For A. Smith, M. Toglia, & J. Lampinen (2018). *Methods, measures, and theories in eyewitness identification tasks*. Taylor & Francis

Ratings-based identification procedures

James D. Sauer

University of Tasmania

Neil Brewer

Flinders University

Supported by funding from Australian Research Council grant DP150101905 to N. Brewer et al.

Corresponding author: James Sauer
School of Medicine (Psychology)
College of Health and Medicine
University of Tasmania
Private bag 30, Hobart
Tasmania, 7001

Email: Jim.Sauer@utas.edu.au

Ratings-Based Identification Procedures

As discussed elsewhere in this volume, and at length in the broader literature, eyewitness identification evidence is common, compelling, and prone to error. Traditional identification procedures—where witnesses either pick a lineup member or reject the lineup—suffer a number of limitations that contribute to these errors. First, given the nature of the lineup environment, procedures that require categorical identification responses amplify the potential for non-memorial influences acting on witnesses' decision criteria to contribute to identification error. Social, environmental, and metacognitive influences can increase (or decrease) the likelihood a witness will pick someone, independent of the quality of the witness's memory for the culprit or the degree of match between individual lineups members and the witness's memory for the culprit (see Wells, 1993). Second, despite being an intuitively obvious method of testing a witness's memory (or an investigator's hypothesis about the guilt of a suspect), a categorical identification response is often less informative than it might appear (Sauer & Brewer, 2015). Although a categorical identification presumably indicates that the identified person best matched the witness's memory of the culprit, it provides no information on (a) the degree of match (other than that it exceeded some latent threshold) or (b) the extent to which that person was favored over the alternative lineup members. Further, a rejection provides no information about the extent to which the suspect matched the witness's memory (other than indicating that the degree of match did not exceed the threshold for identification).

In this chapter, we review the applied and theoretical motivation for a novel—and, we argue, more rigorous—approach to collecting identification evidence: Having witnesses provide, for each lineup member, a rating indicating their confidence that this person is the culprit instead of requiring a categorial identification decision. We present empirical evidence from face recognition and eyewitness identification paradigms demonstrating that, compared to categorical responses, confidence ratings provide a richer source of information about the likelihood that the culprit is in the lineup. More importantly, a ratings-based approach to assessing identification encourages an adaptive shift in the way we think about identification evidence: Moving away from perceiving an identification as a categorical indication of guilt, and towards a more nuanced, probabilistic treatment of the evidence. We consider various approaches to analyzing ratings-based identification data and examine what these approaches can tell us in both research and criminal justice settings. Finally, we consider recent research investigating how mock-jurors respond to this evidence.

Eyewitness Identification Error

Eyewitnesses make mistakes. The most obvious type of error is a false identification: The witness views a lineup and picks an innocent person. Given the established contribution of false identifications of innocent suspects to wrongful convictions (see Innocence Project, 2019), and the clear individual and societal costs associated with these errors, the eyewitness identification literature reflects how researchers have focused primarily on reducing false identifications. A less obvious, less researched, but still important problem relates to “misses”. A miss occurs when an eyewitness incorrectly rejects a lineup that contains the culprit. A miss may lead to the wrongful release of an offender, or at least make the prosecution of an offender more difficult, preventing the administration of justice and, potentially, allowing the culprit to continue offending. Misses may also undermine investigative efforts by erroneously suggesting that the initial suspect was not the culprit, and encouraging police to redirect their enquiries, resulting in time and resource mismanagement. To understand the causes of these errors, and how we might address them using a ratings-based approach to identification, we need to consider the underlying decision architecture. In the sections that follow we focus our consideration on decisions made from single-suspect lineups (i.e., lineups containing one suspect and a set of known-innocent fillers), and lineups that are “fair”[1].

Eyewitness Decision-making and Confidence

Rather than providing a detailed critique of the various models that may inform our understanding of eyewitness decision-making and confidence, we introduce the reader to two classes of theoretical frameworks—those grounded in signal detection theory and those grounded in accumulator models—that represent extensions of models originally intended to account for basic psychophysical discrimination tasks. Our intention is not to pit these accounts against each other, but to use them to (a) illustrate some key ideas about recognition and confidence that flow from the literature in general (i.e., that are applicable across both classes of theories), and (b) note the implications of these ideas for understanding the causes of eyewitness error, and potentially attenuating eyewitness error via the use of ratings-based identification evidence. In the following discussions, we focus on decisions that indicate recognition (i.e., I have seen this stimulus/person before) and the mechanisms underlying confidence in these decisions. This approach allows us to introduce the mechanisms underlying confidence in recognition that are relevant to understanding the theoretical motivation for, and proposed value of, ratings-based approaches to collecting identification evidence.

