Developing a Strong(er) Theory of Eyewitness Memory: The Selection, Verification, and Application of Mathematical Models of Identification Decisions

Thesis submitted in fulfillment of the requirements

for the Degree of Doctor of Philosophy (Medicine)

Kym Michelle McCormick

School of Psychology,

Faculty of Health and Medical Sciences

The University of Adelaide

Submission date: July 2022

In loving memory of

Alan Morris Brinkworth (1946—2022)

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

The author acknowledges that copyright of published works contained within the thesis resides with the copyright holder(s) of those works. I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship and by an Australian Research Council grant DP160101048.

Kym Michelle McCormick

9th July, 2022

Abstract

Various formal and informal models of eyewitness memory have been proposed. While serving to guide both the construct and analytical frameworks of research within the field, these models have yet to be critically tested through a process of empirical falsification. This thesis addresses this gap by critically testing four hypotheses: the hypotheses that eyewitness memory possesses both (1) random-scale and (2) monotonic-likelihood representation; the hypothesis that eyewitness memory data is (3) accurately predicted by high-threshold (HT) models; and the hypothesis that a mathematical model of eyewitness identification provides a (4) good representation of the psychological constructs of eyewitness memory and decision making. After investigating the Block-Marschak inequalities test for random-scale and monotonic-likelihood representation and developing a new critical test for the falsification of the high threshold (HT) models, two experiments were conducted online with a total of 5,056 participants recruited from Amazon Mechanical Turk. Experiment 1 collected k-AFC probabilities for lineup sizes $k \in \{2, \ldots, 7\}$. Experiment 2 collected identification and ranking probabilities from a simultaneous 8-item lineup using a 3 (strong, weak, very weak memory) x 2 (low vs high expectation) x 2 (target-present vs target-absent) between-subject experimental design. Eyewitness identification outcomes were shown to have both random-scale and monotonic likelihood representation, thus allowing for development of a mathematical model. The 2HT models of eyewitness memory were falsified and superseded by an alternative surviving model—signal detection theory (SDT). Finally, the predictive ability of the unequal-variance (UV) SDT model of simultaneous lineup identification (assuming a

MAX decision rule) was confirmed, as was the independence of the model's parameters and its generalizability across task structures. It was concluded that the UV-SDT class of models provide an evidence-based account of eyewitness identification behavior, support the measurement of empirical eyewitness identification data, and have facilitated a shift towards the building of stronger scientific evidence.

Acknowledgements

I would like to thank Associate Professor Carolyn Semmler for her support throughout my candidature as my primary supervisor. Her knowledge of the field of study within which I chose to work—eyewitness memory—was vitally important as it ensured that this work remained both grounded within, and relevant to, the applied field. Her continued support also extended beyond what would normally be expected, and for that, I am eternally grateful. I would also like to thank Dr. Melissa Humphries who came to my rescue in times of great need during the final phase of my candidature. As co-supervisor, she worked tirelessly to ensure that I developed the technical expertise necessary to complete this thesis. As a friend and as a colleague, her tireless support encouraged me to persevere through adversity, time and time again. Similar professional support was also provided by Professor Anna Chur-Hansen (Head of School) and Diana Dorstyn (HDR Coordinator), who each provided vital counsel and encouragement.

I also acknowledge the following people, each of whom contributed to this thesis. Firstly, to Dr. Keith Ransom for his support in developing and implementing the software created for the collection of the experimental data. Keith also generously taught me how to collect human data using online platforms, such as Amazon Mechanical Turk.

Secondly, to Associate Professor David Kellen (Syracuse University) for his valuable advice in relation to the role of mathematical models within the applied sciences. As a researcher, David is very generous. He freely provided accessible versions of his R code via OSF, spoke openly about his theoretical ideas, and kindly explained how to conduct a range of complex statistical bootstrapped procedures.

Thirdly, to my initial co-supervisor, Professor John Dunn (University of Western Australia), who encouraged me to think freely and creatively about how mathematical models might be applied to eyewitness identification tasks. He ensured that I developed the necessary mathematical skills required to work within the field and enthusiastically supported me as I developed a critical test for high-threshold representation.

