Psychonomic Bulletin & Review (2021) 28:556–564
https://doi.org/10.3758/s13423-020-01831-7

BRIEF REPORT

Discrete-state versus continuous models of the confidence-accuracy relationship in recognition memory

Christophe G. Delay[1] · John T. Wixted[1]

Accepted: 13 October 2020 / Published online: 27 October 2020

Abstract
The relationship between confidence and accuracy in recognition memory is important in real-world settings (e.g., eyewitness identification) and is also important to understand at a theoretical level. Signal detection theory assumes that recognition decisions are based on continuous underlying memory signals and therefore inherently predicts that the relationship between confidence and accuracy will be continuous. Almost invariably, the empirical data accord with this prediction. Threshold models instead assume that recognition decisions are based on discrete-state memory signals. As a result, these models do not inherently predict a continuous confidence-accuracy relationship. However, they can accommodate that result by adding hypothetical mapping relationships between discrete states and the confidence rating scale. These mapping relationships are thought to arise from a variety of factors, including demand characteristics (e.g., instructing participants to distribute their responses across the confidence scale). However, until such possibilities are experimentally investigated in the context of a recognition memory experiment, there is no sense in which threshold models adequately explain confidence ratings at a theoretical level. Here, we tested whether demand characteristics might account for the mapping relationships required by threshold models and found that confidence was continuously related to accuracy (almost identically so) both in the presence of strong experimenter demands and in their absence. We conclude that confidence ratings likely reflect the strength of a continuous underlying memory signal, not an attempt to use the confidence scale in a manner that accords with the perceived expectations of the experimenter.

Keywords Recognition memory · Signal detection theory · Threshold theory

Introduction

The relationship between confidence and accuracy is important to understand at a theoretical level. For example, in the field of eyewitness identification, it was long believed that confidence in a positive identification was largely unrelated to its accuracy (Sporer, Penrod, Read, & Cutler 1995; Wells & Murray, 1984). The implication was that judges and jurors should disregard confidence and instead concentrate on the fact that eyewitness memory is fallible. Courts across the USA accepted that science-based recommendation, and some presumably still do.[1] However, in recent years, it has become clear that on an initial test of memory from a lineup (early in a police investigation), confidence is in fact highly predictive of accuracy (Wixted & Wells, 2017). In other words, the higher the confidence, the more accurate the identification. The same is true of memory for a list of items in the basic-science laboratory (e.g., Mickes Hwe, Wais, & Wixted, 2011; Tekin & Roediger, 2017). Theoretically, what explains the graded relationship between confidence and accuracy?

Here, we consider two longstanding theoretical frameworks that speak to that issue. The first one, signal detection theory, is widely used in both psychology and neuroscience, and it assumes that recognition decisions are based on a

✉ John T. Wixted
jwixted@ucsd.edu

[1] Department of Psychology, University of California, San Diego, CA, USA

[1] For example, in 2012, the Connecticut Supreme Court stated "Courts across the country now accept that there is at best a weak correlation between a witness' confidence in his or her identification and its accuracy" (State v. Guilbert, 2012).

continuously distributed memory signal (Egan, 1958; Gold & Shadlen, 2007; Karanian & Slotnick, 2018; Kepecs, Uchida, Zariwala, & Mianen, 2008; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2013; Wixted, 2007, 2019). The second one, threshold theory, has a long history in the field of experimental psychology, and it assumes that recognition decisions are based on discrete memory states (Bröder & Shütz, 2009; Kellen & Klauer, 2018; Malmberg, 2002; Province & Rouder, 2012). As illustrated in more detail below, continuous models inherently predict that the relationship between confidence and accuracy for both "new" and "old" decisions should be continuous (i.e., as confidence increases, so should accuracy), whereas threshold models inherently predict that the relationship should either be flat (i.e., confidence is not at all predictive of accuracy) or be characterized by a step function (i.e., confidence should be either low or high). Empirically, the confidence-accuracy relationship is almost invariably continuous, as illustrated in Fig. 1.

Essentially the same issue has been the focus of prior work involving the shape of the confidence-based receiver operating characteristic (ROC), though when conceptualized from that angle, it may seem like a less important issue to understand. Threshold models predict that the ROC should be linear, but recognition memory ROCs are almost invariably curvilinear (e.g., Egan, 1958; Ratcliff, Sheu, & Gronlund, 1992). To accommodate that result (and to accommodate the continuous confidence-accuracy relationship illustrated in Fig. 1), threshold models require additional assumptions about the mapping relationship between discrete psychological states and the confidence rating scale (Malmberg, 2002). For example, when a test item leads to the "detect-old" state, thereby conclusively confirming its prior occurrence on the list (and therefore warranting high confidence), participants are assumed to nevertheless spread their responses continuously across the confidence scale. Such assumptions are unnatural in that they are not required by the model's conception of the nature of underlying memory signals. However, they do allow threshold models to better fit the data well, sometimes even better than signal detection models do (e.g., Kellen, Singmann, Vogt, & Klauer, 2015).

