Using Confidence Ratings to Identify a Target Among Foils

James D. Sauer, Neil Brewer, and Nathan Weber

Flinders University

Author Note

James D. Sauer, Neil Brewer, and Nathan Weber, School of Psychology, Flinders University, Adelaide, South Australia, Australia.

James D. Sauer is now at the Department of Psychology, University of Portsmouth, Portsmouth, United Kingdom.

This research was supported by Australian Research Council Grant DP1093210 and a Flinders University research grant.

Correspondence concerning this paper should be addressed to James Sauer, Department of Psychology, University of Portsmouth, King Henry Building, King Henry I street, Portsmouth, PO1 2DY. Email: James.Sauer@port.ac.uk. Ph: 44-23-9284 6330. Fax: 44-23-9284 6300.

Abstract

Sauer, Brewer, and Weber (2008) advanced a novel procedure for testing eyewitness recognition memory. Rather than providing a single decision (i.e., identifying a lineup member or rejecting the lineup as a whole), participants rated their confidence that each lineup member was the culprit. Classification algorithms determined when patterns of confidence ratings indicated suspect guilt or innocence. Across varied test stimuli, confidence-based classifications equalled or out-performed single decisions. However, Sauer et al.'s classification criteria were designed to optimize performance for the data to which they were applied. If effective classification using confidence ratings requires such idiosyncratic criteria, the applied utility of the confidence procedure is nil. We re-analysed the data from Sauer et al.'s two identification experiments and demonstrated that confidence-based classification performance exceeding that of a traditional lineup task did not depend on uniquely-developed classification criteria. Confidence-rating lineups offer a potentially promising alternative to procedures requiring single decisions from witnesses.

Keywords: eyewitness identification, confidence, memory, decision criteria

Using Confidence Ratings to Identify a Target Among Foils

Eyewitness identification decisions should, ideally, result from witnesses comparing individual lineup members with their memory of the perpetrator. A strong memory, and match between a lineup member and the witness' memory, should produce a positive identification. If no lineup member strongly matches the witness' memory, or if the witness' memory is weak, the witness should reject the lineup or say *don't know*. However, non-memorial influences often distort this process (e.g., Wells, 1993). Lineup environment pressures may lead witnesses to set an inappropriate decision criterion, reducing the extent to which their identification decision reflects the match between their memory for the perpetrator and the lineup member. Sauer, Brewer, and Weber (2008) suggested a radically different lineup procedure - based on a novel application of basic memory theory - designed to reduce the effects of non-memorial influences on eyewitness decisions. The traditional identification test requires a witness to view a lineup and either select a lineup member as the culprit or reject the lineup. In contrast, Sauer et al.'s procedure did not require a decision by the witness. Instead, witnesses rated their confidence (0-100%) that each lineup member was the culprit. Subsequently, classification algorithms designed to optimize classification performance determined criteria for classifying these confidence ratings as indicative of a guilty or innocent suspect (see below for further explanation of the classification procedure). Previous research demonstrated that, compared to single identification decisions, confidence ratings provide a more sensitive index of recognition (Sauer, et al., 2008; Sauer, Weber, & Brewer, in press). However, in previous research, optimum classification criteria were developed separately for each stimulus set. To be useful in applied settings, classification criteria must be applicable across stimuli. Although Sauer et al.'s classification criteria performed well relative to standard identification decisions, it is not clear how they would perform when applied to different samples or stimuli. Here we tested this potentially major limitation of the confidence procedure’s practical utility: namely, the extent to which effective classification performance using confidence ratings depends on applying idiosyncratic classification criteria.

A lineup procedure based on the application of a computerized algorithm to confidence ratings represents a drastic departure from traditional practice. Almost certainly, the confidence procedure would be met with scepticism by researchers, law enforcement, the judiciary, jurors and witnesses (Brewer & Wells, 2011). Indeed, the lay view that “a computer could not possibly know my memory better than me” leaps to mind. However, psychologists know, based on various lines of evidence (e.g., the malleability of reports of memory phenomenology, Bodner & Lindsay, 2003; participants' lack of awareness of their cognitive processes, Nisbett & Wilson, 1977; feedback effects, Wells & Bradfield, 1998; anchoring effects, Wilson, Houston, & Brekke, 1996) that self-report measures do not give privileged access to mental processes and that more sophisticated, often indirect and even implicit, methods are necessary for valid measurement. Thus, based on (a) the initially promising empirical evidence, (b) recognition memory theory, and (c) measurement theory, we believe this procedure deserves thorough investigation. Despite the encouraging evidence we report here, we are not arguing for the procedure’s adoption. Many procedural aspects require investigation before such a recommendation would be prudent. Even if the procedure is empirically validated as effective and practically workable, the issues of its interpretation by law enforcement and jurors would need to be investigated, and the legal and policy ramifications of such a radical change thoroughly explored. Thus, our aim is to present evidence that strongly encourages further close investigation of this technique.

Why is a New Approach to Identification Evidence Needed?

The fallibility of eyewitness identification evidence and the consequences of identification errors are well-established. However, triers of fact find identification evidence compelling and identification evidence is sometimes the primary evidence against a suspect. Thus, eyewitness researchers endeavor to improve the reliability of identification evidence. However, as Wells, Memon, and Penrod (2006) noted, researchers have been “profoundly conservative” in their approaches to testing eyewitness memory (p.68). While researchers *have* investigated how variations in lineup construction and presentation affect the accuracy of identification decisions, they generally *have not* questioned whether a single identification decision is the best way to test a witness’ memory for the culprit, or the extent to which the suspect matches the witness’ memory of the culprit. For example, researchers have long debated the merits of sequential versus simultaneous lineup presentation. However, neither of these approaches produce acceptable accuracy rates (e.g., Brewer & Palmer, 2010; Lindsay, Mansour, Beaudry, Leach, & Bertrand, 2009; Steblay, Dysart, Fulero, & Lindsay, 2001). Similarly, even promising recent advances (e.g., using confidence or response latency to assess identification reliability, Brewer & Wells, 2006; Weber, Brewer, Wells, Semmler, & Keast, 2004; or using free-report procedures to improve the diagnosticity of volunteered responses, Weber & Perfect, 2012) produce significant error rates, or fail to reliably diagnose the accuracy of certain response types (e.g., lineup rejections). While it is possible to improve the reliability of evidence obtained using traditional lineup tasks, there have been no fundamental advances in lineup protocol in over two decades. Wells et al. suggest that, if psychologists developed a technique for testing witness memory based on psychological principles of memory function and testing, the result would probably look “radically different” to the traditional identification task (p. 69). Thus, we argue the merit of investigating novel approaches to testing witness memory.