Finally, to my daughter, Lauren McCormick, for helping me to maintain an appropriate balance of voice and grammar, Lauren has directly assisted me in making significant improvements to my writing skills over the last five years, and has tirelessly proofread my work throughout my candidature.

Table of Contents

List of Figures

List of Tables

Section I: Models of eyewitness identification decisions

Psychological researchers have put forward several theories of eyewitness identification behavior, each of which provide differing hypotheses relating to eyewitness memory and identification task difficulty. These theories inform both the conceptual and analytical frameworks applied within the field, combining into an overarching methodology of study. While certain verbal models have remained relatively unchallenged within the field, newer theories have more recently been introduced. As a result, the field has experienced a significant shift in theoretical perspectives, bringing with it a certain level of controversy as researchers grounded in various schools of thought argue over their methodological differences. However, such competition between competing models and their associated analytical frameworks also provides an opportunity to strengthen eyewitness identification theory via the scientific process of falsification.

This thesis seeks to achieve greater clarity in relation to the validity of the concepts underlying new mathematical models of eyewitness identification. The following chapters within Section I briefly describe the competing informal (verbal) and formal (computational and mathematical) models of eyewitness decision making and the methods which might be applied to their selection. Section II presents a paper submitted to the journal *Law and Human Behavior.* This paper addresses three themes. Firstly, the appropriateness of applying mathematical models to eyewitness identification decisions at the population level.[1] Secondly, whether such a model of eyewitness identification should

[1] Experimental eyewitness identification data is almost always collected between subjects, such that each participant provides a single data point. Thus, empirical eyewitness identification data is collected at the population level, rather than at the individual level.

be represented by one grounded in threshold theory, or one grounded in signal detection theory (SDT). Thirdly, whether the surviving model provides a good representation of both the empirical outcomes, and the psychological constructs which it is attempting to formalize. Finally, Section III gives commentary to the paper, discussing the potential benefits provided by the surviving set of models, the issues raised in applying them, and the potential future focus within the field.

Informal models of eyewitness identification

Built using purely verbal descriptions, the use of analogy within informal models of eyewitness identification provides a sense of understanding in ways that are both intuitive and accessible. Importantly, because these models are easily understood by non-scientists, their dissemination across multiple communication channels has been highly successful, and have, until relatively recently, remained unopposed. Each of the following models have advanced the field of eyewitness identification through the provision of a methodological grounding, which has, in turn, informed the conceptual and the analytical frameworks of eyewitness identification research.

Relative and absolute judgements

The model of *relative and absolute judgements* (Wells, 1984) is described as the combination of two decision processes positioned along a single decision continuum. A *relative-judgement process* is one in which eyewitnesses simply identify "the lineup member who most resembles the witnesses' memory relative to the other lineup members" (p. 92). Identifications made using this process result in an increased rate of correct identifications from target-present lineups (i.e., lineups which include a guilty suspect), simply because the guilty suspect will be the lineup member who most resembles the eyewitness's memory of the perpetrator. However, if the suspect is

innocent, the relative-judgement process would also lead to a high rate of false identifications, occurring a probability of $1/k$, where k is the size of a 'fair' lineup[2]. In contrast, the *absolute judgements process* results in the identification of the lineup member if, and only if, the level of match between a lineup member and the eyewitness's recollection of the perpetrator exceeds some critical value. Thus, assuming that this critical value is set conservatively, eyewitnesses who utilize an absolute judgement process will be less likely to falsely identify an innocent suspect than those eyewitnesses who employ a relative judgement process.[3]

Relative and absolute judgement decision processes are also described as being non-dichotomous. That is, they lie along a continuum such that the eyewitness will give greater weight to one process over the other. For example, certain eyewitnesses may favor a set of judgement processes which are more heavily weighted by relative judgements, while others may give greater weight to absolute judgments. Wells suggested that tendencies towards the direction of preferential weighting is primarily influenced by the quality of the eyewitness' memorial information, with the tendency towards utilizing relative judgments when memorial information is weaker and vice versa. Thus, eyewitnesses who hold weak memorial information of a perpetrator are more likely to make a false identification than reject a lineup purely because they are biased towards always choosing the lineup member whom most resembles the perpetrator. Conversely,