Signal detection theory. A basic recognition memory task will have participants study a list of stimuli (e.g., words or faces) and then, after some delay, make a series of recognition judgements to indicate whether a particular test stimulus (in a Yes/No recognition task), or which of a set of test stimuli (in an *n*AFC task), was present in the study list. Let us begin by considering a simple Yes/No recognition task where, for each test trial, a participant

[1] Defining the conditions that must be met for a lineup to be considered “fair” is no simple matter, but for now we use “fair” to mean “not obviously biased against the suspect”.

must decide if a single test stimulus was present at study. In its simplest form, signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 1991) holds that when attempting this task participants will compare the presented test stimulus to their memory of a stimulus presented at study (i.e., a target). This comparison will generate some evidence of recognition (i.e., degree of match between the test item and the target item[2]). The strength of this evidence will fall somewhere along a continuum from very weak to very strong; with previously studied items generating values that (generally) fall towards the higher end of the continuum and previously unseen items producing values (generally) falling towards the lower end of the continuum. According to SDT, a recognition decision is reached by comparing this evidence of recognition against a response criterion. If the evidence of recognition exceeds the criterion, the participant makes a positive recognition response (i.e., indicating that the stimulus *was* in the studied list); if the evidence of recognition does not exceed the criterion, the participant makes a negative response (i.e., indicating that the stimulus *was not* in the studied list). Confidence, then, indexes the extent to which the evidence of recognition exceeds the response criterion. Thus, the stronger the evidence of recognition, the greater the extent to which it will exceed the response criterion, and the higher confidence will be. This evidential basis for confidence produces the well-established relationship between confidence and accuracy for recognition decisions. As discussed in greater detail later in this section, this mechanism for confidence also hints at an alternative use for confidence: as an index for recognition in the absence of an explicit categorical decision.

This basic architecture can then be extended to account for more complex, compound recognition decisions, such as those in the eyewitness identification, where a participant must

[2] This value has been referred to as a likelihood ratio or function of a likelihood ratio (e.g., Green & Swets, 1966; Stretch & Wixted, 1998), as an index of stimulus familiarity (e.g., Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996), or as a strength effect (e.g., Balakrishnan & Ratcliff, 1996).

decide which *if any* of an array of test stimuli has been previously seen (e.g., Horry, Palmer, & Brewer, 2012; Palmer & Brewer, 2012; Smith, Wells, Lindsay, & Penrod, 2017). For example, when viewing a lineup, an initial inspection of the lineup members may reveal one lineup member that the witness believes might be the perpetrator. The witness can then compare this individual to their memory of the offender, generating some evidence of recognition. The strength of this evidence, when compared to the witness's response criterion, determines the identification response (i.e., an identification or lineup rejection), and confidence in this response.

Accumulator models of recognition. In the context of a simple Yes/No recognition decision, accumulator models generally assume that a comparison of a test stimulus with a memorial representation of a studied stimulus will provide evidence favoring either of two response alternatives: this stimulus *was* in the studied list or this stimulus *was not* in the studied list. Evidence favoring each of these two alternatives is stored in independent accumulators. Each accumulator has a preset criterion and a response is made when the evidence in one of the accumulators exceeds the criterion (e.g., Van Zandt, 2000). Accumulator models typically employ a balance of evidence mechanism to explain confidence (Vickers, 1979). According to this view, confidence indexes the difference between the amounts of evidence in the competing accumulators when the response is made. If the difference between the amounts of evidence in the two stores is large, confidence will be high; if the difference is small, confidence will be low. Thus, in a Yes/No recognition task, if the participant has a strong memory for a studied item and a test item provides a high degree of match to that memory, we would expect the evidence favoring a *Yes* response (i.e., this stimulus *was* in the studied list) to far exceed the evidence favoring a *No* response (i.e., this stimulus *was not* in the studied list), thus producing high confidence in recognition.

Again, this logic can be extended to more complex recognition tasks involving choices from arrays that may or may not include a target, as in the eyewitness identification domain (see Horry & Brewer, 2016). When viewing a lineup, a witness will compare lineup members to their memory of the culprit, and probably to each other in terms of relative similarity to this memory. These comparisons will accumulate evidence of recognition, stored in separate accumulators for each lineup member. When one accumulator reaches criterion, a decision is made. Confidence then indexes the evidence for identified lineup member relative to the evidence for the non-identified lineup members (Horry & Brewer, 2016). Again, a strong memory for the culprit and a high degree of match between the favored lineup member and that memory will produce (a) a greater discrepancy between the evidence for the identified and non-identified lineup members, and (b) higher confidence in recognition. A separate accumulator will also be required for a “Not present” response such that, if evidence in this accumulator reaches criterion first, the witness rejects the lineup. However, as mentioned, our focus here is on confidence in recognition as this provides the foundation for understanding ratings-based identification procedures.

Recognition, decision-making, and confidence. Although the frameworks introduced above differ in terms of the specific mechanisms thought to underly an identification decision and confidence in recognition, there are a number of important implications shared by these accounts. First, recognition is not an “all or nothing” process. Instead, according to both signal detection and accumulator frameworks, a categorical recognition *decision* is the result of comparing some latent strength-of-recognition variable—representing a point on an evidential continuum—against a response criterion.