Why Use Confidence-Rating Lineups Rather Than Single Identification Decisions?

There is theoretical and empirical support for using confidence ratings, rather than single decisions, to assess eyewitness memory. First, various theories of confidence processing for recognition memory decisions - based on signal detection theory (SDT) (e.g., Wixted & Mickes, 2010) and accumulator models of decision making (e.g., Van Zandt, 2000) - suggest that confidence should index the match between a presented lineup member and the witness' memory for the culprit (or *ecphoric similarity,* Tulving, 1981). Thus, the use of confidence ratings may avoid the potentially detrimental effects of non-memorial influences on the witness' decision criterion, and allow a more direct assessment of the degree of match between presented stimuli and the witness' memory. Further, if a witness provides multiple confidence ratings, investigators can not only determine which lineup member best matches the witness' memory (information also provided by an identification, though not by a rejection), but also the extent to which this member is favored over the alternatives. More importantly, investigators can specifically assess the extent to which the *suspect* matches the witness' memory. This information is not available if the witness provides a single identification decision. Second, across multiple stimulus sets and recognition memory tasks, previous research demonstrates that confidence ratings provide classification performance that equals or exceeds that of traditional identification tasks, and face recognition tasks (Sauer et al., 2008; Sauer, et al., in press). This evidence demonstrates that, when asked to provide a single identification response, witnesses do not make optimal use of the memorial information available to them. Compared with single identification decisions, confidence ratings offer a (a) richer source of diagnostic information, and (b) more sensitive index of memory, resulting in a more effective method of classification. Finally, although we acknowledge the novelty of the confidence rating approach in the forensic context, testing a latent construct (e.g., an individual's memory for a complex stimulus) using a single data-point departs from established principles of psychological measurement. Thus, in the broader psychological context, the standard identification task is the unusual testing procedure and, of course, a fallible one (see Brewer & Wells, 2011).

Threats to the Forensic Utility of Confidence-Rating Lineups

While the confidence procedure does not require a single response from witnesses, its intent is still to provide information about the likely guilt of a suspect. Thus, the procedure's forensic utility depends on identifying a criterion or critical value to determine when a pattern of confidence ratings, can be taken to indicate a positive classification. That is, when do a witness' confidence responses suggest that the suspect is the offender, and when do these responses suggest that the suspect is not the offender?

Adapting a procedure used by Koriat and Goldsmith (1996) to estimate the criteria participants used in memory control decisions, Sauer et al. (2008) advanced a number of alternative algorithms designed to identify critical values capable of discriminating target from foil stimuli. All algorithms first identified whether or not a participant's confidence ratings included a single-highest confidence value (hereafter referred to as a *max* value) indicating the lineup member the participant thought most likely to be the culprit. Cases for which no *max* value was present were treated as lineup rejections (as, in the absence of a *max* value, there was no reason to select one lineup member over the others as the culprit). Four potential algorithms were designed to determine when a *max* value could be taken as indicating suspect guilt. Classification criterion 1 (C1) considered only the absolute *max* value, indicating the magnitude this value must reach to indicate a positive classification. C2 and C3 indexed relationships between *max* and non-*max* confidence ratings.C2 subtracted the second highest confidence value from the *max* value, indicating the magnitude this difference must reach to return a positive classification. C3 subtracted the mean of the non-*max* confidence ratings from the *max*. Thus, C3 incorporated the additional assumption that there would be some degree of uniformity in the low confidence ratings given to unseen faces. Similarly, C4 considered the negative variance (i.e., the variance multiplied by -1) in non-*max* confidence ratings, indicating the degree of homogeneity required for a positive classification. Investigating C4 was motivated by speculation that, in the presence of a *max* value, the variability in non-*max* ratings may be lower when the *max* rating was assigned to a previously seen stimulus. Using the negative variance ensured consistency when presenting results; specifically, larger absolute values represented more conservative classification of recognition for all criteria. For all criteria, values equaling or exceeding the criterion represented positive classifications (i.e., indicating suspect guilt) while values failing to reach the critical value were treated as lineup rejections. Optimal critical values were those that maximized overall accuracy - with no preferential weighting given to accuracy for positive or negative classifications - and were derived from participants' data. For two large-scale eyewitness experiments the researchers identified critical values capable of providing classification accuracy rates comparable, if not superior, to a control, single identification decision condition. Using the simple algorithms outlined above, the confidence procedure substantially decreased false identification rates from target-absent lineups compared with the single-decision performance. For one experiment, when compared with the single decisions, the confidence procedure reduced overall target-absent false alarm rates from 53% to 15%. For the other experiment, the overall false alarm rate dropped from 36% to 10%.

In some cases the confidence procedure's superior target-absent performance was accompanied by (smaller) decreases in target-present performance compared with single decisions. However, a more sophisticated, hierarchical algorithm was able to improve target-present performance while maintaining target-absent performance comparable with the single decision condition. The hierarchical algorithm used different criteria depending on whether *max* values referred to suspect (H1) or foil (H2) stimuli. Separating suspect from foil *max* values enabled the hierarchical algorithm to use information routinely available in the forensic setting to remove potentially misinforming *max* values. Further, the hierarchical algorithm could then test foil *max* values for exonerating value. Wells and Olson (2002) argued that if a foil provides the best match to a witness' memory of the offender, especially to an extent great enough to merit a positive identification, the likelihood of the suspect being the culprit is greatly reduced[1]. Thus, H2 indicated the critical value that optimized the accuracy of rejections based on foil *max* values. Suspect *max* values reaching the H1 criterion were treated as positive classifications (i.e., identifications). Foil *max* values reaching the H2 criterion were treated as negative classifications (i.e., lineup rejections). Suspect *max* values that failed to reach the H1 criterion and foil *max* values that failed to reach the H2 criterion were classified as 'indeterminate' responses (i.e., cases for which the evidence was not strong enough to support a postivive or negative classification).