[2] A lineup is considered to be fair if the suspect and fillers all equally match an eyewitness's memory of the seen culprit. Unfair lineups (sometimes referred to as biased lineups) occur when an innocent suspect either more closely resembles the description of the culprit, or is presented in a way that indicates a higher probability of guilt in comparison to other lineup members.

[3] The development of the sequential lineup presentation is grounded in the theory of relative and absolute judgements. Lindsay and Wells (1985) hypothesized that sequential presentations limit the eyewitness's ability to make relative judgements and instead force them to use the absolute judgement process only. As a result, the diagnosticity of the test (as measured by the ratio of correct to false identifications) increases. See Section II and Section III for further discussion.

eyewitnesses who hold strong memorial information of a previously seen perpetrator are more likely to reject a lineup containing an innocent suspect, as they are biased toward not choosing someone unless there is a match between the 'most similar' lineup member and their recollection of the perpetrator.

Filler siphoning

Under the *filler siphoning* model (Wells, Smalarz, & Smith, 2015)—also known as *filler shift* (Wells, 1993)—identification errors are spread equally across the lineup, assuming the lineup is fair. As a result, the risk of an innocent suspect being identified is equivalent to the rate of selecting the suspect purely by chance. Since a guilty suspect will always resemble themselves better than any filler or innocent suspect within a lineup, the rate of correct identifications should not be significantly affected by additional of filler (Lindsay & Wells, 1980; Penrod, Garcia, & Robertson, 2005; Wells, 1984). The filler siphoning model is grounded in the assumption that eyewitnesses either detect a guilty suspect and correctly identify them, reject the lineup, or make an identification based on a guess.

Partial memory

The model of *partial memory*, proposed by Wells, Steblay, and Dysart (2012), extends the filler siphoning model by suggesting that correct identifications are a mixture of 'true' identifications and 'educated' guesses. A 'true' identification occurs when the eyewitness accuratly detects and identifies the culprit within the lineup task. When there is a failure to detect the culprit, the eyewitness may still choose to identify someone from the lineup with the probability of suspect selection maximised at $1/k$, where k is the lineup size (Penrod, 2006). However, rather than guessing at random, eyewitnesses may use their partial memory of the culprit to "eliminate some lineup members and then choose

from among the remainder" (Wells, Steblay, & Dysart, 2012, p. 268). For example, because the fillers within a lineup have different levels of similarity to the previously seen culprit, one or more of them may be detected as innocent persons and be removed from contention as a possible target for identification (i.e., a guilty suspect). Then, if the eyewitness does not detect the actual target, either because the suspect is innocent or the eyewitness's memory is poor, then they might choose to guess from the remaining lineup contenders. The stronger the memory of the eyewitness, the greater the number of innocent lineup members (i.e., both the fillers and any innocent suspect) will be eliminated and the greater the overall accuracy of their decision, be it to reject the lineup or identify someone from it.

Thus these 'educated' guesses may have a probability of guilty suspect selection between $1/k$ and 1. Under this model's structure, researchers consider it theoretically possible to achieve perfect levels of eyewitness identification reliablity through the elimination of guessing behaviour. This model also emphasises the importance of presenting a fair lineup. That is, a lineup in which the suspect does not stand out in some way from the lineup fillers, and in which all lineup members equally match the eyewitness's verbal description of the culprit. This is because, when presented with a fair lineup, the eyewitness is equally likely to eliminate the suspect as they are to eliminate any filler lineup member.