Second, therefore, there are multiple mechanisms that can contribute to error. Some identification errors—false identifications or misses—might reflect task difficulty and the imperfect nature of memory. For example, poor encoding conditions, long retention intervals, or substantial changes in the culprit's appearance might lead to weak memories or otherwise undermine the matching process that drives the decision, leaving the witness unable to recognize a culprit who is present in the lineup, and producing a miss. Alternatively, a highly plausible but innocent suspect might produce a very strong sense of recognition, increasing the risk of a false identification. Essentially, some conditions will contribute to errors by affecting the extent to which a lineup member (be it a target or an innocent suspect) *matches* the witness's memory of the culprit (i.e., the strength of the recognition experience). However, once we consider the decision architecture described above—namely, that strength of recognition is a continuum and that identification decisions reflect evidence of match *relative* to malleable response criteria—it seems likely that at least some errors will reflect the witness's placement of their decision criterion, rather than fundamental problems with the quality of their memory or the matching process. If the witness sets an overly lenient response criterion—for example, because they approach the task with a strong assumption that the guilty party is likely to be present in the lineup, and that their "job" is to pick someone—the risk of a false identification will increase. If the witness is too conservative in their criterion placement—because, for example, they appreciate that false identifications can contribute to miscarriages of justice, or because they generally doubt their ability to recognize people—the risk of a miss increases. In more concrete terms, a miss might occur because the witness simply does not recognize the culprit in the lineup (e.g., because the witness's memory for the original event is weak, or because the culprit's appearance has changed substantially since the original event). Alternatively, a miss might occur because the witness *thinks* they recognize a lineup member, but this sense of recognition is not strong enough to exceed the decision criterion for an identification. These later errors might be thought of as more akin to decision-making errors than "failures" of recognition memory.

An attendant implication here is that in some cases a witness might possess useful memorial information that is not captured by a categorical identification response. More specifically, for choices from arrays, where a rejection reflects the outcome of multiple underlying comparisons, collapsing these multiple comparisons into a single categorical response may obscure useful information about the extent to which individual lineup members—and the police investigator's suspect in particular—match the witness's memory for the culprit. Consider the following three situations where, after viewing a crime, a witness views a lineup in which lineup member number 4 is the police suspect: (1) Inigo views a lineup, does not recognize anyone, and rejects the lineup, (2) Fezzik views a lineup, *thinks* "Number 4" looks a lot like the culprit, but is not confident enough to *identify* "Number 4" as the culprit, and (3) Westley who views a lineup, thinks both "Number 4" and "Number 6" look very similar to the culprit, but cannot decide between them and therefore rejects the lineup. In all of these cases, the categorical response outcome is the same: the witness rejects the lineup. And, in none of the cases, is the response "wrong": if the witness is not sufficiently confident in their recognition, it is appropriate to reject the lineup. However, the categorical rejection outcome obscures differences in the recognition states underlying each witness's response. Moreover, if "Number 4" is the police suspect, these differences in the underlying recognition states may have important implications for our understanding of the likely guilt of the suspect. For example, although neither Fezzik nor Westley categorically identified "Number 4", the fact that both thought this lineup member looked similar to the culprit might[3] indicate that the investigator's suspect is worthy of further investigation (cf. the intuitive interpretation that a rejection indicates either that the bad guy is not in the lineup, or that the witness did not recognize the police suspect). Critically, however, the ability to

Commented [JS1]: need to put this in rhyme form.

[3] To foreshadow results discussed later in this chapter, our data certainly suggest this is the case

appreciate this nuance is lacking from the categorical rejections produced by traditional identification tasks[4].

The Informational Value of Identification Evidence

At this stage let us take a step back and think about the purpose of a lineup. A lineup is designed to help police test their hypothesis that their suspect is the culprit (Wells & Luus, 1990). Thus, an effective lineup task must provide information that addresses this question: chiefly by determining the extent to which the witness recognizes the suspect. To what extent do, and can, traditional lineups achieve this aim? There are two ways to think about this question. First, we must consider the extent to which the outcome of a lineup speaks to the underlying construct of interest. Second, we must think about how often a lineup provides diagnostic information.

If our construct of interest is the extent to which the suspect matches the witness's memory of the culprit then, to be valuable, an identification decision must reflect a comparison of individual lineup members with a witness's memory of the culprit[5]. Ideally, a strong memory for the culprit and a strong match between this memory and a lineup member should produce a positive identification. A weak memory or poor matches between the lineup members and the witness's memory should lead the witness to reject the lineup. However, as discussed in the previous section, non-memorial and meta-cognitive influences can operate on witnesses' decision criteria to compromise the fidelity with which the eventual decision indexes the strength of the underlying match. An overly lenient criterion may lead a relatively weak match to produce an identification (increasing the risk of a false identification) and an overly conservative criterion may mean that even a relatively strong match fails to produce an

[4] Moreover, the robust demonstrations of non-systematic confidence-accuracy relations for lineup rejections suggest that a retrospective confidence rating following a rejection is unlikely to help here.

[5] This is not to imply that an identification should be based on a deliberative comparative process rather than a more automatic recognition process. Our point is that the identification decision should be driven primarily by the degree of match between a lineup member and the witness's memory of the culprit rather than non-memorial factors that may contribute to the evidence favouring a decision, or the placement of decision criteria.

identification (increasing the risk of a miss). We should note that, although we use terms like "overly lenient" and "overly conservative", a witness who adopts a lenient or conservative criterion is not necessarily "wrong" to do so, even if the placement of the criterion is sub-optimal: There are valid reasons for witnesses to, for example, be cautious in identifying a lineup member. The key point is that the malleability of witnesses' decision criteria can contribute to errors in categorical identification decisions, and reduce the extent to which the outcome reflects the match between suspect and a witness's memory of the culprit.

