judgments on the same scale for each decision. Thus, across the two sets of materials, subjects studied 100 faces and were tested on 200.

Results

All of the results sections follow the same organisation to explain our analyses: We first provide the analyses for hits (correct *old* responses, Table 1(a,b)) and then the analyses for correct rejections (correct *new* responses, Table 1(c) and 1(d)) and compare the two methods of accuracy calculations in CAC plots (response-based and item-based accuracy) for each type of correct response. Within each section, we indicate the accuracy measure by indicating response-based or item-based.

CAC plots with hits

Figure 1(a) provides CAC plots for correct *old* responses (hits) to targets calculated using response-based accuracy (false alarms as errors, Table 1(a), and item-based accuracy (misses as errors, Table 1(b)). Both methods show a similar pattern: As confidence increased, so did proportion correct, but the CAC plot using response-based accuracy is higher than the CAC plot using item-based accuracy as shown in Figure 1(a). A 3 (confidence bins) x 2 (calculation type) repeated measures ANOVA, with both variables as within-subjects factors, revealed a main effect of confidence, $F(2,140) = 195.14$, $p < .001$, $\eta_p^2 = .74$, and a main effect of calculation type, $F(1,70) = 36.86$, $p < .001$, $\eta_p^2 = .35$. Overall, the calculation using response-based accuracy ($M = .77$, $SE = .01$) demonstrated higher proportion correct than the calculation using item-based accuracy ($M = .66$, $SE = .02$). Interestingly, the interaction was also reliable, $F(2,140) = 4.80$, $p = .010$, $\eta_p^2 = .06$. The calculation using response-based accuracy showed higher proportion correct than the calculation using item-based accuracy across all confidence bins ($ps < .001$); however, this difference decreased across confidence bins (.13, .11, and .06, respectively). The interaction was probably driven in part by a ceiling effect in the highest bins ($M = .95$, $SE = .01$; $M = .89$, $SE = .01$, respectively). Despite the differences in the two calculations, both CAC plots indicate that high confidence is associated with high accuracy.

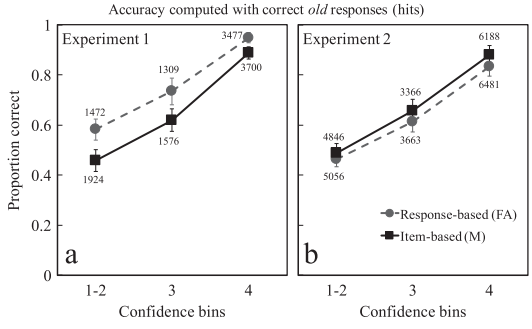

Figure 1. Comparison of CAC plots for accuracy computed with correct *old* responses (hits) across four confidence bins using either false alarms or misses in the denominator of the CAC plot. (a) Experiment 1 with unrelated lures. (b) Experiment 2 with related lures. Error bars indicate 95% confidence intervals. Response-based accuracy was calculated as hits/(hits + FAs) and item-based accuracy was calculated as hits/(hits + misses).

CAC plots with correct rejections

In all of these experiments, when subjects rejected test items by choosing *new*, they also provided confidence judgments. Thus, we also plotted the relationship between confidence and accuracy in two additional ways, with correct *new* responses to lures (correct rejections) in the numerator and with either incorrect *new* responses to targets (misses) or incorrect *old* responses to lures (false alarms) in the denominator. These correspond to response-based, (Table 1(c)) and item-based accuracy (Table 1d), respectively.

Figure 2(a) provides the CAC plots for correct rejections. There are two clear differences between these CAC plots and the ones for hits (Figure 1(a)). First, these CAC plots for correct rejections are noticeably flatter than the CAC plots for hits. Nonetheless, the same trend between confidence and accuracy emerged; as confidence increased, so did proportion correct. Second, the CAC function was higher when item-based accuracy was used relative to response-based accuracy, a flipped pattern compared to Figure 1(a). A 3 (confidence bins) x 2 (calculation type) repeated measures ANOVA confirmed both of these observations. There was a main effect of confidence, $F(1.78,124.87) = 89.90$, $p < .001$, $\eta_p^2 = .56$, and a main effect of calculation type, $F(1,70) = 25.85$, $p < .001$, $\eta_p^2 = .27$. Overall, the CAC plot using item-based accuracy ($M = .84$, $SE = .01$) produced a higher proportion correct than the plot using response-based accuracy ($M = .76$, $SE = .01$). The interaction was not reliable, $F(1.79,125.12) = 2.75$, $p = .074$, $\eta_p^2 = .04$.

Discussion

The results of Experiment 1 indicate that in CAC plots for hits, using response-based accuracy yielded greater accuracy than using item-based accuracy, whereas in CAC plots for correct rejections, using item-based accuracy produced higher overall accuracy. In both these calculations, false alarms were used as recognition errors rather than misses. Why do the two accuracy methods differ? The top section of Table 2 provides the frequency of the four types of recognition decisions and helps to explain why the difference occurred. In Experiment 1, with unrelated faces as lures, the number of misses was much higher than the number of false alarms in all confidence bins. In other words, during the recognition test subjects did not

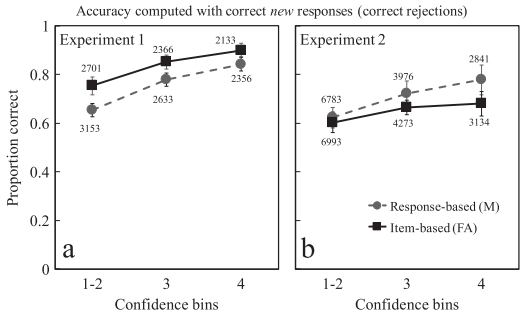

Figure 2. Comparison of CAC plots for accuracy computed with correct *old* responses (correct rejections) across four confidence bins using either false alarms or misses in the denominator of the CAC plot. (a) Experiment 1 with unrelated lures. (b) Experiment 2 with related lures. Error bars indicate 95% confidence intervals. Response-based accuracy was calculated as CRs/(CRs + misses) and item-based accuracy was calculated as CRs/(CRs + FAs).

false alarm to unrelated lures as often as they missed identifying the targets as *old*. Thus, for hits and correct rejections, the CAC plots using item-based accuracy and response-based accuracy, respectively, yielded lower proportions correct. In other words, the accuracy calculations using misses as errors yielded lower proportions correct compared to the accuracy calculations using false alarms because false alarms were fewer in number than misses.

Why did subjects commit more misses than false alarms? The top section of Table 3 shows that overall, subjects had conservative response criteria in Experiment 1, making them less likely to declare a test item to be *old* during the recognition test. More conservative responding decreases the hit rate while increasing the miss rate. At the same time, the false alarm rate decreases while the correct rejection rate increases. This in turn explains the higher number of misses relative to false alarms in Experiment 1. We can only speculate on why subjects responded more conservatively. Previous research has shown that instructing subjects to emphasise one response over the other, asking them to be more or less cautious in saying *old,* offering more reward for one type of response, or even implicitly changing the emphasis on a particular response can affect response criterion (Azimian-Faridani & Wilding, 2006; Han et al., 2010; Mill & O'Connor, 2014; Van Zandt, 2000). We did not explicitly use such manipulations and the test instructions we used were neutral: "Some of the faces will be from the faces you studied before and some will be new faces. For each face, please indicate whether you recognise the face or not. If you have studied the face before, click OLD. If you have not studied the face before, click NEW." Nonetheless, the order of the *old/new* instructions (i.e., the *old* decision is mentioned before the *new* decision) might have still implicitly affected response criterion. Furthermore, although some studies found that complex and rich stimuli such as paintings consistently yield conservative responses (Lindsay & Kantner, 2011), no evidence that suggests that a similar pattern was observed for neutral faces.