Limitations of verbal models

Although verbal models of eyewitness identification decision behavior are constructed from observed truths, and despite their general acceptance, their inherent descriptive imprecision creates ambiguity between their postulates and the hypotheses that have been drawn from them (Smaldino & McElreath, 2016). Unfortunately, without

formalization, what remains is a scientific process devoid of deductive reasoning (Robinaugh, Haslbeck, Ryan, Fried, & Waldorp, 2020) and a set of models which cannot be expected to generate precise predictions (Fried, 2020). Indeed, without formalization, such verbal models cannot be falsified, are easily interpreted in ways that 'explain away' empirical findings and leave the field of eyewitness identification research open to HARKing (i.e., hypothesizing after results known; Kerr, 1998).

Formal models of eyewitness identification

The transformation of eyewitness identification theory-sets from containing primarily verbal to primarily formal models is inherent to the building of a strong theory, as by doing so the underlying assumptions become more salient, predictions more precise, and the testing-risk of hypotheses more profound (Bjork, 1973; Clark, 2008; Hintzman, 1991), by becoming more susceptible to falsification, and yet surviving such tests, a model of a theory may be considered strong and therefore regarded as the best available conceptualization of truth.

Mathematical models provide a multitude of benefits to the field within which they are applied. For example, they become a tool for measurement and evaluation. Through their precise predictions of outcomes, it is possible to estimate the parameters of a given phenomenon, and under the assumption of latent variable independence, changes to these parameters under experimental manipulation conditions may be evaluated under a known level of confidence. Mathematical models also provide a platform for theory development, a common and precise language with which to facilitate expanded academic collaboration and integration, and most importantly, the evaluation of explanations made during the inductive and deductive reasoning processes.

Attempts to formalize verbal models of eyewitness identification began nearly twenty years ago with the introduction of the computational model, WITNESS, by Clark (2003). The development and introduction of mathematical models of eyewitness identification decision processes and tasks have more recently appeared in the literature (for examples, see Amendola & Wixted, 2015b; Dunn, Kaesler, & Semmler, 2022; Kaesler, Dunn, Ransom, & Semmler, 2020; Wixted & Mickes, 2014). Grounded in signal detection theory (SDT) models, a successful model-set appropriated from other psychological fields of study, these models not only provide highly precise predictions of eyewitness identification tasks, but they also parameterize task-relevant psychological constructs, such as decision bias and discriminability[4], without attempting to explain the underlying psycho-physical processes governing memory and perception. However, the focus on SDT models within eyewitness identification overlooks the fact that there exist competing theories of human perception, recognition, and detection—the set of models grounded in high-threshold (HT) theory. These models provide the simplest account of eyewitness identification and are supportive of two of the pre-existing verbal models of eyewitness identification—filler siphoning and partial memory. The following sections provide informal descriptions for both the SDT and HT classes of models. For formal descriptions, please refer to Appendix 1: Formal models of eyewitness memory.

WITNESS model

Attempts to formalize the verbal theories of eyewitness identification behavior began through the computational modelling of absolute and relative judgements (the *WITNESS* model; Clark, 2003) and comparison of this new, formalized model with empirical data. In doing so, new competing hypotheses relating to the weighting of

[4] In the context of an eyewitness identification task, discriminability is defined as the eyewitness's ability to discriminate between a guilty suspect and innocent lineup members.

judgement type and the role of decision bias were discovered, leading to the questioning of the verbal model's assertion of absolute judgement superiority (Clark, Erickson, & Breneman, 2011). This issue was again raised by Fife, Perry, and Gronlund (2014), who found that the model's inclusion of two bias constructs—judgement weighting and decision bias—resulted in unidentifiable model formalizations. A few years later, the WITNESS model was adapted to include the assumptions of filler siphoning (Wetmore, McAdoo, Gronlund, & Neuschatz, 2017). Subsequent data simulations from the modified model not only provided a poor fit to empirical data, but also resulted in contradictory outcome predictions.

Signal detection theory models

The *signal detection theory* (SDT) models (Swets, Tanner, & Birdsall, 1961) of eyewitness memory assume eyewitnesses have access to continuously graded information when making identification decisions. These models assume that each item within a lineup will elicit a subjective value of signal strength, often referred to as a value of 'familiarity', from within the eyewitness's cognitive system through some random process. They also assume that eyewitnesses have direct access to these subjective values, such that, when comparing the true nature of the lineup stimuli, the eyewitness will base their identification decision on the values of familiarity generated by them. As a result, and assuming rationality in decision making holds, the lineup member with the strongest familiarity value will always be judged by the eyewitness as such, regardless of truth.