Let us now consider the second question—the likelihood that a given lineup will produce useful information about the likely guilt of the suspect—with reference to three possible outcomes: a suspect identification, a filler identification, and a lineup rejection. A positive identification of the suspect provides, at face value, the clearest example of useful information. However, this outcome is less informative than it may appear. It is reasonable to assume that a suspect identification indicates that, of the presented lineup members, the suspect provides the best match with a witness' memory of the culprit. However, the aforementioned effects on a witness's decision criterion can increase the chances of a suspect identification without affecting the degree of match between the suspect and the witness's memory for the culprit. Thus, although a suspect identification probably indicates that the suspect provided the best available match, it is much less informative about the quality of this match in an absolute sense (other than that the strength of recognition exceeded a decision criterion that cannot be determined for a single identification decision). The informational value of a filler identification is less clear. There is some evidence that filler identifications can be diagnostic of suspect innocence—because another lineup member provided a better match to the witness's memory of the culprit—but the evidence in support of this notion is mixed (e.g., Clark & Wells, 2008; Wells & Olson, 2002; Wells, Yang, & Smalarz, 2015). Finally, what does a rejection tell us about the likely guilt of a suspect? Again, there is some evidence that rejections provide exculpatory information (e.g., Wells et al., 2015) but, given the aforementioned variation in recognition states that can produce a categorical rejection response, we argue that, on its own, the informational value of a rejection is limited when it comes to assessing the guilt of a suspect.

Given that at least some errors from traditional identification procedures are likely to reflect non-memorial influences on criterion placement, and some decisions are likely to provide minimal (or even obscure potentially) useful memorial information, we must ask: is it possible to attenuate criterion-related errors and obtain a more direct index of the construct of interest (the extent to which the police suspect matches the witness's memory of the culprit)? Thus, our motivations for testing a ratings-based approach to identification evidence were twofold. First, we hoped to reduce the impact of criterion-related noise on decision-making, reduce errors resulting from idiosyncratic influences on witnesses' criterion placement, and gain a more direct index of the construct of interest: The extent to which the suspect matches the witness's memory for the culprit. Second, we sought to improve the frequency with which lineups produce informative outcomes.

Using Confidence Ratings to Index Recognition

We now discuss what may appear to be a radical departure from traditional identification practice: using ratings rather than categorical identification decisions to index recognition. In this section, we discuss various paradigms that have been used to investigate the utility of ratings-based approaches to assessing recognition and identification, and some analytical approaches employed to convert ratings-based recognition data into formats that can be informative about the likely guilt of a suspect. We also consider the questions these approaches can answer for researchers and for decision-makers in legal settings. Clearly, a ratings-based approach involves a drastic change in the way identification evidence is collected and interpreted, but we argue that it provides a more informative index of recognition, and represents a more valid approach to “obtaining” probative information from witness memory (Brewer, Weber, & Guerin, 2019). We also note two further points. First, from a psychological perspective, ratings-based measures of recognition have a long tradition in the basic memory research literature (Egan, 1958). Thus, it is the area of application, rather than the approach itself, that is novel here. More broadly, a basic appreciation of psychometrics suggests that the idea that a latent construct (recognition) can best be assessed with a single datapoint (a categorical identification decision) is questionable, and that multiple ratings are likely to provide a richer, more informative index of the participant or witness’s underlying recognition. Second, traditional approaches to collecting identification evidence contribute to high error rates (e.g., Steblay, Dysart, & Wells, 2011) and, for one reason or another, often tell us relatively little about our key construct of interest (the extent to which the witness recognizes the suspect).

As discussed previously, theories of how confidence estimates are derived for recognition memory hold that confidence indexes the degree of match between a presented item and an image in memory. This suggests an interesting possibility for assessing witness recognition. Specifically, avoiding categorical identification responses (and thereby attenuating idiosyncratic influences on witnesses’ decision criteria) and, instead, simply asking the witness to rate their confidence (on a scale from 0-100%) that each lineup member is the culprit may provide a more sensitive and informative index of recognition, and a more direct assessment of the degree of match between individual lineup members and the witness’ memory of the culprit (Brewer et al., 2019; Brewer, Weber, Wootton, & Lindsay, 2012; Sauer, Brewer, & Weber, 2008, 2012a; Sauer, Weber, & Brewer, 2012b).

Preliminary empirical support for the potential diagnostic value of confidence ratings in the absence of categorical recognition decisions came from a series of word recognition studies demonstrating the ‘recognition without identification phenomenon’ (Cleary, 2002;

Cleary & Greene, 2000, 2005; Cleary, Langley, & Seiler, 2004; Peynircioglu, 1990). In these experiments, participants studied a list of words before, at test, words or word fragments were presented very briefly on a computer screen. For each test stimulus, participants attempted to identify the word and rated their confidence that this word came from the studied list. The recognition without identification phenomenon refers to the finding that, even when participants were unable to identify the displayed word, higher confidence ratings were consistently given to previously studied words than to non-studied words. Thus, differences in confidence ratings were to some extent able to differentiate between previously viewed and previously unseen items, even when participants could not reliably discriminate between studied and unstudied items based on categorical responses. Despite differences between the cognitive processes underlying face and word recognition tasks (Farah, Wilson, Drain, & Tanaka, 1998; McKone, 2004), these findings provided some early empirical motivation for testing the utility of ratings-based evidence in face recognition and eyewitness identification tasks.