Regardless of the response criteria, both for hits and correct rejections, the CAC plots using response-based and item-based accuracy showed similar results, with greater confidence indicating greater accuracy. The similarity in CAC results for hits occurs despite the fact that for a given subject the miss rate is dependent on the hit rate (i.e., hits plus misses must add up to 100% across all confidence bins), whereas the number of false alarms is independent of the number of hits (i.e., the number of non-studied items out of 100 that received *old* responses). For CAC plots with correct rejections, this case is reversed: The correct rejection and false alarm rates are constrained to sum to 100%, but now the miss rate can vary. We consider the theoretical implications of our findings below in the General Discussion. We next compare the different kinds of CAC plots when related lures are used instead of unrelated lures.

Table 2. Number of Observations and Percentages for Hits, Misses and False Alarms for Experiment 1(top section) 2 (middle section) and 3 (bottom section).

	Experiment 1: Unrelated lures						
Confidence	1–2		3		4		Total
Response	n	%	n	%	n	%	n
Hit	868	17.0	957	18.7	3288	64.3	5113
Miss	1056	50.6	619	29.7	412	19.7	2087
FA	604	52.8	352	30.7	189	16.5	1145
CR	2097	34.6	2014	33.3	1944	32.1	6055
	Experiment 2: Related lures						
Confidence	1–2		3		4		Total
Response	n	%	n	%	n	%	n
Hit	2425	23.9	2203	21.7	5505	54.3	10133
Miss	2421	56.7	1163	27.3	683	16.0	4267
FA	2631	51.9	1460	28.8	976	19.3	5067
CR	4362	46.7	2813	30.1	2158	23.1	9333
	Experiment 3: Unrelated and related lures						
Confidence	1–2		3		4		Total
Response	n	%	n	%	n	%	n
Hit	524	13.6	482	12.6	2834	73.8	3840
Miss	935	64.9	275	19.1	230	16.0	1440
Unrelated FA	316	57.8	116	21.2	115	21.0	547
Related FA	777	38.1	598	29.4	662	32.5	2037
Unrelated CR	1867	39.4	967	20.4	1899	40.1	4733
Related CR	1623	50.0	792	24.4	828	25.5	3243

FA indicates false alarms. CR indicates correct rejections. n stands for number of observations. Percentages refer to percentage of responses in a particular confidence bin. Percentages should not be compared across confidence bins due to widely different numbers of observations.

Experiment 2

To recap, Experiment 1 indicated that the two types of CAC plots with hits revealed similar CAC functions for correct *old* responses, although the function based on response-accuracy was higher than the function based on item-accuracy. The reverse was true for CAC plots for correct rejections: The CAC function using item-based accuracy was higher than the function using response-accuracy. Both of these findings were due to their being fewer false alarms than misses, caused by subjects' conservative placement of their response criteria. In Experiment 1, the set of lures was not created to be particularly similar to

Table 3. Hit rates, false alarm rates, d-prime and response criterion scores for Experiment 1(top section) 2 (middle section) and 3 (bottom section).

	Experiment 1: Unrelated lures			
	Hit	FA	*d′*	*C*
	0.71	0.16	1.69	0.25
	Experiment 2: Related lures			
	Hit	FA	*d′*	*C*
	0.70	0.35	1.02	−0.08
	Experiment 3: Unrelated and related lures			
Lures	Hit	FA	*d′*	*C*
Total	0.73	0.24	1.42	0.03
Unrelated	0.73	0.10	2.06	0.40
Related	0.73	0.39	1.00	−0.18

FA indicates false alarm rates. *d′* indicates discrimination. *C* indicates response criteria.

the targets or to be "deceptive" as others have called related lures (Brewer et al., 2005; Koriat, 2008).

In Experiment 2, we tested subjects on recognition memory for words, and the lures were primary associates of the targets (e.g., study *chair* and have *table* as a lure or vice versa). In a recognition experiment with strongly related lures, we expected the false alarm rate to be relatively high and the correct rejection rate correspondingly lower. Given these conditions, we expected that CAC plots using the same four methods of calculation would differ in predictable ways compared to the results of Experiment 1. In particular, we predicted that false alarms would outnumber misses due to higher similarity of lures and a more lenient criterion placement, and thus when CAC plots are constructed for hits, response-based accuracy they will be lower than when constructed item-based accuracy. Likewise, the CAC plots for correct rejections should flip relative to performance of Experiment 1 with unrelated lures. Of course, another major difference between Experiments 1 and 2 besides relatedness of lures is the stimulus set used, namely, pictures versus words. However, we replicated the results of Experiment 1 using words in Experiment 3, so we believe that the relatedness of lures is the critical variable and not the type of stimuli.

Method

Seventy-two undergraduates from Washington University participated for either payment or course credit. Two-hundred associated word pairs (400 words) were selected from the Nelson et al. (2004) norms, with all associates being one of three primary associates of the target word (e.g., subjects studied *knee,* and *leg* served as the lure in the recognition test, although all items on the test were presented in random order so the two might be widely separated in their presentation). The two-item sets were counterbalanced across study and test phases, such that all words served equally often as targets and lures across subjects. As in Experiment 1, subjects received two study lists and two tests in sequence, with a 10-min filler task between the study and test for each list. Each study list consisted of 100 target words and the recognition test of 200 words (targets and their associates as related lures); thus, across two lists, subjects studied 200 words and were tested on 400 words, making a confidence rating after each decision. Critically, unlike in Experiment 1, the lures were highly related to the targets. The presentation rate and other characteristics of Experiment 2 were the same as in Experiment 1, except for the different types of material.

Results

CAC plots with hits

The results of Experiment 2 for correct *old* responses are shown in Figure 1(b), and the CAC plots replicate the confidence-accuracy relationship from Experiment 1 shown in Figure 1(a): a strong relationship between confidence and accuracy is apparent, and, once again, high confidence indicates high accuracy (even though subjects were tested on 400 words across two lists). However, unlike in Experiment 1, the CAC plot using item-based accuracy produced a slightly, but reliably, elevated function relative to the CAC plot with response-based accuracy. A 3 (confidence bins) x 2 (calculation type) ANOVA revealed a main effect of confidence bins, $F(1.80,127.57) = 179.38$, $p < .001$, $\eta_p^2 = .72$, and a main effect of calculation type, $F(1,71) = 6.17$, $p = .015$, $\eta_p^2 = .08$. Overall, using item-based accuracy in the CAC plot ($M = .67$, $SE = .02$) produced greater proportion correct than the plot with response-based accuracy ($M = .64$, $SE = .01$). The interaction was not reliable, $F < 1$. At the highest level of confidence, accuracy was .88 and .84 for the item-based and response-based CAC plots, respectively. These values seem quite high given that altogether subjects studied 200 targets and the lures were highly related to the targets. Despite these features, average accuracy at the highest level of confidence was .86 across the two types of measures.

CAC plots with correct rejections

The results of Experiment 2 for correct *new* responses are shown in Figure 2(b). As in Experiment 1, the CAC functions for correct rejections are flatter than those for hits (Figure 1(b)). In addition, as with the results for hits, the CAC plots for correct rejections are reversed compared to the ones from Experiment 1 (Figure 2(a)): The CAC plot constructed with response-based accuracy displayed a slightly higher function than the CAC plot using item-based accuracy (Figure 2(b)). A 3 (confidence bins) × 2 (calculation type) repeated measures ANOVA revealed a main effect of confidence, $F(1.38,95.39) = 22.11$, $p < .001$, $\eta_p^2 = .24$, and a main effect of calculation type, $F(1,69) = 14.04$, $p < .001$, $\eta_p^2 = .17$. The interaction was also reliable, $F(1.76,121.46) = 5.21$, $p = .009$, $\eta_p^2 = .07$. Overall, the response-based accuracy calculation ($M = .71$, $SE = .02$) revealed higher proportion correct than the item-based accuracy calculation ($M = .65$, $SE = .02$), but the differences

were only significant at the two highest two confidence bins (3 and 4, *p*s < .01).