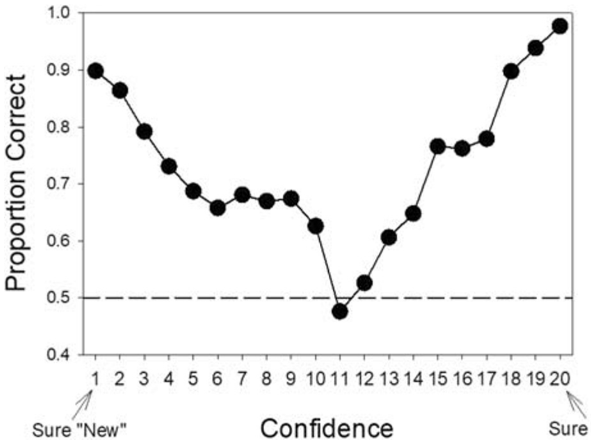

Fig. 1 The relationship between confidence and accuracy for memory tested using a list of words and recognition decisions made using a 20-point confidence scale (1 = "Sure New" to 20 = "Sure Old). Data from Mickes et al. (2011)

From a pure mathematical modeling perspective, goodness of fit (perhaps adjusted for model flexibility) is the ultimate adjudicator between competing models. Thus, if ad hoc mapping parameters allow threshold models to fit the data as well as a signal detection model, then, from this perspective, it means that the confidence-based recognition memory data are inconclusive. As such, to comparatively evaluate the models, other methods must be used (e.g., Bröder & Schütz, 2009). However, this line of reasoning simply takes confidence off the table, the very behavior we seek to theoretically explain here.

What explains the graded relationship between confidence and accuracy in recognition memory, both in the lab and in the real world? From the threshold perspective, a variety of possibilities have been offered. As summarized by Bröder, Kellen, Schütz, and Rohrmeier (2013), these possibilities include: (1) demand characteristics induced by instructing participants to spread their responses across the confidence rating scale; (2) sequential dependencies, where a response made to one test item carries over to the next; and (3) scale biases, such as anchoring effects. However, none of these possibilities has been experimentally investigated in the context of a recognition memory experiment to determine whether they meaningfully contribute to the confidence-accuracy relationship. As noted by Pazzaglia, Dubé, and Rotello (2013; Dubé, Rotello, & Pazzaglia, 2013), without knowing which variable (if any) explains the mapping parameters, there is no sense in which they add to our understanding of how confidence is related to accuracy. Here, we describe an effort to test whether demand characteristics account for the hypothesized mapping relationships. Before doing so, we briefly sketch out the competing theoretical accounts (Green & Swets, 1966; Macmillan & Creelman, 2005; Wixted, 2019).

Theoretical accounts of confidence in recognition memory decisions

Signal detection theory Signal detection theory (SDT) is illustrated in Fig. 2a. According to this account, items on a recognition memory test generate underlying signals that are normally distributed and therefore vary continuously in strength across items. The confidence rating applied to an "old" or "new" decision is determined by the strength of the memory signal relative to a decision criterion. The farther it falls above or below the decision criterion, the higher the confidence. Figure 2b shows the continuous confidence-accuracy relationship predicted by the signal detection model depicted in Fig. 2a.

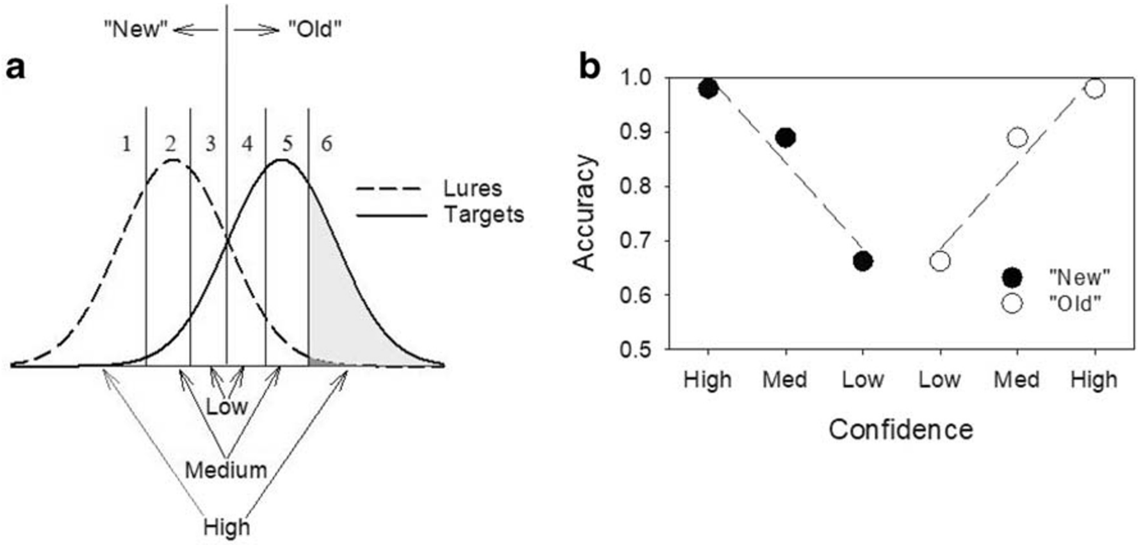

Fig. 2 **(a)** Standard equal-variance signal detection model of an old/new recognition memory test in which targets and lures generate Gaussian memory strength distributions, and responses are taken using a 6-point scale, where 1 = Sure New and 6 = Sure Old. Memory strength is represented on the x-axis, and test items that generate a signal strong enough to exceed the decision criterion are declared to be "old," whereas test items that do not are declared to be "new." As illustrated by the shaded regions, high-confidence "old" decisions are rarely made to lures but are frequently made to targets (thus, high-confidence accuracy should be high). **(b)** Continuous confidence-accuracy relationship predicted by the signal detection model depicted on the left (with $d' = 2$ and confidence criteria placed -.5, .5, 1, 1.5, and 2 standard deviations away from the mean of the lure distribution)