Sauer et al.(2008) used various stimulus sets. But one vital issue was not addressed: How generalizable are the optimal critical values across conditions? Thus far, classification performance has only been assessed using criteria specifically designed to optimize performance, and by applying these criteria to the stimuli/data from which they were derived. Both of these factors may have inflated performance. Further, the extent to which the levels of performance observed *depended* on applying unique, "optimal" criteria is unclear. In the applied setting, variations in encoding conditions and retention intervals are likely to influence a witness' memory and, consequently, their confidence ratings. Additionally, variations in test stimuli and conditions can vary greatly across crimes and witnesses, and may also alter witnesses' confidence ratings. These effects may shift the critical value required to optimize classification accuracy. However, it is impossible to calculate the optimal classification criterion on a case-by-case basis. The applied value of the confidence procedure will depend on its ability to provide an effective method of suspect classification across encoding and test conditions. Thus, it is important to test that reliable classification performance does not depend on applying idiosyncratic, optimal criteria.

SDT holds that discrimination (d') remains constant across variations in decision criteria. Does this prediction hold for the present context? Although this claim is central to the rationale for the confidence procedure, it is unclear that basic recognition theory adequately accounts for performance on complex, global memory tasks. Thus, it is necessary to test whether predictions derived from basic memory research account for applied memory phenomena. This research contributes to a growing literature exploring the applicability of basic memory theory to eyewitness identification performance. Further, analytically, d' does not index changes in memory quantity (i.e., the number of ‘responses’ provided). However, practitioners care about both the number and accuracy of classifications returned by an identification protocol. Thus, variations in performance using the hierarchical algorithm (which permits indeterminate classifications), must be assessed with reference to effects on classification accuracy and *quantity*. Simply demonstrating stable d' is insufficient.

We re-analysed the data from each of Sauer et al.’s (2008) identification experiments (Experiment 3 and Experiment 4), and charted changes in overall accuracy rates as the criteria for classification were shifted away from the identified ideal values (i.e., as the critical value becomes more or less conservative). Criterion stability was assessed for the four initial classification algorithms (C1–C4), and for the more sophisticated, hierarchical algorithm (H1 and H2). Demonstrating stable classification performance despite shifts in critical value would strengthen the applied potential of the confidence procedure.

Method

The data analysed here were from two eyewitness identification experiments (Experiments 3 and 4) reported by Sauer et al. (2008). For both experiments the presentation of crime and lineup stimuli was computerized. Each lineup consisted of a 2 × 4 array of color photographs. For both experiments, lineup instructions clearly stated that the offender may or may not be present, and participants in the single decision condition were provided with an explicit *not present* option (a *Not Present* button was presented on-screen below the lineup).

Experiment 3

Experiment 3 used four stimulus sets (each including a simulated crime event and associated target-present and -absent lineups). Clip A showed a young Caucasian male entering a residential property, and removing a VCR. Target A was visible for 29 s, with full or partial views of his face available for 9 s. Archival accuracy rates (based on published and unpublished research in our laboratories) for Target A are 76% and 44% for target-present and -absent lineups, respectively. Clip B showed a young Caucasian male attempting to break into a car. Target B was on camera for 14 s, with views of his face available for 8 s. Archival accuracy rates for Target B are 63% and 34% for target-present and -absent lineups, respectively. Clip C showed a middle-aged Caucasian male handing a bank teller a note and a brown paper bag, instructing the teller to follow the directions on the note and waiting as the teller filled the bag with money, and then exiting the bank. Target C was in view for 42 s, with his face visible for 37 s. Archival accuracy rates for Target C are 34% and 70% for target-present and -absent lineups, respectively. Clip D showed a young Caucasian woman shoplifting from a supermarket. Target D was presented for 43 s, with a full facial view available for 9 s. Archival accuracy rates for Target D are 24% and 60% for target-present and -absent lineups, respectively. Archival choosing rates for Targets A, B, C, and D are 74%, 74%, 53% and 44%, respectively. Thus, classification accuracy for the confidence and single decision conditions could be compared across tasks of varying difficulty.

Participants. Participants (*N* = 480) were undergraduate students paid for their participation.

Procedure. Participants were randomly allocated to view one of the four simulated crime videos. Exposure duration of the (Caucasian) targets varied from 14 s to 43 s. After viewing the crime participants did a puzzle task for 25 minutes, and were then presented with an eight-member simultaneous lineup. Participants in the single decision condition (a) clicked the photo of the lineup member they believed to be the target to indicate a positive identification or clicked the *Not Present* button to reject the lineup, and (b) provided a confidence rating in the accuracy of their decision. Participants in the confidence condition provided, for each lineup member, a confidence rating reflecting their belief that the lineup member was the culprit. Participants in both conditions registered their confidence ratings by typing a number in an on-screen box. Confidence ratings were free to vary from 0-100% and participants in both conditions could edit their confidence ratings before concluding the experiment by clicking a *Done* button.

Experiment 4

Experiment 4 attempted to replicate the findings from Experiment 3 while also examining the effect of foil similarity on confidence ratings from, and subsequent classification accuracy of, the confidence procedure. It used one simulated crime clip. The clip ran for 1 min 2 s, and showed four young, Caucasian adults breaking in to and robbing a self-storage facility. We used two of these four individuals as targets. Targets A and B were in view for 18 s and 29 s, respectively, with full or partial facial views available for 5 s and 7 s, respectively. Novel lineups were constructed for this experiment. Thus, no archival accuracy data were available for these stimulus sets. Sauer et al. (2008) reported target-present accuracy rates of 45% and 10% for Targets A and B, respectively, and target-absent accuracy rates of 60% and 68% for Targets A and B, respectively. Choosing rates were 25% and 7%, for Targets A and B, respectively.

Participants. Participants (*N* = 480) were undergraduate students paid for their participation.

Procedure. The procedure for Experiment 4 was identical to that for Experiment 3, except that participants watched a 15 minute video clip for the distracter task.