For eyewitness identification tasks, the SDT models also assume that, where an eyewitness holds memory of a seen perpetrator, the distribution of familiarity values generated by an image of that perpetrator will lie on a 'continuum of familiarity' at some point above the distributions of familiarity generated by previously unseen persons. As a

result, a guilty suspect is more likely to be judged by the eyewitness to be the most familiar lineup member. Furthermore, the model also assumes that, because both innocent suspects and fillers have not been previously seen by the eyewitness, they should be equally unfamiliar. Therefore, their sampling distributions are assumed to perfectly overlap.

Interestingly, although the SDT models allow for familiarity values to vary along a continuum, they do not measure these values using an externally defined scale. Instead, their parameters are always expressed in the units of the non-target sampling distribution, which is typically transformed to the standard normal Gaussian distribution to maintain identifiability. As a result, all other parameters, including those representing discriminability and choice bias, are expressed in the standard deviation units of the non-target (e.g., filler) distribution.

When an eyewitness performs a showup or lineup identification task, they may either identify someone or reject the lineup altogether. Thus, the SDT models also include a *decision criterion* at some specified point along the continuum of familiarity (Tanner & Swets, 1954). This criterion is set by the eyewitness in a way that increases the likelihood that their decision goal is achieved. For example, the *ideal observer*[5] eyewitness may want to minimize overall error. To achieve this, they will select a decision criterion that minimizes the risk of both the false identification of an innocent suspect and the missed identification of a guilty one. However, it is more often that the eyewitness's goal is to favor the minimization of false identification risk, and so the will set their decision criterion in a position much higher on the continuum. In doing so, these eyewitnesses are

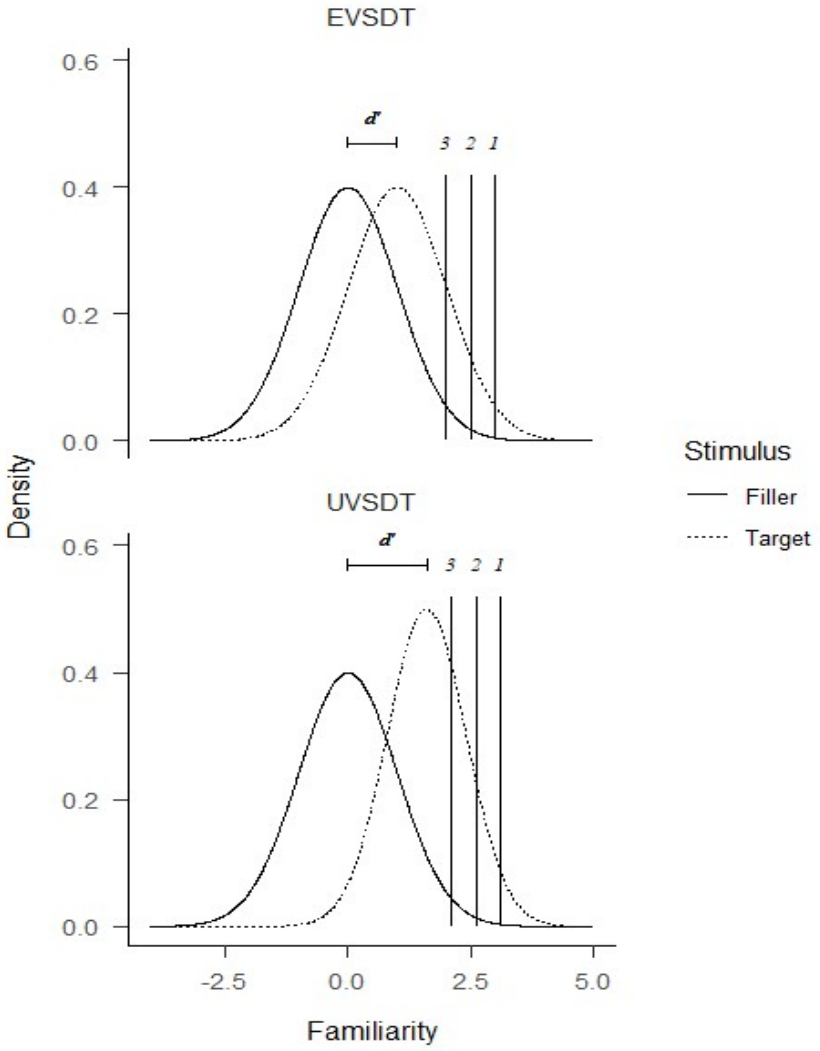
[5] Note that the term 'ideal observer' refers only to a decision maker who achieves the minimization of overall decision error—false identifications and missed identifications. It does not refer to an eyewitness who makes 'ideal identification decisions', since such judgements are relative to the decision goals set by the eyewitness, the investigating officers, and the courts.