High-threshold theory The original threshold theory, known as high-threshold theory (HTT), instead assumes that test items are either recognized (the "detect-old" state) or they are not recognized (the "non-detect" state). Test items enter the detect-old state when the memory signal they generate exceeds a "high" threshold. As illustrated in Fig. 3a, because only targets can exceed that threshold, an above-threshold signal is perfectly diagnostic of the item's status as a target. It therefore follows that any target that generates a memory signal strong enough to exceed the threshold should be declared "old" and with high confidence (i.e., 6 for a standard 6-point scale).

Some targets fail to generate a memory signal strong enough to exceed the threshold, and all lures fail to do so. Critically, for these below-threshold targets and lures, there is no diagnostic signal upon which to base the recognition decision. Under such conditions, it is not obvious what the confidence rating should be, but even if those ratings are spread out along the confidence scale (as they almost always are), there should be no predictive relationship between

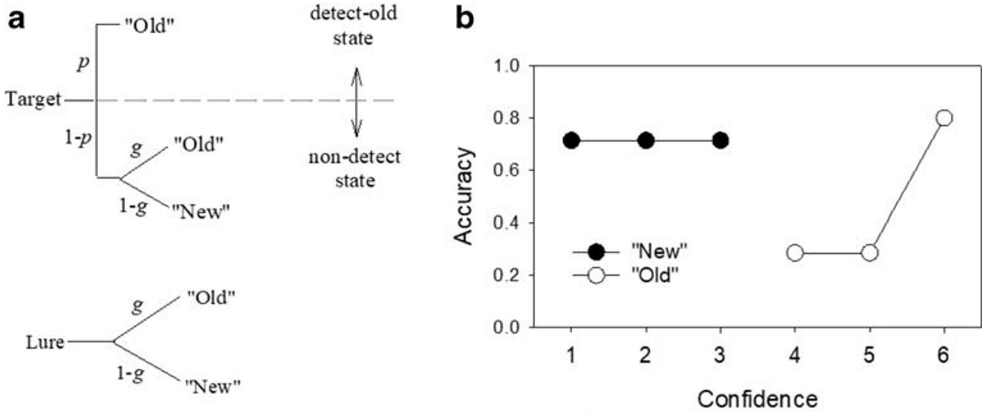

Fig. 3 **(a)** Depiction of the high-threshold theory (HTT) model, where p represents the probability that a target will enter the detect-old state, and g represents the probability that an item in the non-detect state is guessed to be old. **(b)** The relationship between confidence and accuracy predicted by the model shown in panel A assuming that (1) $p = .60$, with confidence = 6 in the "detect" state, and (2) $g = .50$, with confidence spread evenly across the scale in the "non-detect" state. Note that although the model depicted here assumes a .50 probability of guessing "old" in the below-threshold state, for low- and medium-confidence decisions, accuracy is below .50 for old items and above .50 for new items because 60% of old items are above threshold (reducing the number of old items available to receive decisions made with low or medium confidence)

confidence and accuracy. Figure 3b provides an example of what the model depicted in Fig. 3a naturally predicts. The predicted confidence-accuracy relationship is flat for "new" decisions (reflecting the fact that there is no diagnostic signal below the threshold) and is a step-function for "old" decisions.

2-high-threshold theory Like HTT, 2-high-threshold theory (2-HTT) assumes that only targets can exceed the threshold required to achieve the detect-old state, thereby generating a signal that is perfectly diagnostic of the item's status as a target (Macmillan & Creelman, 2005). However, unlike HTT, 2-HTT assumes a second high threshold, one that can be exceeded only by lures, thereby giving rise to the "detect-new" state (Fig. 4a). Because only lures can exceed this threshold, the detect-new state is perfectly diagnostic of the item's status as a lure. Thus, the predicted confidence-accuracy relationship for lures is the mirror-image of the predicted confidence-accuracy relationship for targets (Fig. 4b). That is, for both "old" and "new" decisions, 2-HTT naturally predicts a step-function relationship consisting of high accuracy for decisions made with the highest level of confidence, and low accuracy otherwise.

Low-threshold theory Finally, low-threshold theory (LTT) is similar to HTT except that it assumes that lures can also sometimes incorrectly exceed the (low) *detect-old* threshold (Luce, 1963). Note that this assumption differs from the 2-HTT assumption that lures can correctly exceed the high *detect-new* threshold. LTT has enjoyed a recent resurgence (e.g., Kellen, Erdfelder, Malmberg, Dubé, & Criss, 2016; McAdoo & Gronlund, 2019; Starns & Ma, 2018), and for good reason. First, it can provide a good fit to even apparently curvilinear ROC data. And second, like SDT, but unlike HTT and 2-HTT, it has the advantage of assuming that false alarms reflect consciously experienced memory signals, in agreement with the widely held view that false memories and true memories are subjectively indistinguishable (e.g., Bernstein & Loftus, 2009).