Discussion

The CAC plots of Experiment 2 were as predicted, based on the assumption that the false alarm rates would be higher than the miss rates because of the strongly related lures. The middle section of Table 2 shows that the number of false alarms was greater than the number of misses in Experiment 2. Subjects provided more false alarms to related lures than they missed identifying old items at every level of confidence in Experiment 2, contrary to the pattern observed in Experiment 1. Thus, for hits, CAC plots using item-based accuracy yielded higher proportions correct compared to CAC plots using response-based accuracy because false alarms were greater in number than misses. For correct rejections, the opposite pattern emerged because of the very same reason: CAC plots using response-based accuracy yielded higher proportions correct compared to CAC plots using item-based accuracy due to more false alarms. The high false alarm rate in Experiment 2 can be explained by more difficult discrimination between targets and lures and a more liberal response criterion than in Experiment 1. The middle section of Table 3 confirms that relative to Experiment 1, subjects were less able to differentiate between targets and lures in Experiment 2, most likely due to the relatedness of lures. Furthermore, subjects seemed to exhibit slightly more liberal response criteria, increasing their likelihood of responding *old* overall. However, what seems to be more liberal response criteria in Experiment 2 might rather be due to the rightward shift in the lure distribution (from relatedness) even if the criterion did not change across experiments.

Nonetheless, as with Experiment 1, the CAC plots using both accuracy measures show similar results. Of course, the total number of observations was much greater in Experiment 2, because the number of items studied and tested doubled relative to Experiment 1. The difference in false alarm and miss rates between experiments probably occurred because the lures in Experiment 2 were related and those in Experiment 1 were not; however, the type of material – faces or words – differed between experiments, too, as did the total number of items subjects studied and were tested on. The change in the relative number of misses and false alarms in Experiment 2 relative to Experiment 1 created the corresponding differences in CAC plots. The differences were relatively small, especially for hit CAC plots, but were consistent across confidence bins except for the lowest confidence bin in Figure 2(b).

Experiment 3

Experiment 1 employed a recognition test with unrelated lures, whereas in Experiment 2 the lures were highly related to the targets. We obtained different CAC plots using item-based and response-based accuracy in the two experiments, both for hits and correct rejections. In Experiment 3, we investigated whether both patterns would be obtained within the same experiment if both related and unrelated lures were used. Once again, a published experiment by DeSoto and Roediger (2014, Experiment 1) permitted us to perform these novel analyses. In DeSoto and Roediger, subjects studied words belonging to common semantic categories (e.g., birds) and then took a recognition test that included targets, related lures (other category members), and unrelated lures (words from categories not used in the study phase). Subjects made *old/new* responses and then reported accuracy on a 100-point scale. The authors reported positive, null, and negative correlations between confidence and accuracy within this single paradigm, computing correlations between confidence and accuracy using items as the unit of analysis. In the following section, we reanalyzed their data using the four different methods of computing CAC plots; no CAC plots were reported in the original paper. As with Experiment 1 and 2, the confidence ratings were binned into four bins, and the lowest two bins were combined.

Method

Forty-four undergraduates from Washington University participated for either payment or course credit. The items used in the experiment were the 20 most frequent words from 12 categories (240 items in total), selected from the Van Overschelde et al. (2004) category norms. For each category, subjects studied 10 items (either the even or odd items from the norms) and the complementary set of 10 items served as related lures on the recognition test. The two groups of items were counterbalanced, and thus all words served as both targets and lures across subjects. Another unrelated 120 words drawn from non-studied categories served as unrelated lures for all subjects.

During the study phase, subjects first heard a category name and then heard each member of that category for 2 s, until they had heard all 12 category names and, altogether, 120 items. They then proceeded to a filler task that lasted approximately 10 min. Finally, they were given an *old/new* recognition test on 360 items (120 targets, 120 related lures, and 120 unrelated lures), and gave item-by-item confidence judgments on a 100-point scale with 100 being the highest confidence possible. Items were presented in random order on the test. This experiment permits us to examine CAC plots in all four ways using a within-subjects design with both unrelated and unrelated lures. We sought to replicate the results of the first two experiments in a single within-subject experiment.

Results

CAC plots with hits

As in the first two experiments, the confidence ratings were first combined into three bins for the CAC plots (1–50, 51–75, and 76–100, to remain consistent across experiments, we labelled these bins as 1–2, 3 and 4, respectively). We used these three bins because *old/new* judgments made with low confidence were less frequent than those made with high confidence. The design of this experiment permitted us to calculate response-based accuracy in CAC plots for hits using both related and unrelated false alarms as errors, separately. Of course, for item-based accuracy (i.e., when the method used misses as errors in its denominator), the CAC plots stayed the same across the unrelated/related lure dimension, because these CAC plots ignored responding with lures (i.e., false alarms) altogether. Thus, the main contrast of interest involved the comparisons using different false alarm rates that arose from the different types of lures (the 120 unrelated lures and the 120 related lures). For each comparison, we conducted a 3 (confidence bins) x 2 (calculation type) within-subjects ANOVA. Results of the two comparisons of interest can be observed in Figure 3, which we consider in turn.

Figure 3(a) provides a CAC plot using response-based accuracy with false alarm rates for unrelated lures, and Figure 3(b) shows the same type of plot but with related lures (other members of the category). To repeat, the CAC function using item-based accuracy is the same in both Figure 3(a,b). Figure 3(a) shows that the CAC plot using response-based accuracy is much higher than that using item-based accuracy. This outcome provides a replication of the effect of calculating CAC plots using unrelated lures across an entirely different set of materials (faces in Experiment 1, categorised words in Experiment 3). The strongly positive relationship between confidence and accuracy was again observed, $F(1.49,61.10) = 71.38$, $p < .001$, $\eta_p^2 = .64$, as was the main effect of calculation type (using false alarms or misses in the denominator of the CAC plot). This pattern also replicates the results of Experiment 1 in that using response-based accuracy led to an overall higher proportion correct ($M = .83$, $SE = .02$) than did the CAC plot using item-based accuracy ($M = .66$, $SE = .02$), $F(1,41) = 61.18$, $p < .001$, $\eta_p^2 = .60$. Further, the interaction was reliable, $F(1.72,70.48) = 25.01$, $p < .001$, $\eta_p^2 = .38$ and, as in Experiment 1, probably driven by a ceiling effect for high confidence responses. Pairwise comparisons revealed that the response-based calculation produced a greater proportion correct in the lower two confidence bins than the item-based CAC calculation (*p*s < .001); however, for the highest bin, this difference was not significant with a difference of only .03, $p = .057$, ($M = .95$, $SE = .01$, $M = .92$, $SE = .02$, for the response-based and item-based calculations, respectively).

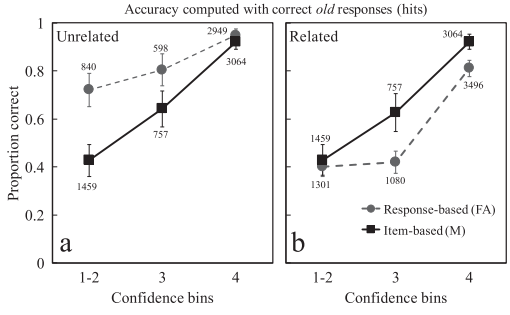

Figure 3 .#Comparison of CAC plots for accuracy computed with correct *old* responses (hits) across four confidence bins using either false alarms or misses in the denominator of the CAC plot. (a) Experiment 3 with unrelated lures. (b) Experiment 3 with related lures. Error bars indicate 95% confidence intervals. Response-based accuracy was calculated as hits/(hits + FAs) and item-based accuracy was calculated as hits/(hits + misses).