Results

The terms ‘classification criterion/criteria’ and ‘algorithm’ refer to the measure used to determine whether a confidence rating (or pattern of confidence ratings) was indicative of the suspect’s guilt. Classification criteria reduce participants’ confidence data to a single numerical value (see Sauer et al., 2008). The term ‘critical’ value refers to the magnitude that this value must reach in order to return a positive (or negative) classification. The term ‘optimal critical value’ refers to the critical value that maximizes overall classification accuracy. Finally, the following sections discuss findings in relation to ‘targets’, however the findings obviously reflect the properties of the stimulus sets as wholes.

Across targets, approximately 20% of trials were rejected for not containing a *max* value (see Table 1). Table 2 presents the optimal critical values for each stimulus set, and for data collapsed across stimulus sets in each experiment, for the initial criteria (C1-C4), and the two criteria associated with hierarchical algorithm (H1 for positive classifications based on suspect *max* values, and H2for negative classifications based on non-suspect *max* values). Variation in optimal critical values is evident across targets. However, the extent to which this jeopardizes the applied utility of the confidence procedure will depend on the extent to which classification performance varies as the applied critical value is shifted away from the optimal critical value.

The Effects of Varying the Critical Value on Classification Accuracy

Initial criteria. Figure 1 plots variations in overall accuracy according to the critical value applied[2]. Our analyses focus on the effect of varying critical values on overall accuracy rates because our algorithms were designed to maximize overall accuracy. However, the effects of varying critical value on the accuracy of positive and negative classifications are also indicated[3]. Overall accuracy rates for the control condition are provided to show how variations in the critical value applied affect overall accuracy relative to the control condition, as well as relative to optimal performance. To gauge the stability of classification performance as critical values were varied, a series of 2 × 2 contingency table analyses compared the number of correct and incorrect classifications generated by each possible critical value (as derived from the data) with those generated using the optimal critical value. In Figure 1,for each criterion (C1 through to C4), the critical values are plotted on the x-axis such that more lenient critical values are located toward the left extreme of the axis, and more conservative critical values toward the right extreme. The bold vertical lines indicate performance using the optimal critical value. Each point on the function represents performance using a different critical value. The dashed vertical lines indicate the points at which the difference between accuracy for the optimal and applied critical values reached the cut-off for a small effect (Cohen's *w* = 0.10)[4]. As an example, we refer the reader to the panel for Experiment 3, and the function for C1. The optimum critical value is 100 (indicated by the bold vertical line). As the applied critical value departed from the optimal value, in this case becoming more lenient, overall performance remained relatively stable until the applied critical value reached approximately 70 where performance worsened. Performance then remained relatively stable until it worsened again when the applied critical value reached approximately 55. The dashed vertical line indicates that, at this point, the difference between performance using the applied (i.e., 55) and optimal (i.e., 100) critical values reached the cut-off for a small effect. Analyses were conducted on data from each target, and on data collapsed across targets. The overall level of performance, and the relative contributions of positive and negative classification accuracy to overall performance, varied across targets. However, the stability of overall classification performance - despite variations in the applied critical values - was similar across targets. Thus, for ease of interpretation, we present data for these analyses collapsed across targets for Experiments 3 and 4.

Figure 1 reveals three important patterns of change in overall accuracy with critical value. First, the optimal critical value is not associated with a notable peak in overall accuracy. Second, results from the contingency table analyses show that the applied criterion can depart considerably from the optimal criterion before producing even a small effect on overall accuracy. Third, for all criteria the optimal critical value for Experiment 3 is included in the identified range for Experiment 4, and vice versa. Thus, the applying the optimal critical value for one dataset to another would not produce a notable departure from optimal performance. Moreover, while there is, for each target, an optimal critical value capable of maximizing overall accuracy, this value does not generally represent a pronounced improvement. Instead, performance tends to improve gradually as the applied critical value approaches the optimal value. This suggests that the diagnostic utility of the confidence procedure does not require applying unique, optimal critical values for individual stimulus sets.

Hierarchical algorithm. Criteria C1-C4 reduce participants' confidence ratings to a binary outcome. If a critical value is reached or exceeded, a positive classification is made (a lineup member is identified as the target); if not, a negative classification is returned (the lineup is rejected). However, the hierarchical algorithm employs a more sophisticated classification system. As previously outlined, the hierarchical algorithm applies a separate criterion for cases in which the *max* value refers to the suspect (H1) and cases in which the *max* value refers to a foil (H2). In the forensic context, using an appropriately constructed single-suspect lineup means that the foils in the lineup are known to be innocent. By separating suspect *max* values from foil *max* values, information routinely available in the forensic setting can be used to eliminate many potentially misleading *max* confidence values from our analyses. Further, whereas previously a negative classification merely indicated the absence of sufficient evidence to indicate guilt, the hierarchical algorithm is able to test non-suspect *max* values for positive, exculpatory evidence. Cases that failed to meet either of the criteria received an indeterminate classification. This third classification option has important applied and analytical implications. The applied implications will be considered in the Discussion. The analytical implications are as follows.

Indeterminate classifications are conceptually similar to *don't know* responses. They indicate that the available information is insufficient to reliably indicate a suspect's guilt or innocence. This third classification option means that the accuracy of classifications and the number of correct classifications generated can vary independently. Thus, when assessing the effects of different critical values on classification performance using the hierarchical algorithm, it is important to chart changes in both the accuracy and number (or quantity) of correct classifications generated. Figure 2 plots classification performance as a function of the applied critical value for both positive (H1: left panel) and negative (H2: right panel) classifications, for all targets. A series of 2 × 3 contingency table analyses compared the number of correct, incorrect and indeterminate classifications generated by each possible critical value with those generated using the optimal critical value. Again, the bold line indicates optimal classification performance, and the dashed line indicates the point at which the difference between the optimal and obtained performance reaches the cut-off for a small effect. When considering the hierarchical algorithm results, it should be noted that because our focus is on variations in classification performance as a function of the critical value applied, rather than on classification performance per se, the following analyses include only data from trials for which a *max* confidence value was present – and for which a classification algorithm could be applied. Thus, the accuracy rates presented do not include lineup rejections resulting from the absence of a *max* value.