In LTT, beyond the information about whether a test item is in the detect state or the non-detect state, there is no diagnostic information available. If participants spread their confidence ratings over 4-5-6 when in the detect state and over 1-2-3 when in the non-detect state, the confidence-accuracy relationship would be flat for *both* "old" and "new" decisions (as in Fig. 3B for "new" decisions). Thus, LTT naturally predicts a confidence-accuracy pattern that is even farther removed from the empirical data than HTT.

Mapping relationships between discrete threshold states and continuous ratings

Although none of the threshold models reviewed above inherently explains the continuous confidence-accuracy relationship, they can all accommodate the data by assuming specific mapping relationships between discrete states and the confidence rating scale. We use HTT to illustrate this point, but the same idea applies to each of the threshold models considered above.

As before, assume that 60% of the targets give rise to the detect-old state, whereas 40% of targets and 100% of lures fall into the non-detect state. Further assume that instead of responding with high confidence when in the detect-old state, participants do so with a probability of only .60 and choose ratings of 5, 4, 3, 2, or 1 with continuously declining probabilities (.20, .10, .05, .04, and .01, respectively) and that when in the non-detect state, participants choose confidence ratings of 6 through 1 with the mirror image of those probabilities (namely, .01, .04, .05, .10, .20, and .60, respectively). Thus,

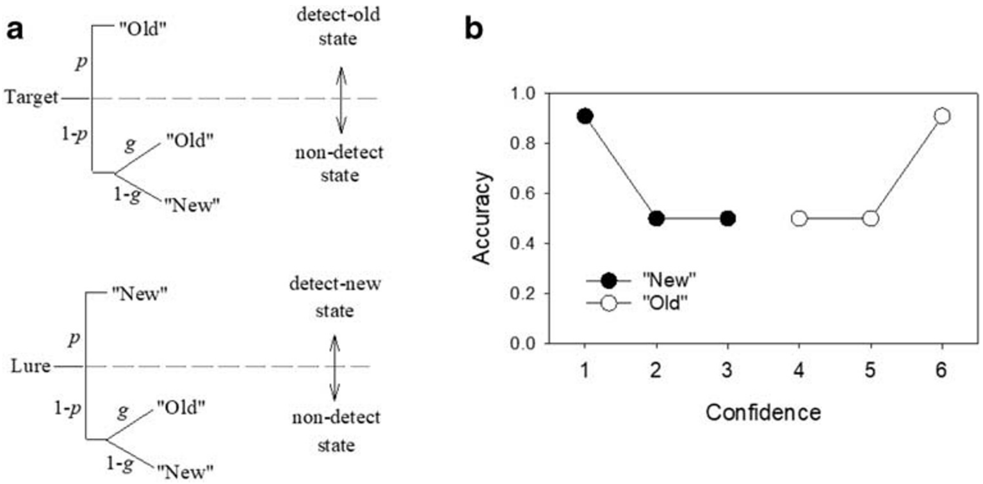

Fig. 4 **(a)** Depiction of the 2-high-threshold-theory (2-HTT) model, where p represents both the probability that a target will enter the detect-old state and the probability that a lure will enter the detect-new state, and g represents the probability that an item in the non-detect state is guessed to be old. **(b)** The relationship between confidence and accuracy predicted by the model shown in panel A assuming that (1) $p = .60$, with confidence = 6 in the "detect-old" state and confidence = 1 in the "detect-new" state, and (2) $g = .50$, with responses spread evenly across the confidence scale in the "non-detect" state

even when in the detect-old state, the model must assume that participants sometimes declare the item to be "new" with some degree of confidence (1-2-3). These are clearly nontrivial assumptions that, in our view, themselves require explanation. Nevertheless, by adopting hypothetical mapping relationships like these, HTT can account for a continuous confidence-accuracy relationship, as illustrated in Fig. 5. With similar assumptions about mapping relationships, 2-HTT and LTT can also accommodate the continuous relationship between confidence and accuracy for both "old" and "new" decisions. However, 2-HTT can do so without having to assume that the mapping relationships extend beyond a natural response category (e.g., in the "detect old" state, only ratings of 4-5-6 must be assumed).

What gives rise to the hypothesized mapping relationships like these? As noted earlier, one possibility is that instructions to participants often stipulate that the ratings should be spread across the confidence scale. Indeed, the mere fact that a 6-point confidence scale has been thrust upon the participant might create an implicit demand to use all available ratings. Given explicit or implicit demand characteristics like these, participants may not use the confidence rating scale in accordance with the underlying memory signals they experience. We tested that idea by manipulating demand characteristics across conditions.

Method

Participants

In total, we recruited 48 participants (75% female, 25% male) through the UC San Diego SONA portal who were asked to read and sign a consent form before proceeding with the study. The SONA portal is an undergraduate research portal that allows students taking psychology classes at the university to receive class credit in exchange for completing a psychology experiment.