Figure 3(b) shows the comparison between the hit CAC plots using either response-based accuracy with related lures or item-based accuracy. Now the CAC plots are like those of Experiment 2 (Figure 1(b)). The CAC plot using item-based accuracy is higher than the one using response-based accuracy. As usual, we obtained a main effect of confidence bins, $F(1.75,73.33) = 127.95$, $p < .001$, $\eta_p^2 = .75$, and a main effect of the calculation type, $F(1,42) = 43.76$, $p < .001$, $\eta_p^2 = .51$. The calculation using item-based accuracy ($M = .66$, $SE = .02$) now revealed a greater proportion correct than the calculation using response-based accuracy ($M = .54$, $SE = .01$). The interaction was again reliable, $F(1.57,66.04) = 10.26$, $p < .001$, $\eta_p^2 = .20$. The two methods produced similar proportions correct at the lowest bin (1–2), $p = .42$, whereas, for confidence ratings of 3 and 4, the item-based calculation revealed higher proportion correct than the response-based calculation, *p*s < .001, with differences of .21 and .11, respectively, for medium and high confidence. To reiterate, the item-based calculation remained the same across the two panels of Figure 3; the differences in outcome were solely driven by which false alarm rate was used for the CAC plots using response-based accuracy.

CAC plots with correct rejections

As with the analyses for hits, we also calculated CAC plots in two ways for correct rejections. Given that these proportion correct calculations use correct rejections as correct responses and the number of correct rejections changed based on relatedness of the lures, we used both different false alarm *and* correct rejection rates across two comparisons reported below. The miss rates stayed the same. Thus, unlike the case for hits, both CAC plot calculations using response-based and item-based accuracy changed across these comparisons (Figure 4). The contrasts again arose from the different types of lures, unrelated and related lures. For each contrast, we again conducted a 3 (confidence bins) x 2 (calculation type) repeated measures ANOVA.

Figure 4 provides CAC plots for correct *new* responses (correction rejections) for the case of unrelated and related lures (Figure 4(a,b), respectively). Figure 4(a)

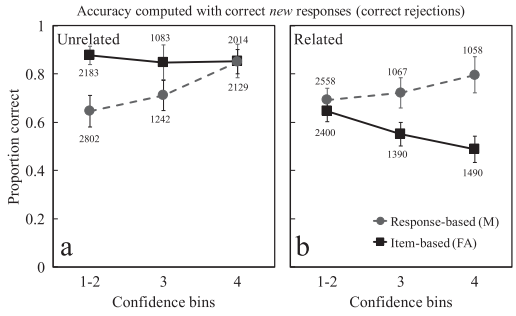

Figure 4. Comparison of CAC plots for accuracy computed with correct *old* responses (correct rejections) across four confidence bins using either false alarms or misses in the denominator of the CAC plot. (a) Experiment 3 with unrelated lures. (b) Experiment 3 with related lures. Error bars indicate 95% confidence intervals. Response-based accuracy was calculated as CRs/(CRs + misses) and item-based accuracy was calculated as CRs/(CRs + FAs).

shows that when the CAC plots for correct rejections were calculated with unrelated lures, the functions are high and flat, especially the CAC plot using false alarms. For the CAC calculation using and item-based accuracy, subjects were relatively accurate in correct rejections regardless of their level of confidence. Using response-based accuracy in the CAC calculation, however, shows an increase in accuracy with increasing confidence. The CAC plot using item-based accuracy ($M = .86$, $SE = .03$) showed higher accuracy than the CAC plot using response-based accuracy ($M = .74$, $SE = .02$), $F(1,41) = 27.54$, $p < .001$, $\eta_p^2 = .40$, just as happened in Experiment 1 with unrelated lures. The two calculations, however, did not differ at the highest confidence bin, $p = .881$. The relationship between confidence and accuracy was significant, $F(1.46,59.73) = 5.74$, $p = .011$, $\eta_p^2 = .12$, but was moderated by a reliable interaction, $F(1.03,70.68) = 21.87$, $p < .001$, $\eta_p^2 = .35$. For the CAC calculation using item-based accuracy, none of the pairwise comparisons between confidence bins were significant, $ps > .05$, whereas for the method using response-based accuracy, the highest confidence bin (4) showed greater accuracy for correct rejections than the other two bins, $ps < .001$, and the other two bins did not significantly differ.

Figure 4(b) provides the comparison between the two methods of calculating CACs with only related lures. Overall, the CAC plot using response-based accuracy ($M = .74$, $SE = .02$) revealed higher proportions correct than the plot using item-based accuracy ($M = .56$, $SE = .02$), $F(1,42) = 58.08$, $p < .001$, $\eta_p^2 = .58$, but the two methods did not differ at the lowest bin, $p = .152$. The main effect of confidence was not significant, $F(2,84) = 1.18$, $p = .312$, $\eta_p^2 = .03$, and was moderated by a reliable interaction, $F(1.60,67.28) = 17.61$, $p < .001$, $\eta_p^2 = .30$. In calculating CAC plots using item-based accuracy, the CAC plot was inverted from its usual form: Responses given low confidence (1–2) led to more accurate responding than did the responses made with higher confidence (3 and 4), $ps < .05$ in each case, whereas the bins expressing higher confidence did not differ, $p = .104$. We discuss this curious outcome – higher confidence leading to poorer accuracy – below.

For the correct rejection CAC plot using response-based accuracy, the usual CAC pattern was obtained, with higher confidence leading to somewhat greater accuracy. In Figure 4(b), the high confidence bin revealed higher accuracy than the other two bins, $ps < .05$, whereas the other two bins did not differ, $p = .560$. For the correct rejection CAC plot using item-based accuracy with related lures, the confidence-accuracy relationship was negative. We turn next to discussing these results in terms of how the number of observations changes depending on the types of lures that are used (although, of course, the confidence expressed in correctly rejecting or false alarming to the lures also matters).

Discussion

The bottom section of Table 2 provides separate accounts for the results from hits and correct rejections depending on whether lures were related or unrelated. In considering CAC plots for hits, because subjects' false alarm rates to unrelated lures were much lower than their miss rates, this outcome produced a higher CAC plot using response-based accuracy for unrelated lures rather than item-based accuracy (Figure 3(a)). As with Experiment 1, this difference between miss rates and false alarm rates to unrelated lures might be driven by conservative response criteria observed for unrelated lures in Experiment 3 (see Table 3). On the other hand, subjects false alarmed to related lures more compared to misses at the higher two confidence bins. Thus, the two calculations differed at those bins, with the CAC calculation using item-based accuracy leading to higher proportions of correct (Figure 3(b)). Table 2 shows that the diverging patterns seen in Figure 3(a,b) are determined by the numbers of false alarms for unrelated and related lures. False alarms for unrelated lures are rare relative to those for related lures. Further, the number of high confidence false alarms showed a similar imbalance. These related lures may be considered deceptive (Brewer et al., 2005; Koriat, 2008), and hence subjects often endorse them with medium or even high confidence.

The calculation of CAC plots with hits and unrelated lures in Experiment 3 (Figure 3(a)) showed a similar pattern to that of Experiment 1 (Figure 1(a)), with high confidence corresponding to high accuracy. Similarly, the hit CAC plots in Figure 3(b) with related lures shows a pattern similar to that seen in Experiment 2 (Figure 1(b) in that the CAC plot using item-based accuracy was higher than that using response-based accuracy; however, the difference is obviously much greater in Experiment 3 (Figure 3(b)) than in Experiment 2 (1b). In Experiment 2, for each target item, only one related (associated) lure occurred, but in Experiment 3, 10 categorised items were studied and 10 appeared as lures. Further, in Experiment 3, related and unrelated lures

were intermixed with target words on the test and with each other. Including the unrelated lures may have created a situation in which the related lures seemed more related to the target items than if the unrelated lures had not been presented (see DeSoto & Roediger, 2014). The bottom section of Table 3 suggests that these differences between the two experiments might have led to more liberal response criteria for related lures in Experiment 3 than in Experiment 2 though the discrimination between related lures and targets was similar to Experiment 2.