Early investigations of the diagnostic value of ratings-based indices of face recognition and identification also provided a number of encouraging findings. To understand these findings, it is necessary to give some consideration to the paradigms used to collect the data. Basic face recognition paradigms (Sauer et al., 2008, Expt 1; Sauer et al., 2012b) involved participants completing blocks of trials, where each block included a study phase (a series of faces presented sequentially), a retention interval of a few minutes, and a test phase (series of faces presented sequentially with individuals indicating, for each face, whether it was viewed in the previous study list). At test, participants responded with either a categorical (Yes/No) decision (control condition) or provided a confidence rating for each face. From these experiments, a number of key findings emerged. First, when participants simply rated their confidence that a face had been studied, there was a generally linear, positive relationship between these ratings and the likelihood that a face had been previously seen (Sauer et al., 2012b). Second, we were able to develop and apply a classification algorithm (based on an approach used by Koriat & Goldsmith (1996) and described in detail in Sauer et al., 2008) to determine when a confidence rating could be treated as a positive recognition decision; that is, an algorithm that allowed us to collapse confidence ratings into a categorical index of recognition. As discussed later, reducing confidence ratings down to categorical responses is not necessarily the most informative use of these data, or the approach we would recommend for applied settings, but it was a useful initial approach to testing the utility of ratings-based evidence against a control, categorical response condition. Ratings-based measures of recognition consistently produced higher classification accuracy than categorical recognition decisions in these face recognition tasks. That is, confidence ratings were more effective than participants' decisions at indicating whether or not participants had seen a presented face before. Further, analyses of classification performance using SDT-based measures of discrimination (i.e., the sensitivity of the index of recognition) and response bias (i.e., the general tendency to return an increased or decreased number of positive classifications) demonstrated that the improved classification accuracy associated with the ratings-based approach (cf. categorical decisions) was attributable to improvements in discrimination rather than any change in bias (Sauer et al., 2012b).

We then extended this work to determine whether ratings-based approaches to recognition could effectively discriminate seen from unseen faces in arrays. Determining whether a pattern of confidence ratings could be taken as indicating a categorical recognition decision was more complicated than determining the criterion described above for the single confidence ratings generated in a standard face recognition paradigm. Across a number of experiments, we developed and tested a variety of algorithms (described in Sauer et al., 2008), but the best two involved (a) determining whether there was a maximum confidence value present in the ratings provided (i.e., whether one member in the array was favored over the others) and, if so, (b) considering the extent to which this array member was favored over the other (considering the max rating either in relation to the next-highest rating, or the average of the non-max ratings). We also extended these algorithms to incorporate the possibility for confidence ratings to return indeterminate responses (i.e., to acknowledge that the available recognition information was not strong to indicate either a positive identification or a rejection of the lineup, see Sauer et al., 2008). Across a series of experiments, using both simultaneous and sequential lineups, we found generally consistent evidence that confidence ratings were more effective than participants' decisions at indicating whether or not presented lineups members had been seen before (Brewer et al., 2012; Sauer et al., 2008). That is, compared to participants' recognition decisions, confidence ratings provided a more sensitive index of recognition. Subsequent work has also extended this procedure to child witnesses and shown that, although confidence ratings do not necessarily confer an advantage over categorical responses when ratings are collapsed in categorical classifications, child witness can use ratings-based approaches to identification (Bruer, Fitzgerald, Price, & Sauer, 2017; see also Hiller & Weber, 2013). Even in the absence of a significant advantage for the ratings-based (cf. categorical) approach in this study, the fact that child witnesses could use a ratings-based approach is encouraging. The value here will become evident in the following sections when we discuss (a) alternative treatments of ratings-based identification data and (b) the flexibility these approaches provide in terms of interpreting identification evidence in the broader context of a case. Specifically, this flexibility may help offset the pervasive tendency of child witnesses to choose too often from target-absent lineups (see discussion of Brewer et al., 2019).

Collapsing ratings into categorical classifications—via the algorithm approach described in Sauer et al. (2008) and Brewer et al. (2012)—allowed us to compare the diagnostic value of ratings-based identification information with performance based on categorical recognition decisions. From a research perspective, this allowed a clear, initial demonstration that participants were not making optimal use of the memory information available to them (i.e., that, using the same memorial information, we could improve classification accuracy if we controlled the criterion for a positive classification). However, this approach also reduced the richness of the recognition information provided. More recent analytical approaches have embraced the richness of the data provided by a ratings-based procedure, and suggested treatments of the data that will be more informative in applied contexts. For example, Brewer et al. (2012) determined, for each lineup, whether there was a single highest, maximum confidence value. For cases that included a maximum confidence value, the researchers examined variations in the likely guilt of the suspect as a function of the discrepancy between this maximum value and next-highest value. Based on this profile analysis approach, two important results emerged across three experiments. First, consistent with theoretical frameworks that view confidence in recognition as an index of the relative similarity of decision alternatives to a previously viewed item (Horry & Brewer, 2016), the likely guilt of the suspect increased almost monotonically as a function of the discrepancy between the maximum and next-highest confidence ratings. Second, when this discrepancy was large (e.g., ≥ 80%) the likely guilt of the suspect was very high (e.g., 80-100%) and, until the discrepancy fell to 30-50%, confidence ratings were a better predictor of suspect guilt than were categorical identification decisions.