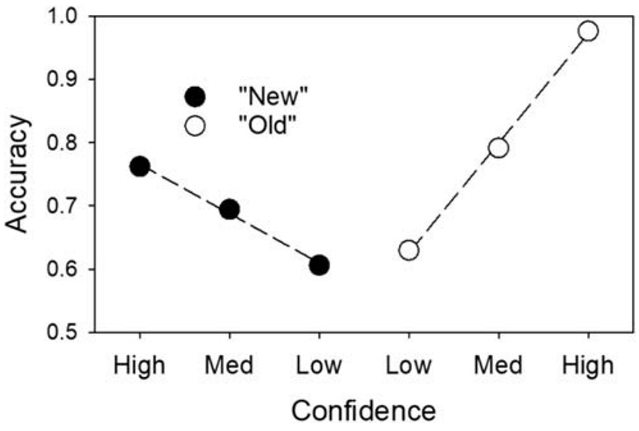

Fig. 5 Confidence-accuracy relationship predicted by the high-threshold theory (HTT) after adding specific assumptions about the mapping relationship between the memory state and confidence ratings

Materials

The lists were created using words drawn from an online database (Nelson, McEvoy, & Schreiber, 1998). The selection criteria consisted of nouns with frequency scores that fall between 100 and 1,000. The 383 words that satisfied these criteria constituted the word pool used to create the study list of 72 words and test lists consisting of those 72 words plus 72 additional words drawn from the word pool that served as lures. A different set of 144 words was randomly drawn from the word pool for each participant. In addition, we created a short practice study list and test list consisting of the following words:

> *Study list:* dog, cat, snow shovel, mountain, car, radio
> *Test list:* dog, penguin, snow shovel, valley, car, television

Design

In Experiment 1, participants were randomly assigned to one of two conditions that differed only in the instructions presented after the presentation of the list and just prior to the recognition test: (1) the *Demand* condition (16 participants) consisted of instructions to spread ratings across the confidence scale, and (2) the *No-Demand* condition (16 participants) consisted of instructions to use the confidence scale as desired. In Experiment 2, a third condition (similar to the *No-Demand* instruction condition) was implemented to further reduce any remaining demand characteristics that the *No-Demand* condition failed to remove. In this *Free* condition (16 participants), the instructions explicitly stated that the confidence scale should be used without regard to what the experimenter might want. Although the *Free* instruction condition was a second experiment, we report it together with Experiment 1.

Procedure

Participants were asked to read a brief introduction to the study, with a paragraph describing the purpose of the study, namely, to test models of recognition memory. After completing the short practice test, all participants studied a list of 72 words presented at a rate of one word per 3 s. For the subsequent recognition test, the 72 targets from the list and 72 lures were randomly intermixed and presented one at a time for an old/new decision. After each old/new decision, participants provided a confidence rating using a 1-to-6 scale, with 1 representing a high-confidence "New" decision and 6 representing a high-confidence "Old" decision.

Before taking the test, those in the *Demand* condition were instructed to spread their ratings more or less evenly across the scale (a common instruction). Those in the *No-Demand* condition were instructed to use the confidence scale as they pleased. In an effort to eliminate any ambiguity, they were specifically told that they did not need to spread their ratings across the scale. Those in the *Free* condition (Experiment 2) were additionally told that the experimenters: "have no expectation or preference whatsoever" and would: "simply like to know" how the participant would personally prefer to use the confidence scale. The instructions that were common across experimental conditions were presented on the screen in black (non-bold) font against a white background, whereas the critical instructions (which differed across conditions) were presented on the screen in bold and in blue. The experimenter was present and stressed the importance of reading all the instructions, additionally emphasizing the importance of the blue text because it provides information about their response strategy. The experimenter also asked if there were any questions before moving on to the next phase.