Considering correct rejections, the bottom section of Table 2 indicates that the numbers of correct rejections for unrelated lures were similar for the lowest confidence bin and highest confidence bin. More critically, the table shows that the number of correct rejections for related lures decreased from the lowest to the two higher confidence bins. Thus, when people accurately rejected related new items as *new*, they tended to have low confidence rather than high confidence, probably because of their semantic similarity to target items. This pattern, along with the more even distribution of related lure false alarms across confidence bins, accounts for why we observe a negative confidence-accuracy relationship in the CAC plot for correct rejections using false alarms to related lures (Figure 4(b)): For the higher confidence bins, the number of correct responses (correct rejections) decreased, whereas the number of incorrect responses (false alarms) did not decrease as fast, resulting in lower accuracy at these higher confidence bins. This negative CAC will likely occur with highly deceptive lures. The response-based CAC calculation for correct rejections demonstrated consistent results across different types of lures because the number of misses did not change when the CAC plots used related or unrelated lures as referents.

General discussion

CAC plots used in eyewitness identification experiments to assess the relationship between confidence and accuracy have consistently found that high confidence corresponds to very high accuracy, often 95% or better (Wixted & Wells, 2017). Our aim in this paper was to adapt the CAC method for use in *old/new* recognition experiments to discern possible consistencies and inconsistencies with multiple test items in recognition experiments relative to single-item eyewitness experiments. We computed proportion correct in two different ways for hits, one using response-based accuracy, and the other using item-based accuracy. The calculation using item-based accuracy is interesting and informative because the comparison is between items that were all studied for equivalent amounts of time and because the comparison raises an interesting challenge for theory, as discussed below. In both cases, CAC plots are derived as a function of confidence level. We also plot CACs for correct rejections using the same two methods, with either response-based accuracy or item-based accuracy.

Briefly, our results can be summarised by saying the CAC plot using response-based accuracy operate differently compared to the CAC plot using item-based accuracy depending on the nature of the lures used in the experiment. For hits, when unrelated lures were employed, the CAC function using response-based accuracy was higher than the one using item-based accuracy (Experiments 1 and 3). Critically, the plots with hits were both strongly positive (Figures 1 and 3(a)), although these CAC plots examine different questions regarding the confidence-accuracy relationship. That is, CAC plots using response-based accuracy demonstrate how confidence is related to accuracy for *old* recognition decisions, whereas CAC plots using item-based accuracy explore the confidence-accuracy relationship for old items per se (i.e., hits and misses). For correct rejections, when unrelated lures were employed, the CAC function using item-based accuracy (i.e., false alarms as errors) was higher than the one using response-based accuracy (i.e., misses as errors; Experiments 1 and 3). CAC plots for correct rejections were flatter compared to CAC plots for hits; however, they still demonstrated a positive confidence-accuracy relationship in most cases (Figures 2 and 4(a)).

When strongly related lures were used in the experiment, the number of false alarms greatly increased relative to using unrelated lures. This flipped the CAC functions using item-based accuracy and response-based accuracy such that the hit CAC plot using item-based accuracy became higher than that using response-based accuracy (Figures 1 and 3(b)). In Experiment 2, with only one related lure per target item, the reversal was small (.03 across confidence bins) but reliable (Figure 1(b)). In Experiment 3, with 10 related lures, the CAC plot using misses was much higher than the CAC plot using false alarms (Figure 3(b)). Experiment 3 is particularly compelling in showing the difference between unrelated and related lures because the same subjects and the same studied items were used in all analyses in within-subject designs.

Aside from Tekin and Roediger (2017), CAC plots have not been applied to standard list-learning recognition memory experiments. Although our study was exploratory, we believe it leads to interesting results. We have shown that in multi-item recognition experiments, conclusions from both methods of calculation converge nicely with findings from the single-item eyewitness identification procedure: For hits (identifying studied items as old), high confidence strongly indicates high accuracy. Even when deceptive lures (Brewer et al., 2005; Koriat, 2008) are used, CAC plots of either type show that high confidence indicates high accuracy. This is true when target words are intermixed with primary associates (Tekin & Roediger, 2017) and when 10 categorised words are recognised amidst 10 lures from the same category (with 12 different categorised sets studied; DeSoto & Roediger, 2014, Experiment 1). The fact that CAC plots reveal the

usual strong relation in the DeSoto and Roediger study is all the more surprising because, for categorised lures, there was a strong negative correlation between confidence and accuracy across items. That is, for lures from a studied category, the more often subjects false alarmed to a particular item, the greater their confidence that they were correct. This outcome supports Koriat's (2012) consensuality principle, which states that the more common a mistake is across subjects, the greater confidence people will have when making it. Thus, even in situations in which subjects might be expected to be confused because of the similarity of lures to targets, CAC plots indicate that confidence is still strongly related to accuracy. One can ask whether such high confidence-accuracy relationship observed in eyewitness experiments using CAC plots weaken or dissolve altogether with large numbers of items. The results reported here show that the answer is *no*, at least if the test is given shortly after the study phase.

In the first experiment reported here (Tekin & Roediger, 2017, Experiment 2), subjects studied 100 faces and received an *old/new* recognition test over 200 faces. Accuracy for the highest level of confidence was .95 when false alarms were used to calculate proportion correct and .89 when misses were used. The corresponding values when subjects studied 200 words and were tested on 400 in Experiment 2 were not quite as high, at .84 and .88, but in this case the lures were related to the targets. Finally, in the DeSoto and Roediger (2014) experiment, the accuracy rates for the two ways of calculating CACs were .87 and .92 (related and unrelated lures combined), although subjects had studied 120 words from 12 categories and been tested on 360. Nevertheless, higher confidence indicated higher accuracy (see Figure 3(b)). Thus, the confidence-accuracy relation for hits clearly does not disappear even with large numbers of items and numerous strongly related lures, and this is true for both methods of calculating CACs (response-based and item-based accuracy). However, CACs with hits using item-based accuracy do not (and cannot) differ as a function of the type of lures, and so only the CACs using response-based accuracy will accurately capture changes in performance as a function of the similarity of lures to targets. Thus, when the similarity of lures to targets is a variable of interest, the CAC function using response-based accuracy is a more appropriate measure. CAC plots using item-based accuracy, on the other hand, might be of interest because they capture incorrect *new* responses to old items given with high confidence. These responses are particularly interesting because in all experiments the recognition tests were administered just after a brief filler task. Nonetheless, in all experiments 16% to 21% of misses were made with high confidence, indicating that subjects were sure that they had not seen items that were just presented to them 10–15 min ago (see Roediger & Tekin, 2020).

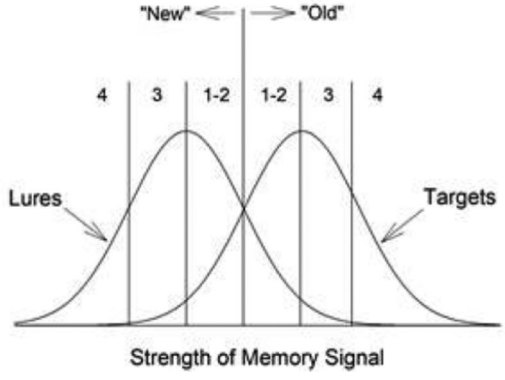

Figure 5. Idealised depiction of the equal-variance signal detection model with symmetrically arrayed confidence criteria. The figure depicts confidence ratings taken using a 4-point scale (1 = low confidence, 4 = high confidence), with ratings of 1 and 2 combined because we combined them in the reported experiments.

Each of the experiments we considered had at least one group of subjects that used a 100-point scale. Thus, we can examine the CAC plot for confidence levels of 90–100 in all three experiments. Reporting hit CAC plots based on item-accuracy first and then the one based on response-accuracy, the accuracy in the 90–100 bin was .94 and .96 for Experiment 1, .92 and .86 for Experiment 2, and .97 and .89 for Experiment 3. When subjects are highly confident, they are highly accurate, and once again this relation holds with word lists when the lures are similar to the targets in terms of association value (Experiment 2) or meaning (Experiments 2 and 3).