classified as *conservative*. In contrast, a *liberal* eyewitness will set a lower criterion so that they may maximize the probability of making a correct identification. Unfortunately, this is achieved at the cost of increasing the risk of falsely identifying an innocent suspect.

Under these models, identification errors occur due to internal and external noise within the sensory system, with some level of error being considered unavoidable, and guessing only occurring where there is a near complete lack of discriminability. Thus SDT models do not assume that false alarm errors always emanate from guessing behavior. Instead, identification decision error occurs solely because sampling distributions of stimuli overlap, creating ambiguity in the interpretation of the sensory information. Indeed, the larger the distributional overlap between stimuli, the lower the decision maker's ability to discriminate between them, and the higher the probability of decision error. Furthermore, if the distribution of values produced by a guilty suspect (the target) does not overlap those of the fillers within the lineup, then eyewitnesses will always be error free in their judgments of target-present lineups.

There are two classes of SDT models, each of which hold different assumptions regarding the distribution of systematic noise. For example, *equal variance* (EV) SDT models assume that systematic noise remains equivalent across all lineup stimuli (i.e., the variance of the target and non-target sampling distributions are equal), while *unequal variance* (UV) SDT models allow these to differ, with this difference being expressed within the target distribution. Figure 1, below, illustrates both EV- and UV-SDT model classes.

Figure 1: *Equal variant and unequal variant signal detection theory (SDT) decision spaces for eyewitness lineup tasks*



Note. The solid curve represents the probability density function of familiarity values for the innocent lineup members (fillers and innocent suspect) and the broken curve represents the probability density function of familiarity values for the target (guilty suspect). Note that the density functions for each innocent lineup member lie on the same curve and are set to the standard normal distribution. The upper figure depicts an equal variant model of a target present lineup in which familiarity values vary equally between the target and fillers and the lower figure depicts an unequal variant model of a target present lineup in which the values of familiarity for the target are shown to vary less than

values of familiarity of the fillers. The distance between distribution means provides the model's *d′* parameter. The vertical lines indicate decision criteria at three levels of bias. The criterion labelled one (1) is the most conservative, lying to the right, and the criteria labelled two (2) and three (3) are more liberal, lying further to the left.

High threshold models

The *high threshold* models (see Bayen, Murnane, & Erdfelder, 1996) are a set of multinomial mathematical models based in *Threshold Theory* (Fechner, 1860/1966). These models assume that stimuli may only be perceived if, and only if, the subjective psychophysical value generated from that target exceeds some fixed limen. Under this assumption, an eyewitness will only have access to a small number of discrete states in which they may be either informed or uninformed by the stimulus. Imagine that an eyewitness has agreed to complete a showup task. They will first generate a sensory value of familiarity for the suspect presented to them, and then based on the strength of subjective familiarity, they will either identify or reject the suspect as being the previously seen perpetrator. The HT models hypothesize that this value will be transformed by some cognitive system into either a *detect* or a *non-detect* state. If the eyewitness enters a detect 'guilty' state when presented with a guilty suspect, then they will *accurately* identify them as the perpetrator. Likewise, if the eyewitness enters a detect 'innocent' state when presented with an innocent suspect, then they will *accurately* reject them.