An inspection of the accuracy rates for positive (H1) classifications suggests that applying increasingly conservative critical values increased classification accuracy but decreased the number of correct positive classifications. Importantly, these reductions in correct classifications were not accompanied by increased incorrect classifications. Rather, the increase in accuracy was associated with an increase in indeterminate classifications. While the dotted lines indicated reductions in performance (i.e., quantity) reaching the cut-off for a small effect, the number of data points supporting the analysis for each target is relatively small (*N* = 13-40). When these data points are split among the three possible classifications, small changes in performance can lead to relatively large changes in proportions and, consequently, effect sizes. Thus, our estimates are conservative. Increases in accuracy were generally modest as, across targets, even with lenient critical values, accuracy tended to be high. Unlike Figure 1, Figure 2 presents data for individual targets because the varied patterns observed for H2 classifications are masked when data are collapsed. Visual inspection of Figure 2 does not suggest notable peaks in performance associated with optimal critical values and changes in performance appear gradual and systematic. In almost all cases, optimal critical values for individual targets are included in the identified range of values for other targets, indicating that applying the optimal critical value from one dataset to another would generally not produce a notable departure from optimal performance .Thus, these data suggest that the utility of the classification procedure may not depend on applying idiosyncratic critical values.

For negative (H2) classification performance (right panel), data for all targets show a reduction in the number of correct classifications (with an accompanying increase in indeterminate classifications) as the critical value for negative classifications is increased. However, the associated effects on the accuracy of negative classifications varied. Applying more stringent critical values improved classification accuracy for targets 3A, 4B and, possibly, 3C, but decreased accuracy for 3D. Targets 3B and 4A show little variation in accuracy associated with changes in critical value. These results are consistent with Clark et al.'s (2008) finding that the diagnosticity of foil identifications can vary according to the nature of the lineup stimuli.

Discussion

Previous research suggests that a radical departure from traditional lineup practice – asking witnesses to rate their confidence that each lineup member is the culprit – may a) allow a more direct index of the extent to which individual lineup members match the witness' memory of the offender, and b) reduce identification errors by ameliorating the biasing effects of non-memorial factors on witness decisions (Sauer, et al., 2008). Previous research compared participants' confidence ratings with unique critical values developed for individual stimulus sets (to determine when they suggest that the suspect is or is not the offender) and found overall classification performance similar or superior to that for a single decision control condition. In the applied setting, however identifying critical values on a case-by-case basis is impossible. If effective classification using the confidence procedure requires identifying and applying specific, optimal critical values (and these values vary across stimuli), the applied utility of the procedure (as a replacement for the traditional identification task) would be nil. However, if classification performance shows consistency despite variation in the critical value applied, the forensic utility of the confidence procedure is enhanced. Here we showed that effective classification using the confidence procedure did not depend on developing and applying unique, optimal critical values.

Specifically, we found that the applied critical value could depart substantially from the ideal before producing any notable effect on overall accuracy (i.e., before the difference between obtained performance and optimal performance reached the cut–off for a small effect). While optimal values differed across stimulus sets, applying non-optimal critical values often produced negligible differences in classification performance. The systematic – and, thus, predictable – effects on overall accuracy of varying the critical value, and the absence of any notable peak in performance associated with the optimal value, strengthen this conclusion. Additionally, criteria that considered the difference between a participant's *max* confidence rating and other confidence ratings (e.g., C2 and C3) – reducing the impact of individual difference factors that have a main effect on confidence ratings– showed improved stability in optimal critical values across stimulus sets. The relative stability of classification performance despite shifts in the critical values applied suggests that criteria developed from one dataset could provide effective classification when applied to novel stimulus sets, increasing the applied potential of the procedure.

From a theoretical perspective, the generous boundaries within which critical values could vary before producing more than a negligible effect on overall accuracy support the rationale underlying the confidence procedure. Classifications made using the confidence and single decision procedures rely on the same memorial information. When using the confidence procedure, critical values could vary substantially before a) affecting overall accuracy rates, and b) with the exception of C1 and C4 for Experiment 4, before overall accuracy dropped below that for the single decision, control condition. Thus, at least in some cases, control condition participants' criterion placement must have been considerably sub-optimal, as has also been demonstrated for decisions from both simultaneous and sequential lineups(Palmer & Brewer, 2011). Therefore, a procedure attempting to index ecphoric similarity directly, and use this information without requiring the witness to set a decision criterion, may offer substantial practical benefits. While basic memory theory (e.g., SDT) would predict this outcome, we needed to demonstrate that this finding from basic memory research would generalize to performance on more complex memory tasks.

Results for the hierarchical algorithm were also encouraging. When *max* values indicated the suspect, the optimal critical values for discriminating guilty from innocent suspects (H1) were remarkably consistent across targets. However, relatively small shifts in the applied critical value appeared capable of affecting classification performance. While this finding may appear problematic, it requires further interpretation. Across targets, applying more conservative critical values reduced the number of correct positive classifications. However, this reduction was accompanied by increases in the number of indeterminate classifications generated, rather than in the number of false identifications generated. Sauer et al.(2008) have previously argued that indeterminate classifications should be viewed as legitimate indications that the available evidence is insufficient to assess the likely guilt of the suspect. Given the well established problem of mistaken identifications, a classification procedure providing the opportunity to forego identifying the suspect without either a) ruling out the suspect, or b) discrediting the witness, may be valuable in the applied setting (see Weber & Perfect, 2012, for a similar argument regarding explicit don't know response options in identification procedures). Thus, the forensic implications of the observed effects on ‘quantity’ may be less severe than they appear. Rather than increasing the chance of error, applying a more conservative critical value simply increases the chances that the identification evidence will fail to meet the level required for a positive classification.