Theoretical implications

The findings reported here were not designed to test a specific theory, but we can ask if any theories naturally anticipate our results and whether any other theories struggle to explain them. Signal detection models, one version of which (the equal variance model) is illustrated in Figure 5, hold that (1) confidence increases for both *old* and *new* decisions the farther a test item's memory-strength signal deviates from the point of indifference (which is where the *old/new* decision criterion is placed) and (2) the probability that the *old* or *new* decision is correct increases accordingly. Accuracy is predicted to increase with confidence because the farther a test item's memory signal falls to the right or left of the decision criterion (leading to higher confidence), the more likely the item is to have been drawn from the target or lure distribution, respectively. Thus, our basic finding of a strong confidence-accuracy relationship for both *old* and *new* responses is naturally anticipated by SDT.

By contrast, our findings seem less compatible with threshold models of recognition memory. Threshold models hold that if the memory signal associated with a test item exceeds a certain value (i.e., if it exceeds a fixed *threshold*), it will be recognised as *old*. However, if the

signal falls below the threshold, then there is simply no information upon which to base a decision. One version of threshold theory – high threshold theory – holds that lures never generate a signal that falls above the threshold. Only targets can do that (Blackwell, 1953). Thus, this version of threshold theory would have to assume that the high false alarm rates generated by categorically-related lures in our experiments do not reflect strong memory signals but instead reflect random guessing (which is odd given that some of these false alarms are made with high confidence). However, another version of threshold theory – low threshold theory – holds that lures sometimes yield false, above-threshold memory signals (Kellen et al., 2016; Luce, 1963). Thus, low-threshold theory can accommodate the high false alarm rates generated by categorically related lures more easily than high-threshold theory can. Nevertheless, both threshold theories agree that when a signal falls below the threshold, it provides no information upon which to base a decision. Therefore, because such decisions are not based on a continuous memory strength signal, *new* decisions should not be associated with a graded confidence-accuracy relationship, yet they clearly are (Figures 2 and 4, using misses). Still another threshold model known as the two-high-threshold theory (Bröder et al., 2013) holds that targets are recognised in binary fashion (i.e., recognised as old or not) and lures are as well (i.e., recognised as new or not). Strong evidence in favour of this theory would come from a binary confidence-accuracy relationship for both *old* and *new* decisions, but no such pattern was observed.[1] Thus, it seems fair to say that our findings are not naturally anticipated by any of these threshold models.

Another issue that needs to be addressed is the different CAC functions observed for hits and correct rejections across the experiments. In all experiments, the CAC plots for correct rejections showed flatter functions than the CAC plots for hits (a finding reported in Tekin & Roediger, 2017). In Experiments 1 and 2, accuracy increased across confidence bins for both hit and correct rejection CAC plots (Figures 1 and 2), though subjects were better calibrated for hits (i.e., a steeper function exists for hits). In Experiment 3, the hit CAC plots again demonstrated a positive confidence-accuracy relationship for both related and unrelated lures; however, the correct rejection CAC plots using false alarms demonstrated a negative confidence-accuracy relationship with related lures in Experiment 3.

As noted by Mickes et al. (2011), the differing slopes associated with the CAC functions for hits (Figure 1, steep functions) and correct rejections (Figure 2, flatter functions) can be explained by the unequal-variance signal detection model, which has long been shown as a better model for recognition memory than the equal-variance signal detection model (Egan, 1958). In the unequal variance model, the standard deviation of memory strength associated with targets is greater than the standard deviation associated with lures, resulting in greater variance for the target distribution than the lure distribution. An unequal-variance model would arise if targets are conceptualised as lures that have had varying amounts of memory strength added to them during encoding of the list (whereas an equal-variance model would arise if every item on the list implausibly had exactly the same memory strength added). This additional memory strength to targets might be driven by the process of recollection. That is, although familiarity contributes to recognition of targets and lures, recollection mostly contributes to recognition of targets, producing greater variability for targets relative to lures (for a detailed discussion, see Wixted & Mickes, 2010). This unequal variance of memory strength between targets and lures, in turn, leads to different CAC functions for hits and correct rejections, as discussed below.

Figure 6(a,b) depict the target and lure distributions based on the unequal-variance signal detection model for both related and unrelated lures. According to this model, proportion correct for a given level of confidence in an *old* decision (hits or false alarms) corresponds to the area under the target distribution for a particular band of confidence (hits), divided by the sum of that area plus the corresponding area under the lure distribution (false alarms). Of course, for the alternative measurements using item-based accuracy, the area under the lure distribution is replaced by the area under the target distribution for the corresponding confidence band associated with a *new* decision (misses). For example, for hit CAC plots using unrelated lures to calculate response-based accuracy, the proportion correct for an *old* decision with a confidence rating of 3 is equal to the area under the target distribution corresponding to that confidence bin (i.e., the proportion of targets associated with hits made with a confidence rating of 3) divided by the sum of the areas under the target and lure distributions for that same bin (i.e., the proportion of targets receiving correct *old* responses with a confidence rating of 3 plus the proportion of unrelated lures receiving incorrect *old* responses with a confidence rating of 3). For hit CAC plots using item-based accuracy, on the other hand, the proportion correct for a confidence rating of 3 is equal to the area under the target distribution associated with correct *old* decisions made with a confidence rating of 3 (as before) divided by the sum of that area plus the area under the target distribution associated with incorrect *new* decisions made with a confidence rating of 3. Thus, in the latter calculation, both of the areas are under the target distribution. The same principle applies when calculating proportion correct for correct rejections.

Figure 6(a) shows that when the lures are unrelated as in Experiment 1, the target and lure distributions are further apart than when the distributions are related (Figure 6(b)). In both cases, the variance of the target distribution exceeds that of the relevant lure distribution. Because of the greater variance of the target distribution,

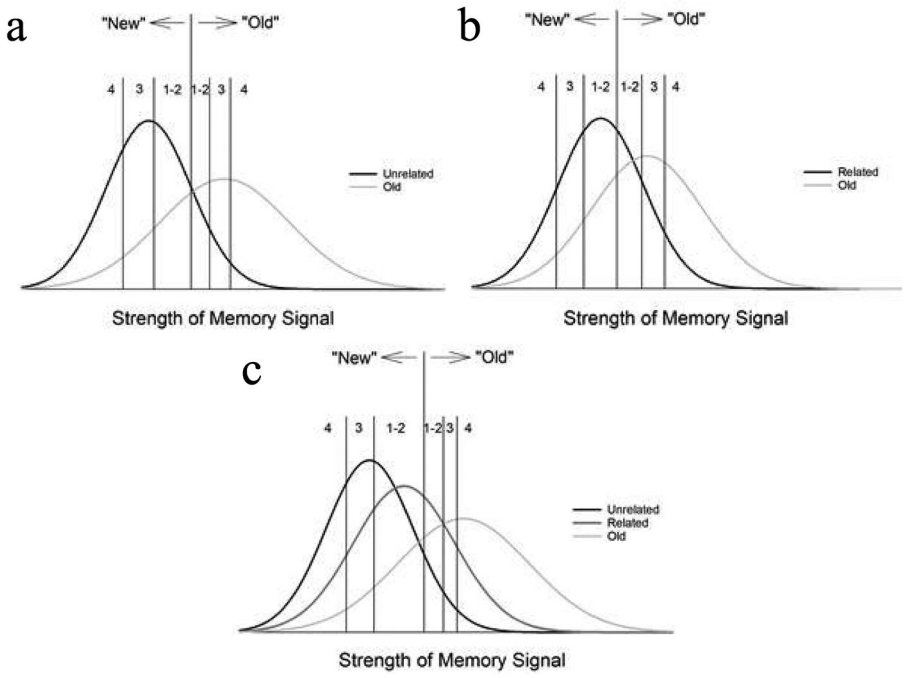

Figure 6. The unequal variance signal detection model based on empirical data from the three experiments we report. (a) Experiment 1 with unrelated lures. (b) Experiment 2 with related lures. (c) Experiment 3 with both related and unrelated lures. The relative variances of the target and lure distributions were estimated by fitting a straight line to the group z-ROC data from each experiment, and the locations of the confidence criteria were estimated using the z-transformed cumulative group false alarm rates (i.e., the group false alarm rate for *old* decisions made with a confidence rating of 4, the group false alarm rate for *old* decisions made with a confidence rating of 3 or 4, etc.).

the area under the lure distribution for *old* decisions (false alarms) decreases more rapidly with increasing confidence than the corresponding area under the target distribution (hits). This pattern leads to a steep increase in proportion correct for hit CAC plots. However, the opposite is true for correct *new* decisions (correct rejections), leading to a shallower confidence-accuracy relationship.