Under the HT models, identification errors occur when eyewitnesses choose to make an identification even although they have failed to enter a detect state. The occurrence of such guessing behavior will depend on the eyewitness' own goals regarding performance accuracy requirements. For example, if the eyewitness feels that it is especially important to 'catch the perpetrator', then they will be more inclined to identify the suspect, even although the culprit was not detected. This may be because they are either naturally inclined to be liberal in their decision making, believe that the suspect must be guilty because they have been presented to them in an identification task (Behrman & Davey, 2001; Davis, Valentine, Memon, & Roberts, 2015), or simply wish to please the task administrator by behaving in the 'expected way'. However, if the

eyewitness feels that it is more important to not falsely accuse an innocent person, then they will be more likely to reject the suspect.

There are two primary forms of high-threshold models—the single high-threshold model and the double high-threshold model. This chapter briefly presents each of these models in turn, including their assumptions and adaptations to the lineup task. For more formal descriptions, please refer to Appendix 1: Formal models of eyewitness memory.

Single high-threshold model

Blackwell (1953) provided researchers in psychophysics with an early model of detection and recognition—the *single-high threshold* (1HT) model. As discussed above, the model assumes that eyewitnesses will either enter a detect state if the latent-variable of familiarity of the target (a guilty suspect) exceeds the relevant target limen, else they will remain in a non-detect state. Thus, only subjective values of familiarity for a previously seen target may exceed the limen. Such detections will always result in the correct identification of a guilty suspect. Otherwise, eyewitnesses will either identify from the lineup by 'guessing' or simply reject the lineup.

The lineup identification task is more complex as it presents the eyewitness with a lineup containing a single suspect who is either guilty or innocent embedded amongst a set of fillers who are known to be innocent. Assuming that an innocent suspect and each of the fillers are equally as likely to be identified by the eyewitness, the risk of falsely identifying an innocent suspect through a guessing strategy is spread evenly across the lineup. Furthermore, where a suspect is guilty and *the chance of their detection is assumed to be independent from the size of the lineup*, the addition of fillers will decrease the rate of false identifications and 'lucky guesses' without impacting the rate of 'true' identifications. For example, imagine that the eyewitness's memory of the culprit is relatively strong such that the probability the guilty suspect is detected from a six item

lineup at the rate of 66.67%. If we were to increase the lineup size from six to eight via the addition of two fillers, there is a maximal absolute risk reduction of educated guessing of 5.7% and a relative risk reduction in false identifications of 28.5%. The expected effect of adding these fillers to a target absent (innocent suspect) lineup would result in the same absolute reduction in correct identifications, the relative risk reduction of correct identification is much smaller at 6.6% simply because the model assumes independence between lineup size and target detection via a threshold mechanism.[6] In this way, the 1HT model is in direct alignment with the filler siphoning model proposed by Wells, Smalarz, and Smith (2015).

Double high-threshold model

Snodgrass and Corwin (1988) noted several inherent difficulties within the 1HT model and so described a new model – the *double high threshold* (2HT) model – to strengthen its predictive ability. This new model simply extends the 1HT by the addition of a parameter representing the probability that non-targets (either a filler or an innocent suspect) are correctly detected by the eyewitness as being innocent. As such, the model assumes an additional detection threshold and parameter to quantify the observer's discriminable ability. However, adjusting the 2HT model to represent the eyewitness identification task is more complicated than for the 1HT model. Firstly the 2HT model

[6] Note that the absolute risk of chance identification (guessing) is $1/5$ for a six item lineup and $1/7$ for an eight item lineup. The absolute risk reduction is thus $1/5 - 1/7$ and relative risk reduction is $\frac{1/5-1/7}{1/5}$. The absolute risk of true identification (detecting a guilty suspect) is $(2/3 + 1/5) - (2/3 + 1/7)$ for a six item lineup and $2/3 + 1/7$ for an eight item lineup. Absolute risk reduction is $1/5 - 1/7$ and relative risk reduction is $\frac{(2/3+1/5)-(2/3+1/7)}{(2/3+1/5)}$.