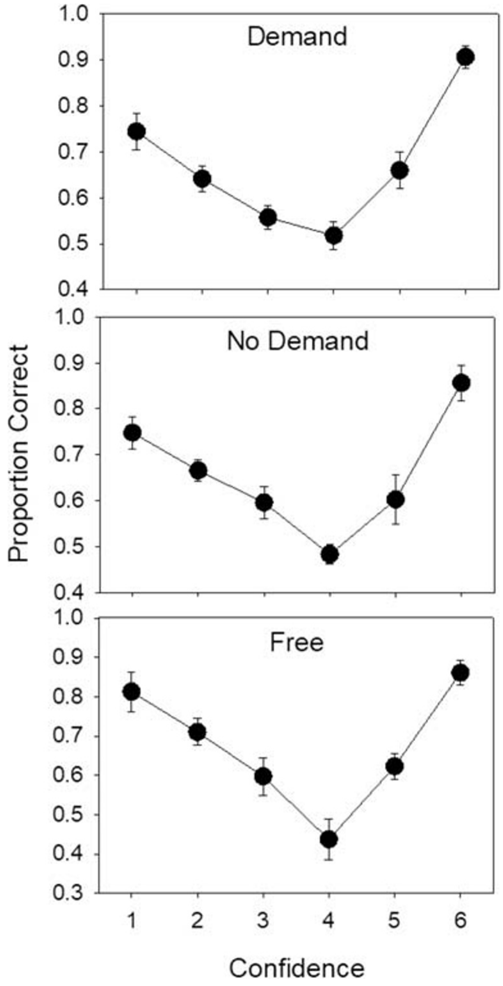

Fig. 6 Confidence accuracy relationships for the Demand and No-Demand conditions of Experiment 1 and for the Free condition of Experiment 2. Accuracy for "Old" decisions is equal to HR_c / (HR_c + FAR_c), where the subscript *c* refers to a particular level of confidence. An analogous accuracy score for "New" decisions is based on the correct rejection rate (*CR*) and the miss rate (*MR*) and is equal to CR_c / (CR_c + MR_c). For the equal base-rate situation used here (i.e., an equal number of targets and lures), these equations represent the posterior probability of being correct. Error bars represent standard errors

Results

Confidence-accuracy analyses For each individual participant, we computed accuracy (percent correct) separately for "Old" and "New" decisions and, within those decision categories, separately for each level of confidence (Mickes, 2015). The results of this analysis are shown in Fig. 6, and it is clear that there is a continuous confidence-accuracy relationship for both "old" and "new" decisions. Moreover, and critically, the trends do not differ in any appreciable way across the three instructional conditions.

To more precisely quantify these apparent continuous trends, we next computed slopes across the three levels of confidence within each decision category. More specifically, for each participant, we separately computed a "new" response slope (based on their accuracy scores over confidence ratings of 1–3) and an "old" response slope (based on their accuracy scores over confidence ratings of 4–6). The function fit to the "new" decisions was $\widehat{a}_i = p_1\ a_i + p_2$, where $\widehat{a}_i$ represents predicted accuracy for confidence rating *i*, a_i represents the observed accuracy, p_1 and p_2 represent the slope and intercept, respectively, and *i* ranged from

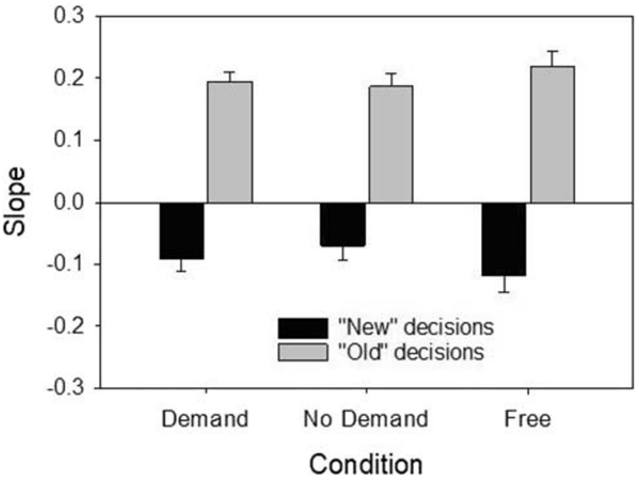

Fig. 7 Average "new" and "old" slopes over all participants for the Demand, No-Demand and Free conditions. For the "new" responses, the slopes were significantly less than 0 in all three conditions: $t(15) = -4.35$, $p = 0.0006$, Cohen's $d = -1.09$, $t(15)= -3.20$, $p= 0.0059$, Cohen's $d = -0.80$, and $t(14) = -3.90$, $p = 0.0016$, Cohen's d = -0.98, for the Demand, No-Demand, and Free conditions, respectively. For the "old" decisions (ratings between 4–6), the slopes were significantly positive in all three conditions, $t(15)=12.24$, $p < 0.001$, $d = 3.06$, $t(15)=8.98$, $p < 0.0001$, $d = 2.25$, and $t(14) = 8.35$, $p < 0.0001$, $d = 2.09$, for the Demand, No-Demand, and Free conditions, respectively

1 to 3. The same function was fit to the accuracy data for "old" decisions except i ranged from 4 to 6.

Figure 7 shows the average "new" and "old" slopes over all participants within their respective conditions (excluding one participant in the *Free* condition who only gave 1 and 6 ratings, making it impossible to compute a slope). The results clearly indicate that as confidence increases for "new" and "old" decisions, so does accuracy. This is shown by the fact that the slopes for old decisions are all positive (accuracy increases with increasing confidence from 4–6 on the scale), and the slopes for new decisions are all negative (accuracy decreases with decreasing confidence, i.e. as the numerical scale increases from 1–3).

The results of the slope analysis are consistent with SDT but are also consistent with threshold models that predict a dichotomous step-function relationship between confidence and accuracy (the prediction made by HTT for "old" decisions and by 2-HTT for both "old" and "new" decisions). Therefore, in addition to fitting a two-parameter straight line to the confidence-accuracy data from each participant, we also fit a two-parameter step function. We did so separately for "new" decisions and "old" decisions. For "new" decisions, the dichotomous function was $\widehat{a}_i = p_1\ a_i$ if i = 1 and $\widehat{a}_i = p_2\ a_i$ if i = 2 or 3, where $\widehat{a}_i$ represents predicted accuracy for confidence rating i, a_i represents observed accuracy, and p_1 and p_2 represent free parameters. For "old" decisions, the dichotomous function was $\widehat{a}_i = p_1\ a_i$ if i = 6 and $\widehat{a}_i = p_2\ a_i$ if i = 4 or 5. Both functions allow for high-confidence decisions to have higher accuracy than for decisions made with lower confidence (which are theoretically equal except for random error).