Figure 6(b) shows that when the lures are related to the targets as in Experiment 2, the lure distribution moves closer to the target distribution because related lures provide a stronger memory signal than unrelated lures. This change increases the area under the lure distribution for false alarms and decreases the area for correct rejections. This pattern results in flatter correct rejection CAC functions using item-based accuracy (between the ratings of 3 and 4) in Experiment 2 compared to that of Experiment 1. Given that the nature of hits and misses remains the same, the shape of the hit CAC functions using item-based accuracy does not change across Experiments 1 and 2. Furthermore, because the relatedness of lures causes their distribution to be closer to the target distribution (i.e., less discrimination from Experiment 1 to Experiment 2), now the area under the distribution for the false alarms is greater compared to that of misses, leading to the lower proportion correct for hit CAC calculations using response-based accuracy rather than item-based accuracy and vice versa for correct rejection CAC calculations in Experiment 2. Critically, the response criteria shift from conservative to more liberal across Experiments 1 and 2 might have also contributed to the increase in false alarms.

Because Experiment 3 included related and unrelated lures, both of the aforementioned considerations for Experiments 1 and 2 apply to Experiment 3, and the experiment is within the same subjects in the same experiment. In Experiment 3 we now have three distributions (Figure 6(c)): the target distribution, the related lure distribution, and the unrelated lure distribution. For related lures in Experiment 3, we observed a negative confidence-accuracy relationship for correct rejections when the CAC function is calculated using item-based accuracy (Figure 4(b), straight line), although the same relationship was not negative in Experiment 2, which also employed related lures (albeit many fewer related lures were used in Experiment 2 than in Experiment 3). We can then ask why a negative relationship exists in Experiment 3 but not in Experiment 2. The answer, when the data are interpreted in terms of SDT, is that the confidence criteria were symmetrically placed with respect to the related lure distribution for *old* and *new* decisions made with high confidence (4) but were asymmetrically placed with respect to low-confidence decisions (made with ratings of 1–2). Thus, high-confidence decisions were approximately 50% correct, but low-confidence decisions were well above 50% correct. The same asymmetry accounts for the flat (indeed, slightly negative) confidence-accuracy relationship for correct rejection CAC plots using item-based accuracy with unrelated lures in Experiment 3 (Figure 4(a), straight line). Conceivably,

subjects may have been inclined to often express low confidence in their *new* decisions and place a more conservative *old/new* decision criterion because they realised (more than they did in Experiment 2) that some of the test items falling below the decision criterion nevertheless seem surprisingly familiar (related lures) given that many other test items fall below the criterion (unrelated lures). Table 3 confirms that subjects were overall more conservative in Experiment 3 than Experiment 2. Subjects might have realised the deceptiveness of related lures more in Experiment 3 because related lures were intermixed with unrelated lures and because there were 10 related lures per category. Experiment 2, on the other hand, only had related lures and one related lure for each target item. This criterion-shift in Experiment 3, however, did not prevent subjects from false alarming to related lures more so than in Experiment 2. In fact, when the response criterion was calculated based on related lures only, subjects' response criteria were still liberal (i.e., more likely to endorse *old* for related lures, see Table 3). Alternatively, decreasing the proportion of old items in the recognition test might have also contributed to the shift towards more conservative responding in Experiment 3 (Rhodes & Jacoby, 2007; Van Zandt, 2000).

Extension to eyewitness memory

Before the widespread use of CAC plots was introduced to study the accuracy of eyewitness confidence, psychologists had claimed for over 30 years that eyewitness confidence was uncorrelated with accuracy or, later, not correlated well enough to be useful in a forensic setting (see Wixted et al., 2015). CAC plots not only show a strong confidence-accuracy relationship on an initial test of an eyewitness, but they also suggest that high-confidence IDs are highly accurate (Carlson et al., 2017; Seale-Carlisle et al., 2019; Semmler et al., 2018; Tekin et al., 2018; Wixted et al., 2016). We have shown that a strong confidence-accuracy relationship using CAC plots even in recognition experiments with hundreds of items. This suggests that in eyewitness cases that require multiple identifications, high confident IDs might still be highly accurate.

In eyewitness identification studies, CAC plots are calculated using response-based accuracy and focusing on eyewitnesses who chose incorrect (false alarms) or correct suspects. In the current paper, we argued that misses are also important recognition errors, and demonstrated a similar relationship between response-based accuracy (using false alarms) and item-based accuracy (using misses). Critically, failing to identify a perpetrator in a lineup (i.e., miss) is also an important recognition error in a real-life eyewitness scenario because such a decision might set the perpetrator free. Therefore, future eyewitness studies should examine whether response-based and item-based accuracy led to similar confidence-accuracy relationships in eyewitness situations.

Interestingly, in the eyewitness memory literature, studies using lineups have found a graded confidence-accuracy relationship for positive identifications of the suspect (analogous to *old* decisions here) but a flat confidence-accuracy relationship for lineup rejections (analogous to *new* decisions here, except that for the slope is usually positive for correct rejection CAC plots in our research). For example, this flat confidence-accuracy relation was the pattern reported by Brewer and Wells (2006) in their lineup study. On the surface, the flat relationship for *new* responses seems to accord with both versions of threshold theory considered above. That is, theoretically, the confidence-accuracy relationship is flat for lineup rejections because there is no memory signal upon which to base the decision. However, such a conclusion may be premature. Unlike a positive identification of a suspect in a lineup, which is made in relation to a particular face, a negative decision is made in relation to the entire set of faces in the lineup. That fact may be why the predicted confidence-accuracy relationship is observed only for positive decisions. To test this possibility, when a lineup is rejected, but before confidence is measured, the suspect in the lineup could be singled out. The eyewitness would then be asked, "how sure are you that the highlighted individual is *not* the person you saw commit the crime." Now, the negative confidence rating will be made in relation to a specific face, and the predicted confidence-accuracy relationship (similar to the one observed here for *new* decisions) should emerge. Such a result would substantially increase the forensic utility of a police lineup. Not only would confidence ratings sometimes provide compelling evidence of guilt (namely, when a high-confidence positive ID of a suspect is made), but, unlike now, they would also sometimes provide compelling evidence of *innocence* (namely, when a high-confidence negative ID of a suspect is made). Whether that predicted result will be observed remains to be seen, but the findings reported here in Figures 2 and 4 suggest that it probably will be. In sum, confidence in correct rejections and misses (in lists or lineups) may be of both theoretical and applied interest.

Conclusion

The relation between confidence and accuracy can be assessed in several different ways (Roediger et al., 2012; Roediger & DeSoto, 2015). The introduction of CAC plots into the study of eyewitness recognition changed the conclusion about how witness confidence is related to witness accuracy, by showing that (on an initial test with an unbiased lineup), confidence is strongly related to accuracy. We applied CAC methods using two different methods to standard recognition memory for both hits and correct rejections as indices of correct responding. For hits, we found that confidence is always highly related to accuracy in CAC plots using either misses or false alarms in the denominator to create the plot. This high confidence-accuracy relationship occurred even when subjects studied hundreds of words and when the

lures were unrelated or related to the target items (although related lures reduce accuracy at the highest confidence levels a bit). On the other hand, when considering correct rejections as the unit of correct responding, CAC plots are relatively flat or, with many related lures on the test, even negative. These outcomes provide a challenge for theory, and we proposed an unequal variance signal detection model to account for the results. In sum, applying CAC plots to standard recognition memory tasks reveals new findings that enhance our understanding of the confidence-accuracy relationship in the recognition process, and we believe that CAC plots prove a useful tool for investigating recognition memory.