While applying more conservative critical values did increase H1 classification accuracy, the relatively high H1 accuracy rates (even when critical values were lenient) meant increases tended to be small. These consistently high accuracy rates suggest that, when the *max* value implicates the suspect, it need not be very high in order to reliably discriminate between guilty and innocent suspects. If consistent, this would be a promising finding for the applied utility of the procedure. Consider the following scenario: a witness views a lineup and thinks that lineup member number four might be the culprit. However, the witness is not confident and does not wish to make a false identification. Thus, s/he rejects the lineup. The witness is correct to reject the lineup if unsure (assuming a *don't know* response option is unavailable). However, if lineup member number four is the culprit, this incorrect rejection could lead to police releasing a guilty person from custody, potentially misleading subsequent investigative efforts, and reducing the likelihood of a conviction if the case ever reaches court. If relatively low suspect *max* confidence values can reliably discriminate between guilty and innocent suspects, the confidence procedure may be capable of avoiding situations where a guilty suspect is removed from further investigation because, although the witness favoured them over the alternative candidates for selection, the witness' criterion placement meant they were not confident enough to make a positive identification.

While the present results are encouraging, two caveats are required. First, although our stimulus sets were diverse enough to produce variations in identification performance, across experiments there was considerable homogeneity in terms of target age and ethnicity, encoding conditions, retention interval, and testing format. Thus, various factors that may affect memory performance were not accounted for. Further, it is unclear if (how) such effects would translate into effects on the *stability* of classification performance across variations in the applied critical value. This may limit the generalizability of the results obtained. Second, while our analyses suggest that the applied critical value can vary considerably from the optimal critical value before producing a small effect on classification performance, the window within which the critical value can vary while still producing performance superior to the control condition will be smaller. This is an important consideration relating to the practical value of the confidence procedure. However, it does not undermine our conclusions that (a) these data demonstrate stable classification performance across variations in the applied critical value, and (b) these findings provide a strong impetus for further development of this alternative to traditional practice.

Moving away from binary assessments of suspect guilt (i.e., traditional lineup practices) may benefit forensic investigations in a number of ways. It allows uncertain witnesses to forego positive identification without acting to rule out a potentially guilty suspect. Similarly, it counteracts the pressures inherent in the identification environment which can lead an uncertain witness to lower their criterion and pick incorrectly. It also allows for varied degrees of strength of evidence against a suspect. For example, a witness may favor the suspect over all other lineup members by a large degree, or may give the suspect a *max* confidence rating of 100. This may offer strong evidence that the suspect is guilty. In such cases, investigators may be confident in their decision to charge the suspect. However, a witness may favor the suspect over other lineup members but to a lesser degree, or give the suspect a moderate *max* confidence rating. These circumstances may encourage investigators to continue pursuing the current suspect, while suggesting that more evidence is required before charges should be laid. Single identification decisions lack these subtleties. For example, building on the uncertain witness scenario outlined above, a witness who is 30% confident that a lineup member (who happens to be the suspect) is the culprit may be sensible to reject the lineup. However, our results suggest that a moderate suspect *max* value (e.g., 30%) can, at least under some conditions, reliably discriminate guilty from innocent suspects. Wells and Luus (1990) argue that a lineup is a method for testing an investigator/s' hypothesis that a suspect is guilty. A lineup rejection and a suspect *max* confidence rating of 30% could be based on identical degrees of match between the suspect and the witness' memorial representations of the culprit. However, they clearly offer different degrees of support for the underlying hypothesis. Thus, similar levels of witness confidence in the guilt of a suspect can have importantly distinct implications for an investigation, depending on how the identification evidence is elicited. As Sauer et al. (2008) argued, while the clarity afforded by single responses may be appealing, probabilistic responses may improve evidential quality and informativeness.

In sum, while classification performance using the confidence procedure and the critical values that optimize classification performance vary according to the nature of the crime and lineup stimuli used, our results suggest that the diagnostic utility of this procedure, particularly in terms of discriminating guilty from innocent suspects, need not depend on developing and applying unique, critical values for individual stimulus sets. However, our conclusion is not that the confidence procedure is currently suitable for the applied setting. Nor is it that any of the specific criteria, critical values, or algorithms identified thus far are the best available. Instead, we argue that we have provided strong evidence against what would have been a fatal limitation of this method of suspect classification. Demonstrating that the utility of the confidence procedure need not require applying specifically-determined critical values for individual stimulus sets represents an important step in assessing the potential practical value of this *type* of classification procedure.

References

Bodner, G. E., & Lindsay, D. S. (2003). Remembering and knowing in context. *Journal of Memory and Language, 48*, 563-580. DOI: 10.1016/s0749-596x(02)00502-8.

Brewer, N., & Palmer, M. A. (2010). Eyewitness identification tests. *Legal and Criminological Psychology, 15*, 77-96. DOI: 10.1348/135532509x414765.

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

Brewer, N., & Wells, G. L. (2011). Eyewitness identification. *Current Directions in Psychological Science, 20*, 24-27.

Clark, S., Howell, R., & Davey, S. (2008). Regularities in Eyewitness Identification. *Law and Human Behavior, 32*, 187-218. DOI: 10.1007/s10979-006-9082-4.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490-517.

Lindsay, R. C. L., Mansour, J. K., Beaudry, J. L., Leach, A. M., & Bertrand, M. I. (2009). Sequential lineup presentation: Patterns and policy. *Legal and Criminological Psychology, 14*, 13-24. DOI: 10.1348/135532508x382708.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

Palmer, M. A., & Brewer, N. (in press). Sequential lineup presentation promotes less biased criterion setting but does not improve disciminability. *Law and Human Behavior* DOI: 10.1037/h0093923.

Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General, 137*, 528-547.

Sauer, J. D., Weber, N., & Brewer, N. (in press). *Using ecphoric confidence ratings to discriminate seen from unseen faces: The effects of retention interval and distinctiveness*.

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 25*, 459-473.

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

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 582-600.

Weber, N., Brewer, N., Wells, G. L., Semmler, C., & Keast, A. (2004). Eyewitness identification accuracy and response latency: The unruly 10-12-second rule. *Journal of Experimental Psychology: Applied, 10*, 139-147.

Weber, N., & Perfect, T. J. (2012). Improving eyewitness identification accuracy by screening out those who say they don’t know. *Law and Human Behavior, 36*, 28-36. DOI: 10.1037/h0093976.

Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist, 48*, 553-571.

Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*, 360-376.

Wells, G. L., & Luus, C. (1990). Police lineups as experiments: Social methodology as a framework for properly-conducted lineups. *Personality & Social Psychology Bulletin, 16*, 106-117.