The linear and dichotomous fits each yielded a residual sum of squares (*RSS*), and we computed a difference between them ($RSS_{linear} - RSS_{dichotomous}$). Thus, positive values would support a continuous detection view, whereas negative values would support the discrete-state 2-HTT model. As shown in Fig. 8, in all three conditions, and for both "old" and "new" decisions, the linear function provided a numerically better fit.

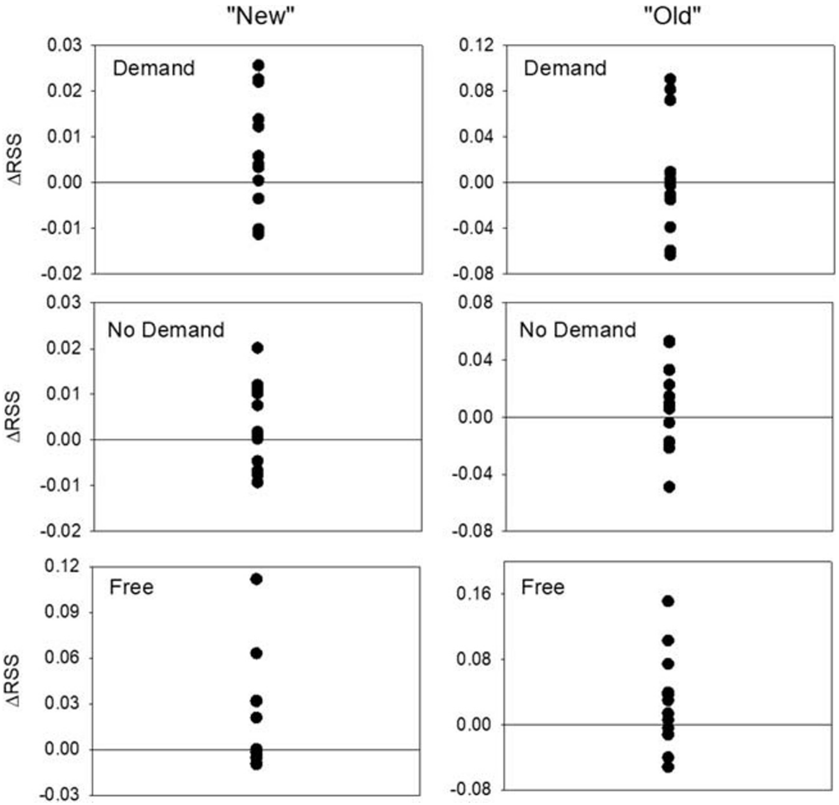

Fig. 8 Difference in *RSS* (where $\Delta RSS = RSS_{linear}$ - $RSS_{dichotomous}$) for two-parameter least-squares fits to confidence-accuracy data (linear – dichotomous) for "new" decisions (left column) and "old" decisions (right column). Positive values indicate a better fit for the linear function, whereas negative values indicate a better fit for the dichotomous function. The mean ΔRSS score was greater than zero for six out of six comparisons, p = .031 (a result that should be interpreted cautiously as we did not attempt to adjust for possible differences in model flexibility). Of greater relevance, according to independent-sample t-tests, for both "new" decisions and "old" decisions, none of the pairwise comparisons of the ΔRSS distributions (e.g., Demand vs. No Demand for "New" decisions) approached significance (obtained p-values ranged from .137 to .909). To increase power, we collapsed the data over "old" and "new" decisions and performed three pairwise comparisons (e.g., Demand vs. No Demand for "old" and "new" decisions combined). None of three comparisons approached significance (obtained p-values ranged from .154 to .811). Thus, any difference that might exist across conditions is likely too small to appreciably affect the confidence-accuracy relationship

Moreover, the results were not measurably affected by the removal of demand characteristics for either "new" or "old" decisions.

General discussion

The main purpose of the present investigation was to measure the confidence-accuracy relationship at the level of the individual participant, both in the presence of a demand to spread responses across the rating scale and in the absence of such a demand. A continuous confidence-accuracy relationship for both "old" and "new" decisions, the pattern inherently predicted by SDT, was observed regardless of the instructions. In fact, the results were essentially identical whether participants were explicitly instructed to spread their responses evenly across the scale or were instead emphatically encouraged to use the scale as they see fit. We interpret these findings as evidence that participants used the rating scale to gauge the strength of an underlying diagnostic memory signal, not because of demand characteristics.

An alternative explanation for ad hoc mapping relationships in threshold models is sequential dependencies, where a rating given to one test item tends to be repeated for the next test item (Bröder et al., 2013). However, although there is little doubt that sequential dependencies exist, the continuous confidence-accuracy relationship is almost certainly not the result of them. In recent years, it has become abundantly clear that the relationship between confidence and accuracy is invariably strong and continuous even in studies in which only one response is collected per participant, such as in a typical study of eyewitness identification (Wixted & Wells, 2017). This is true even of eyewitnesses tested in the real world (Wixted, Mickes, Dunn, Clark, & Wells, 2016). Obviously, no sequential dependencies exist when participants make only a single recognition memory decision.