Extending these findings, Brewer et al. (2019) charted how the probability of suspect guilt varied as a function of the *max* confidence rating provided for an array. A number of elegant and intuitive findings emerged. First, if the suspect did not receive the max value they were more likely innocent than guilty. Second, if the suspect alone received the max rating, the likely guilt of the suspect increased systematically and monotonically with the magnitude of the max rating (extending the relationship previously demonstrated in a face recognition paradigm to choices from arrays). Brewer et al. also demonstrated that variations in the suspect confidence ratings provided by child witnesses predicted the probable guilt of the suspect.

As well as these intuitive findings, Brewer et al. reported some striking, and initially counter-intuitive results. Specifically, variations in max confidence ratings offered useful information about the likely guilt of the suspect in a number of conditions under which a traditional procedure would have been unlikely to return a categorical suspect identification. For example, when the max confidence value was relatively low (e.g., if the suspect received a max rating of 30, provided the value was unique to the suspect, probable guilt was above 60%[6]). Thus, probative information may be obtained from witnesses who for a variety of reasons (e.g., perhaps they have reservations about the quality of their memory for the culprit or, from a dispositional perspective, tend to have reservations about judgments they make) may not feel confident enough to choose the best matching lineup member.

And when the witness gave a max value to multiple lineup members, or even when the max was given to a filler but the suspect also received a high rating, variations in the confidence rating given to the suspect were reliably associated with the probable guilt of the suspect, albeit with lower estimates of probability of guilt than obtained when the suspect alone received the max value. The last two scenarios—involving multiple max values or cases where the suspect was not the max but did receive a high rating—appear counter-intuitive but actually serve to highlight the benefits of a ratings-based approach in terms of working with, rather than constraining recognition memory. The idea that a witness might be unable to discriminate between two, highly plausible alternatives, or that a witness might

[6] Although Brewer et al.'s analyses determine specific probabilities of guilt associated with specific max values, we caution against the temptation, evident in some recent treatments of retrospective confidence judgments, to provide firm estimates, or an absolute sense, of the likely guilt of a suspect given a particular confidence rating.

think someone other than the suspect best matches their memory of the culprit might, at first glance, appear to suggest problems with the witness's memory or their ability to discriminate effectively between response alternatives. Along these lines, such memory states may lead to misses or filler identifications from traditional lineup procedures. However, when we consider that recognition is continuous, and that the ratings given to different lineup members do not represent the outcome of a zero-sum game, it is perfectly conceivable that (a) in some cases, the suspect *and* a filler might both provide a strong match to the witness's memory of the culprit, and (b) this need not preclude a systematic relationship between the rating given to the suspect and the probable guilt of that suspect.

Brewer et al.'s (2012, 2019) findings have two important implications. First, they show that patterns of confidence ratings can offer reliable diagnostic information about suspect guilt for individual witnesses, without the need to sacrifice the richness of the data by collapsing these patterns down to a single, categorical classification. Second, the monotonic positive relationship between the probability of suspect guilt and (a) Brewer et al.'s (2012; 2019) discrepancy measure and (b) max values in Brewer et al.'s (2019) study suggest the plausibility of avoiding categorical classifications entirely, in favor of a probabilistic treatment of identification evidence. This is further supported by Brewer et al.'s (2019) demonstrations that ratings-based approaches to identification provide diagnostic information about suspect guilt even in conditions under which a traditional lineup procedure would likely have returned a rejection, or even a filler identification.

Categorical approaches to identification reduce a witness' recognition memory 'output' to categories (e.g., "It's that guy" or "He's not there" or "I'm not sure"). Based on the findings reported above, we argue the legal system can benefit from a ratings-based approach in two ways when considering what this information says about the *likely* guilt of the suspect/defendant. First, it maximizes the amount of information available about the strength of the witness' recognition of the suspect. Second, it allows investigators to interpret this identification evidence alongside other evidence. For example, imagine a situation where the ratings-based approach returns modest evidence against the suspect in terms of the magnitude of the rating given to the suspect, or the extent to which the suspect if favored over the alternatives. If the police have no other evidence against the suspect, this identification may indicate the merit of pursuing the current course of the investigation, but not warrant prosecuting the suspect at present. However, the same ratings-based identification evidence, if paired with converging evidence of the suspect's guilt, might warrant charging and prosecuting the suspect. Thus, when compared to traditional approaches to identification, a ratings-based approach (a) provides more information and (b) better allows for the probabilistic—and more appropriately nuanced—interpretation of this information alongside other evidence in the case.

The boundary conditions for ratings-based approaches to identification, and the best ways to interpret the information they provide, require further investigation. Nonetheless, this approach seems to have the potential to address a number of the problems that have plagued traditional identification practices. First, in line with our first aim, avoiding categorical identification decisions attenuates non-memorial influences on criterion placement, and better indexes the degree of match between the suspect and the witness' memory of the culprit (i.e., the construct of interest). Second, this approach provides legal decision-makers with a richer source of information upon which to base assessments of likely guilt. As mentioned previously, a single decision provides little information about the extent to which the identified lineup member matches the witness' memory of the culprit. Further, if the suspect is not identified, the eventual decision provides no information about the degree of match between the suspect and the witness' memory of the culprit (other than that the degree of match did not exceed the criterion for identification in the case of a lineup rejection, or that the suspect was not the best match in the case of a foil identification). In contrast, in line with our second aim, in all cases a ratings-based approach provides investigators with useful information about the extent to which the suspect matches the witness' memory of the culprit. Further, Brewer et al. (2019) demonstrate that this information can be diagnostic of suspect guilt even when the underlying recognition state may have prompted a rejection, or a potentially-erroneous guess, from a traditional lineup (e.g., when two lineup members are equally and highly similar to the witness's memory of the culprit).