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

Wells, G. L., & Olson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied, 8*, 155-167.

Wilson, T. D., Houston, C. E., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General, 125*, 387-402.

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

Footnotes

[1] Clark, Howell and Davey (2008) subsequently demonstrated that the exculpatory value of foil identifications depends on lineup composition and the method of foil selection. We return to this issue when discussing our results.

[2] Three points are worth noting with regard to the panels presenting data for C4. First, for C4, x-axis values represent variance in non-*max* confidence ratings. Unlike C1-C3, these values need not vary between 0 and100. Second, unlike C1-C3, smaller absolute values represent more conservative criterion placement (as greater uniformity is required in non-*max* ratings to return a positive classification). Finally, Figure 1 presents the negative variance values so that, as with C1-C3, conservativeness increases from left to right along the x-axis.

[3] In the absence of a compelling reason to prioritize accuracy for positive classifications over accuracy for negative classifications, or vice versa, we analysed variations in overall accuracy.

[4] We assessed changes in performance in terms of effect size because, unlike inferential statistics, effect size measures are independent of sample size. For all initial criteria, the difference between classification performance for the optimal critical value and the most conservative possible value did not reach the cut-off for a small effect.

Table 1

Proportion (SE) of Trials not Containing a Max Value According to Target-Presence, for each Stimulus Set

	Target-Presence		
Stimulus set	Target Present	Target Absent	Overall
Experiment 3			
A	.03 (.03)	.27 (.08)	.15 (.05)
B	.17 (.07)	.23 (.08)	.20 (.05)
C	.10 (.05)	.17 (.07)	.13 (.04)
D	.10 (.05)	.37 (.09)	.23 (.05)
Overall	.10 (.03)	.26 (.04)	.18 (.02)
Experiment 4			
A[a]	.13 (.05)	.33 (.06)	.23 (.04)
B[a]	.18 (.04)	.28 (.06)	.23 (.04)
Overall	.16 (.03)	.31 (.04)	.23 (.03)

Note. [a] Stimulus set labels A and B correspond to the Male and Female stimuli, respectively, in Sauer et al. (2008).

Table 2

The Identified Optimal Critical Values, for the Four Initial Criteria (C1-C4) and for the Hierarchical Algorithm (H1 and H2), for each Stimulus Set

	Initial criteria				Hierarchical algorithm	
Stimulus set	C1	C2	C3	C4	H1	H2
Experiment 3						
A	98.00	70.00	77.14	-15.48	30.00	5.00
B	85.00	80.00	76.57	-14.29	30.00	20.00
C	100.00	100.00	82.86	-57.14	30.00	10.00
D	90.00	70.00	70.71	-95.24	10.00	40.00
Overall	100.00	70.00	76.57	-17.95	30.00	10.00
Experiment 4						
A[a]	95.00	80.00	77.14	-57.14	4.00	20.00
B[a]	100.00	100.00	100.00	0.00	30.00	70.00
Overall	95.00	90.00	93.57	0.00	4.00	70.00

Note. [a] Stimulus set labels A and B correspond to the Male and Female stimuli, respectively, in Sauer et al. (2008).

Figure Captions

Figure 1. Overall accuracy rates (and accuracy for positive and negative classifications) according to the critical value applied for the initial criteria, for data from Experiments 3 and 4collapsed across targets. Horizontal dotted lines indicate overall accuracy for the control, single identification decision comparison. Vertical bold lines indicate optimal criteria. Vertical dotted lines indicate the point at which the difference between the observed and optimal overall accuracy rates reaches the cut off for a small effect size (*w* = 0.10).

Figure 2. Quantity (as a proportion of total trials for which the classification type would be correct) and accuracy of positive (Panel A) and negative (Panel B) classifications using the hierarchical algorithm, according to the critical value applied, for each target. Vertical bold lines indicate optimal criteria. Vertical dotted lines indicate the point at which the difference between the observed and optimal overall accuracy rates reaches the cut off for a small effect size (*w* = 0.10).

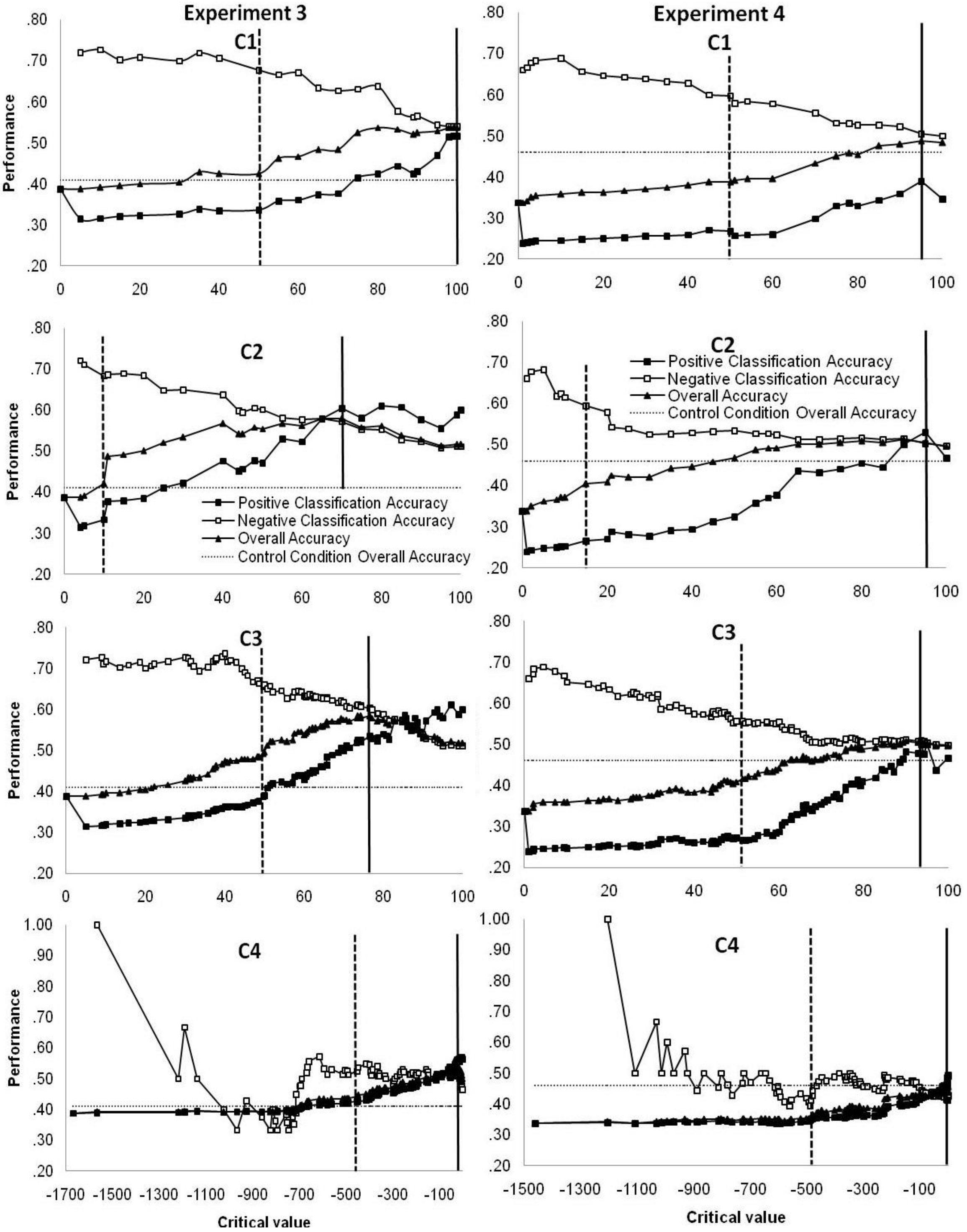
Experiment 3
Experiment 4
C1
C2
C3
C4
Performance
Critical value
.80
.70
.60
.50
.40
.30
.20
1.00
.90
0
20
40
60
80
100
-1700
-1500
-1300
-1100
-900
-700
-500
-300
-100
Positive Classification Accuracy
Negative Classification Accuracy
Overall Accuracy
Control Condition Overall Accuracy