Another possible source of the hypothesized mapping relationships is that the confidence scale itself might introduce scale biases (e.g., Bröder et al., 2013). For example, the wording associated with the confidence bins may encourage or discourage participants from using extreme responses. However, across a number of recent studies, it is striking how consistently the continuous confidence-accuracy relationship emerges over an extremely wide range of both verbal and quantitative scales (e.g., Dodson & Dobolyi, 2015; Weber, Brewer, & Margitich, 2008). For example, recently, Tekin and Roediger (2017) tested 4-, 5-, 20-, and 100-point scales and found that the different scales yielded similar (continuous) confidence-accuracy plots. In their words: "The scales seem convertible from one to the other, and choice of scale range probably does not affect research into the relationship between confidence and accuracy" (p. 2).

In summary, with regard to explaining confidence in recognition memory decisions, the evidence weighs against threshold models that assume demand characteristics, sequential dependencies, and/or scale biases. Thus, from the perspective of threshold models, we still have no idea what the ad hoc mapping relationships actually reflect. Until we do, it cannot be reasonably argued that, for confidence-based data (either ROC data or the confidence-accuracy relationship we focused on here), threshold models and signal detection models offer equivalent accounts merely because they both fit the data approximately equally well. In our view, that perspective loses sight of the purpose of models, which is to increase our understanding of behavioral phenomena of interest. For the models under consideration here, only signal detection theory helps us to understand the continuous relationship between confidence and accuracy.

Open Practices Statement The data and materials for the experiments reported here are available at the Open Science Framework (https://osf.io/53vhq/); none of the experiments was preregistered.

References

Bernstein, D. M., & Loftus, E. F. (2009). How to tell if a particular memory is true or false. *Perspectives on Psychological Science, 4*, 370–374.

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

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear—or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 587.

Dodson, C. S. & Dobolyi, D. G. (2015). Misinterpreting eyewitness expressions of confidence: The featural justification effect. *Law and Human Behavior, 39*, 266–280.

Dubé, C., Rotello, C. M., & Pazzaglia, A. M. (2013). The statistical accuracy and theoretical status of discrete-state MPT models: Reply to Batchelder and Alexander (2013). *Psychological Bulletin, 139*, 1213–1220.

Egan, J. P. (1958). *Recognition memory and the operating characteristic.* (Tech Note AFCRC-TN-58-51). Bloomington, IN: Indiana University, Hearing and Communication Laboratory.

Gold, J. I. & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30*, 535–74.

Green D.M. & Swets J.A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Karanian, J. M. & Slotnick, S. D. (2018) Confident false memories for spatial location are mediated by V1. *Cognitive Neuroscience, 9*, 3-4, 139–150.

Kellen, D., Erdfelder, E., Malmberg, K. J., Dubé, 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.

Kellen, D., & Klauer, K. C. (2018). Elementary signal detection and threshold theory. In J. T. Wixted (Ed.) & E. J. Wagenmakers (Vol.

Ed.) *Stevens' handbook of experimental psychology and cognitive neuroscience*, 4th Edition, Vol. V (pp. 1–39). Wiley.

Kellen, D., Singmann, H., Vogt, J., & Klauer, K.C. (2015). Further evidence for discrete-state mediation in recognition memory. *Experimental Psychology, 62*, 40–53.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455*, 227–231.

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review, 70,* 61–79.

Macmillan N. A. & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd). Mahwah, NJ: Erlbaum.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 380 – 387.

McAdoo, R. M., & Gronlund, S. D. (2019). Exploring Luce's (1963) Low-Threshold Model: Recognition Memory is Mediated by Discrete and Continuous Processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi: https://doi.org/10.1037/xlm0000731.

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.

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.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Retrieved from http://www.usf.edu/FreeAssociation/.

Pazzaglia, A. M., Dubé, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin, 139*, 1173–1203.

Pleskac, T. J. & Busemeyer, J. R. (2010).Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review 117*, 864–910.

Province, J. M. & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences of the United States of America, 109*, 14357–14362.

Ratcliff, R., Sheu, C.-f., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*, 518–535.

Ratcliff, R., & Starns, J.J. (2013). Modeling response times, choices, and confidence judgments in decision making: recognition memory and motion discrimination. *Psychological Review, 120*, 697–719.

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

State v. Guilbert. Supreme Court of Connecticut, 306 Conn. 218 (2012).

Starns, J. J., & Ma, Q. (2018). Guessing versus misremembering in recognition: A comparison of continuous, two-high-threshold, and low-threshold models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(4), 527.

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*, 49.

Weber, N., Brewer, N., & Margitich, S. (2008). The confidence-accuracy relation in eyewitness identification: Effects of verbal versus numeric confidence scales. In K. H. Kiefer (Ed.), *Applied psychology research trends* (pp. 103–118). Hauppauge, NY: Nova Science Publishers.

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

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

Wixted, J. T. (2019). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*, 210–233.

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E. & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences, 113*, 304–309.

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*, 10–65.