Brewer et al. (2019) argued that, despite the possibility that a ratings-based procedure is likely to encounter strong resistance from the police and courts, "Reducing the likelihood of people being convicted of crimes they did not commit and of criminals avoiding detection and conviction are, in our view, objectives that warrant not only the careful consideration of psychological researchers but also concerted attempts on their part to translate the relevant psychological science into policy reform". Further, echoing Brewer et al.'s (2019) conclusions, we contend that a ratings-based approach to collecting and interpreting identification evidence offers an opportunity to abandon the problematic conceptualization of identification evidence as some absolute indication of guilt in favor of a conceptualization more aligned with out scientific understanding of recognition memory: namely, as yet another piece of probabilistic evidence, which must be considered alongside the available corroborating evidence when evaluating the likely guilt of a suspect/defendant. A key question then becomes: How do triers of fact respond to this evidence?

How do jurors respond?

Jurors face a difficult task. Jurors must synthesize complex, ambiguous, and often contradictory information in order to reach a decision with substantial consequences. Given these conditions, it is not surprising that eyewitness identification evidence—where a witness is prepared to provide, under oath, a categorical indication that they recognize the defendant as the culprit—is so compelling: In an environment filled with ambiguity, a categorial identification provides an apparently unambiguous indication of the defendant's guilt. However, this *perceived* clarity is likely to be at least partly to blame for the established contribution of identification error to wrongful conviction. As discussed earlier in this chapter, categorical identifications are both prone to error, and less informative than they appear. By providing a richer source of information about the extent to which the defendant matches the witness's memory for the culprit—in both an absolute sense, and relative to other lineup members—ratings-based identification evidence may assist jurors to make better use of identification evidence presented to them.

Specifically, a ratings-based approach can provide three types of information that may be useful for jurors, but lacking from a categorical identification. First, compared to a categorical identification, ratings-based identification provides a more nuanced assessment of the strength with which the suspect-come-defendant matched the witness's memory of the culprit. Second, a ratings-based approach provides information about the extent to which the suspect-come-defendant was favored over the other lineup members. Although Brewer et al.'s (2019) recent work demonstrates that a single max value is not a necessary pre-condition for ratings reliably indicating suspect guilt, it is true that a pattern of ratings where the suspect receives a high rating and all other lineup members receive low ratings is likely to indicate a strong memory for the suspect *and* good ability to discriminate between the suspect and other lineups members (assuming the lineup is fair). Finally, consistent with proposed benefits for investigators' decision-making, a ratings-based approach to identification evidence allows jurors to interpret the strength of identification evidence alongside the other evidence presented. When we consider that preparing a witness for cross-examination can inflate the witness's confidence in their testimony (Shaw, McClure, & Dykstra, 2007), it is likely that typical presentations of identification evidence in court imply a level of certainty that misrepresents the underlying recognition state at the time the identification is made. In contrast, a ratings-based approach allows jurors to appreciate the witness's recognition of the suspect in a more appropriately nuanced way.

Thus, there are a number of reasons why ratings-based evidence may help jurors make a more nuanced evaluation of a witness's identification evidence. However, it is also possible that jurors might balk at identification evidence that does not involve a witness categorically identifying the suspect, or that jurors may simply lack the necessary framework for interpreting ratings-based identification evidence. The relationship between confidence ratings and underlying memory states might be obvious to cognitive psychologists, but not to members of the broader community.

With these considerations in mind, we recently tested how mock-jurors interpreted ratings-based identification, attempting to answer two key questions: (1) are mock-jurors resistant to noncategorical identification evidence? and (2) can mock-jurors interpret and apply ratings-based evidence in an adaptive way? Across a number of experiments, we presented mock jurors with written vignettes involving a witness providing evidence relating to a crime and their identification of suspect. We manipulated the patterns of ratings in two ways. First, the rating given to the suspect was either high (90%) or moderate (50%). Second, the ratings given to non-suspect lineup members evidenced either good or relatively poor discrimination (i.e., indicating that the suspect was favored over the alternatives by a large or small extent, respectively). We compared perceptions of defendant guilt in these ratings-based conditions with perceptions of guilt in conditions presenting mock-jurors with a categorical identification alongside high (90%) or moderate (50%) confidence.

When comparing mock jurors' perceptions of defendant guilt based on categorical vs. ratings-based identification evidence, three key findings emerged. First, across experiments, when the rating given to the suspect was high, there was no evidence that the absence of a categorical identification decision undermined the persuasiveness of the evidence against the defendant. There was no evidence that mock jurors simply dismissed ratings-based identification evidence. Second, when considering evidence provided by a single witness, mock-jurors seemed insensitive to the additional information (i.e., ratings given to non-suspect lineups members) provided in the ratings-based evidence conditions. However, when they were able to compare the patterns of ratings provided by three different witnesses—one providing a categorical identification, another providing ratings showing good discrimination, and a third showing poor discrimination—mock-jurors' evaluations of defendant guilt were sensitive to evidence suggesting the witness was able to clearly discriminate between the suspect and foils in the lineup. Finally, when provided with instructions that guided their interpretation of patterns of ratings, mock-jurors were able to apply this information in an adaptive way (i.e., to benefit from the additional information provided by non-suspect ratings) when evaluating evidence provided by a single witness.