Note

1. Two-high-threshold models can accommodate a continuous confidence-accuracy relationship if one assumes various confidence rating scale biases, but it seems fair to say that theory naturally predicts a binary relationship for both old and new decisions.

Acknowledgements

We thank Wenbo Lin, Oyku Uner, Jeremy Yamashiro and Steven Dessenberger for helpful comments on an earlier version of this manuscript.

Ethics approval and consent to participate

Both studies were approved by Washington University Institutional Review Board. At the beginning of the study, informed consent to participate was obtained from each subject. In addition, each subject was debriefed at the end of the study.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research was funded by James S. McDonnell Foundation.

List of abbreviations

CAC, confidence-accuracy characteristic

Availability of data and material

The datasets for Tekin and Roediger (2017) are available in the Open Science Framework repository, http:// osf.io/ 7f5ph.

The datasets for DeSoto and Roediger (2014) are may be found at http://pss.sagepub.com/content/by/supplemental-data

Authors' contributions

ET developed the method of calculating CAC plots with misses and conducted Experiments 1 and 2, with HLR. KAD conducted the experiment listed here as Experiment 3, with HLR. HLR suggested the analyses across three sets of data to determine if results would agree. JHW aided theoretical interpretations of our findings. Initial drafts of the paper were by ET and HLR, with KAD and JHW also contributing to the writing. JHW provided the threshold models in the Introduction and the signal detection models in the General Discussion. All four authors worked on revisions of the paper and approved the final draft.

References

Azimian-Faridani, N., & Wilding, E. L. (2006). The influence of criterion shifts on electrophysiological correlates of recognition memory. *Journal of Cognitive Neuroscience*, *18*(7), 1075–1086. https://doi.org/10.1162/jocn.2006.18.7.1075

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language*, *51*(2), 159–172. https://doi.org/10.1016/j.jml.2004.04.001

Blackwell, H. R. (1953). *Psychological thresholds: Experimental studies of methods of measurement* (Bulletin No. 36). University of Michigan, Engineering Research Institute.

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*(4), 618–627. https://doi.org/10.1016/j.jml.2005.01.017

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. https://doi.org/10.1037/1076-898X.12.1.11

Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, *21*(8), 916–944. https://doi.org/10.1080/09658211.2013.767348

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*(1), 26–48. https://doi.org/10.3758/BF03210724

Carlson, C. A., Dias, J. L., Weatherford, D. R., & Carlson, M. A. (2017). An investigation of the weapon focus effect and the confidence–accuracy relationship for eyewitness identification. *Journal of Applied Research in Memory and Cognition*, *6*(1), 82–92. https://doi.org/10.1016/j.jarmac.2016.04.001

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition*, *22*(3), 273–280. https://doi.org/10.3758/BF03200854

DeSoto, K. A., & Roediger IIIH. 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. https://doi.org/10.1177/0956797613516149

Dobbins, I. G., Kroll, N. E., & Liu, Q. (1998). Confidence–accuracy inversions in scene recognition: A remember–know analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(5), 1306–1315. https://doi.org/10.1037/0278-7393.24.5.1306

Egan, J. P. (1958). *Recognition memory and the operating characteristic (tech. Note No. AFCRC-TN-58-51)*. Indiana University, Hearing and Communication Laboratory.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence.

Psychological Review, *98*(4), 506–528. https://doi.org/10.1037/0033-295X.98.4.506

Han, S., Huettel, S. A., Raposo, A., Adcock, R. A., & Dobbins, I. G. (2010). Functional significance of striatal responses during episodic decisions: Recovery or goal attainment? *Journal of Neuroscience*, *30* (13), 4767–4775. https://doi.org/10.1523/JNEUROSCI.3077-09.2010

Kantner, J., & Dobbins, I. G. (2019). Partitioning the sources of recognition confidence: The role of individual differences. *Psychonomic Bulletin & Review*, *26*(4), 1317–1324. https://doi.org/10.3758/s13423-019-01586-w

Kellen, D., Erdfelder, E., Malmberg, K. J., Dube, C., & Criss, A. H. (2016). The ignored alternative: An application of Luce's low-threshold model to recognition memory. *Journal of Mathematical Psychology*, *75*, 86–95. https://doi.org/10.1016/j.jmp.2016.03.001

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, *77*(3), 217–273. https://doi.org/10.1016/0001-6918(91)90036-Y

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 945–959. https://doi.org/10.1037/0278-7393.34.4.945

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113. https://doi.org/10.1037/a0025648

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge University Press.

Lindsay, D. S., & Kantner, J. (2011). A search for influences of feedback on recognition of music, poetry, and art. In P. Higham & J. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honor of Bruce Whittlesea* (pp. 137–154). Palgrave Macmillan.

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, *70*(1), 61–79. https://doi.org/10.1037/h0039723

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. https://doi.org/10.1016/j.jarmac.2015.01.003

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General*, *140* (2), 239–257. https://doi.org/10.1037/a0023007

Mill, R. D., & O'Connor, A. R. (2014). Question format shifts bias away from the emphasised response in tests of recognition memory. *Consciousness and Cognition*, *30*, 91–104. https://doi.org/10.1016/j.concog.2014.09.006

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, *36* (4), 630–633. https://doi.org/10.3758/BF03206543

Murdock, B. B., & Dufty, P. O. (1972). Strength theory and recognition memory. *Journal of Experimental Psychology*, *94*(3), 284–290. https://doi.org/10.1037/h0032795

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. https://doi.org/10.3758/BF03195588

Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 305–320. https://doi.org/10.1037/0278-7393.33.2.305

Robinson, M. D., & Johnson, J. T. (1996). Recall memory, recognition memory, and the eyewitness confidence–accuracy correlation. *Journal of Applied Psychology*, *81*(5), 587–594. https://doi.org/10.1037/0021-9010.81.5.587

Roediger, H. L., & DeSoto, K. A. (2015). Understanding the relation between confidence and accuracy in reports from memory. In D. S. Lindsay, C. M. Kelley, A. P. Yonelinas, & H. L. I. I. I. Roediger (Eds.), *Remembering: Attributions, processes, and control in human memory: Papers in honor of Larry L. Jacoby* (pp. 347–367). Psychology Press.

Roediger, H. L., & Tekin, E. (2020). Recognition memory: Tulving's contributions and some new findings. *Neuropsychologia*, *139*, Article 107350. https://doi.org/10.1016/j.neuropsychologia.2020.107350

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

Seale-Carlisle, T. M., Colloff, M. F., Flowe, H. D., Wells, W., Wixted, J. T., & Mickes, L. (2019). Confidence and response time as indicators of eyewitness identification accuracy in the lab and in the real world. *Journal of Applied Research in Memory and Cognition*, *8*(4), 420–428. https://doi.org/10.1016/j.jarmac.2019.09.003

Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied*, *24*(3), 400–415. https://doi.org/10.1037/xap0000157

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. https://doi.org/10.2307/1413103

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), 41. https://doi.org/10.1186/s41235-018-0134-3

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), 49. https://doi.org/10.1186/s41235-017-0086-z

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 479–496. https://doi.org/10.1016/S0022-5371(81)90129-8

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the norms. *Journal of Memory and Language*, *50*(3), 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 582–600. https://doi.org/10.1037/0278-7393.26.3.582

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. https://doi.org/10.1037/0021-9010.88.3.490

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied*, *10*(3), 156–172. https://doi.org/10.1037/1076-898X.10.3.156

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 201–233. https://doi.org/10.1037/xlm0000732

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review*, *117*(4), 1025–1054. https://doi.org/10.1037/a0020874

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L. III. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*, *70*(*6*), 515–526. https://doi.org/10.1037/a0039510

Wixted, J. T., Read, J. D., & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence–accuracy relationship. *Journal of Applied Research in Memory and Cognition*, *5*(2), 192–203. https://doi.org/10.1016/j.jarmac.2016.04.006

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. https://doi.org/10.1177/1529100616686966