Figure 1

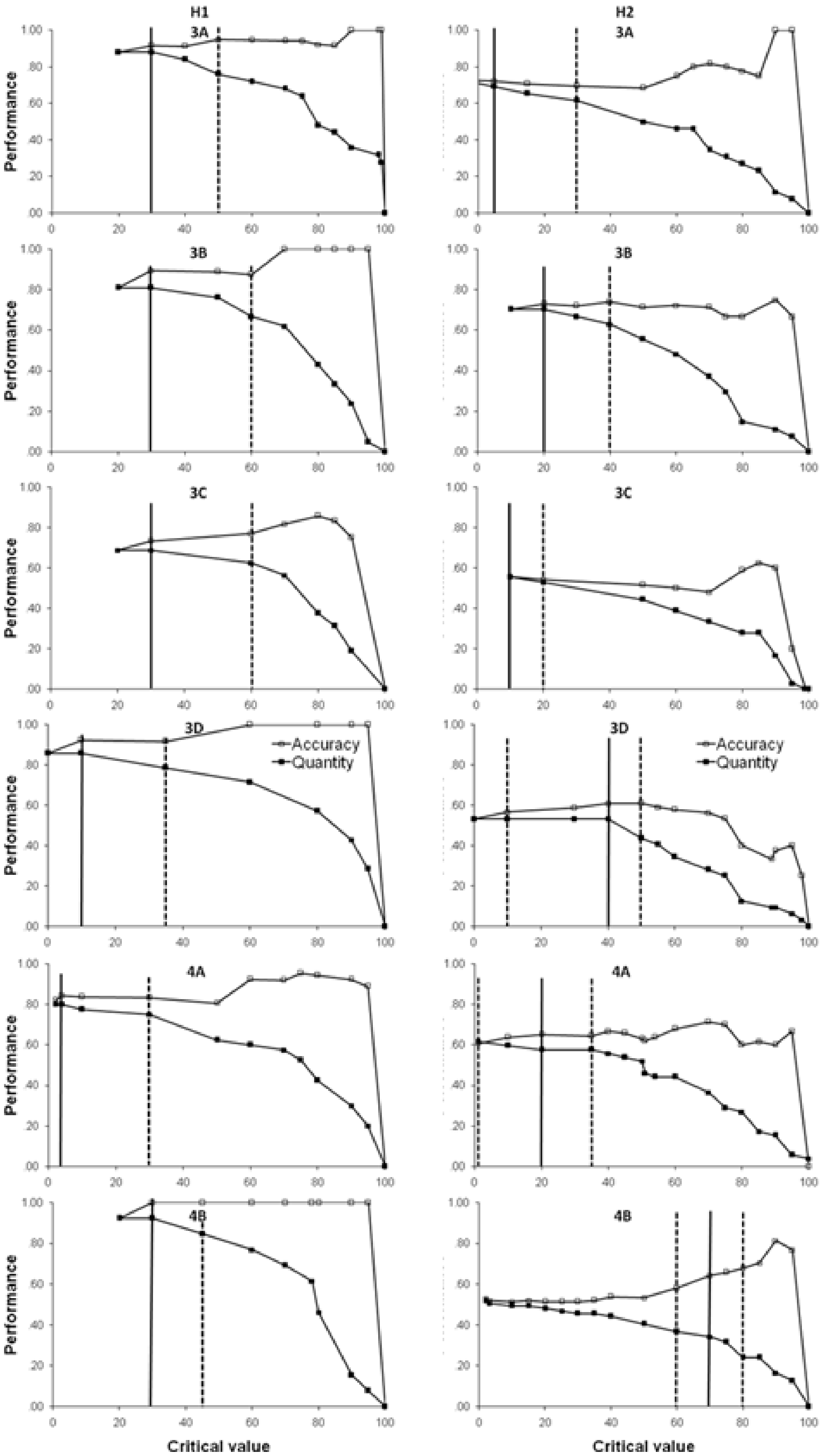
H1
H2
3A
3B
3C
3D
4A
4B
Performance
Critical value
Accuracy
Quantity
1.00
.80
.60
.40
.20
.00
0
20
40
60
80
100

Figure 2