Conclusion

Traditional approaches to collecting identification evidence are sub-optimal in terms of (a) the frequency with which they produce decision errors, (b) the validity with which the decisions they produce index the underlying construct of interest, and (c) the informational value of the decisions they produce. We argue there is solid theoretical support, and a growing body of robust empirical support, for an alternative approach grounded in the science of recognition memory: the use of ratings-based identification evidence. Although a shift to a ratings-based approach to identification requires a fundamental change in the way decision-makers must think about identification evidence, we contend that this approach has the potential to ultimately improve the utility of identification evidence and avoid at least some of the costly consequences of identification error.

References

Balakrishnan, J., & Ratcliff, R. (1996). Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 615-633.

Brewer, N., Weber, N., & Guerin, N. (2019). Police lineups of the future? *American Psychologist*. doi:doi.org/10.1037/amp0000465

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science, 23*(10), 1208-1214. doi:10.1177/0956797612441217

Bruer, K. C., Fitzgerald, R. J., Price, H. L., & Sauer, J. D. (2017). How sure are you that this is the man you saw? Child witnesses can use confidence judgments to identify a target. *Law & Human Behavior, 41*, 541-555.

Clark, S. E., & Wells, G. L. (2008). On the diagnosticity of multiple-witness identifications. *Law and Human Behavior, 32*(5), 406-422. doi:http://dx.doi.org/10.1007/s10979-007-9115-7

Cleary, A. M. (2002). Recognition with and without identification: Dissociative effects of meaningful encoding. *Memory & Cognition, 30*, 758-767.

Cleary, A. M., & Greene, R. L. (2000). Recognition without identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1063-1069.

Cleary, A. M., & Greene, R. L. (2005). Recognition without perceptual identification: A measure of familiarity? *The Quarterly Journal of Psychology, 58*, 1143-1152.

Cleary, A. M., Langley, M. M., & Seiler, K. R. (2004). Recognition without picture identification: Geons as components of the pictorial memory trace. *Psychonomic Bulletin & Review, 11*, 903-908.

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

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review, 105*, 482-498.

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

Hiller, R. M., & Weber, N. (2013). A comparison of adults' and children's metacognition for yes/no recognition decisions. *Journal of Applied Research in Memory and Cognition, 2*(3), 185-191. doi:http://dx.doi.org/10.1016/j.jarmac.2013.07.001

Horry, R., & Brewer, N. (2016). How target-lure similarity shapes confidence judgments in multiple-alternative decision tasks. *Journal of Experimental Psychology: General, 145*(12), 1615-1634. doi:10.1037/xge0000227

Horry, R., Palmer, M. A., & Brewer, N. (2012). Backloading in the sequential lineup prevents within-lineup criterion shifts that undermine eyewitness identification performance. *Journal of Experimental Psychology: Applied, 18*(4), 346-360. doi:10.1037/a0029779

Innocence Project. (2019).   Retrieved from http://www.innocenceproject.org

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

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* New York: Cambridge University Press.

McKone, E. (2004). Isolating the special component of face recognition: Peripheral identification and a mooney face. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 181-197.

Palmer, M. A., & Brewer, N. (2012). Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law and Human Behavior, 36*(3), 247-255. doi:10.1037/h0093923

Peynircioglu, Z. F. (1990). A feeling-of-recognition without identification. *Journal of Memory and Language, 29*, 493-500.

Sauer, J. D., & Brewer, N. (2015). Confidence and accuracy of eyewitness identification. In T. Valentine & J. P. Davis (Eds.), *Forensic facial identification: Theory and practice of identification from eyewitnesses, composites and cctv* (pp. 185-208). Chichester: Wiley Blackwell.

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., Brewer, N., & Weber, N. (2012a). Using confidence ratings to identify a target among foils. *Journal of Applied Research in Memory and Cognition, 1*(2), 80-88. doi:10.1016/j.jarmac.2012.03.003

Sauer, J. D., Weber, N., & Brewer, N. (2012b). Using ecphoric confidence ratings to discriminate seen from unseen faces: The effects of retention interval and distinctiveness. *Psychonomic Bulletin & Review, 19*(3), 490-498. doi:10.3758/s13423-012-0239-5

Shaw, J. S., III, McClure, K. A., & Dykstra, J. A. (2007). Eyewitness confidence from the witnessed event through trial. *The handbook of eyewitness psychology, Memory for events, 1*, 371-397.

Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2017). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law & Human Behavior, 41*(2), 127-145.

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology Public Policy and Law, 17*(1), 99-139. doi:10.1037/a0021650

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning Memory, & Cognition, 24*, 1397-1410.

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

Vickers, D. (1979). *Decision processes in visual perception*. New York: Academic Press.

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

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

Wells, G. L., Yang, Y., & Smalarz, L. (2015). Eyewitness identification: Bayesian information gain, base-rate effect equivalency curves, and reasonable suspicion. *Law & Human Behavior, 39*(2), 99-122.

Yonelinas, A. P., Dobbins, I., Szymanski, M. D., Dhaliwal, H. S., & King, L. (1996). Signal-detection, threshold, and dual-process models of recognition memory: Rocs and conscious recollection. *Consciousness and Cognition, 5*, 418-441.