allows for the possibility of detecting one or more innocent lineup members as innocent, with an innocent suspect being just as likely to be detected by the eyewitness as any one of the fillers in the lineup. If the detection of one or more innocent lineup member(s) occurs, the lineup is effectively reduced by that number and only undetected lineup members remain available for subsequent identification. The size of this reduced lineup is often referred to as the *effective lineup* size (Malpass, 1981). Thus, the 2HT model is in direct alignment with the partial memory model proposed by Wells, Steblay, and Dysart (2012), as it also suggests a form of 'educated guessing', such that the probability of accurately identifying a guilty suspect may exceed chance.

Limitations of formal models

While there are many benefits to the development and application of formal models in the psychological sciences, a number of limitations remain inherent to the use of formal models within the behavioural sciences. Firstly, human behavior is neither precise nor homogeneous, and is in consequence, difficult to predict on any level (Sen, 1986). Secondly, in attempts to improve the predictive performance of formal models, modelers often resort to the development of model structures which are over-complex, difficult to interpret, and open to issues regarding identifiability[7]. For example, if a formal model fails, theorists may be tempted to save it by modifying one or two constraints or by adding a new parameter. However, such attempts may render the model unidentifiable and/or psychologically implausible (Moran, 2016). Yet without a good understanding of the underlying structure of the model, or the ability to assess it, researchers within the field may be forced to 'take the theorist at their word' and must accept a theory which instead ought to be rejected. Thus, not only do formal models require in-depth

[7] A model becomes unidentifiable when more than one set of parameters are able to predict identical outcomes.

understanding of computational and mathematical concepts, making them difficult to communicate to persons without such understandings, their inappropriate development, interpretation, and application without interrogation can lead to the false conclusions regarding empirical outcomes and further misunderstanding of the phenomena.

Thirdly, while a generalized model of behavior may be useful as a tool for measurement, or for informing our understanding of specific behaviors at the population level, without information regarding *individual* performance indicators, population-level models of eyewitness identification cannot be used for the purpose of estimating the probative value of any *single* piece of eyewitness identification evidence.

The scientific endeavor of model selection

To develop a scientific understanding of eyewitness identification also requires an understanding of *what science is*. Interestingly, knowing what science is, and conversely what it is not, is not necessarily straight-forward matter. Indeed, great thinkers within the philosophy of science, such as Plato, Aristotle, Popper, Bacon, et cetera. have all found diverse ways to consider science, both as a philosophy and as an activity in the pursuit of 'truth'. However, the overarching feeling is that purpose of science to advance our understanding of phenomena such that we might see the world not only for what it is, but more importantly, for what it is not. To achieve this, scientific enquiry should utilize reasoning processes that develop hypotheses that may be empirically tested, that is, hypotheses that are falsifiable. Unfortunately, untestable hypotheses remain at the core of eyewitness memory theory, which exists as a set of informal (verbal) and formal (computational and mathematical) models. Where there exist two competing models within the theory-set, the failure of one model will bring about support for the other, at least until a more viable model is discovered.

The best practice in hypothesis testing should involve the development of a program of empirical studies based on the full range of model predictions within a theory set (Lakatos, 1980). Unfortunately, some models do not provide predictions that can be empirically tested. In such cases it should at least be possible to challenge them through the testing of their axillary hypotheses and predictions, such as predictions made under different task structures. This chapter describes various avenues for the testing and selection of models within the field of eyewitness identification. These include the application of critical tests (falsification), testing the fit of a model's predictions to empirical data, testing a model's parameters through constraint, and considerations of psychological plausibility.

Critical tests

The falsification of a hypothesis or model requires the development and application of a critical test: the identification of a dichotomous prediction, which if not met under empirical conditions, cause the model to fail and thus be rejected as a representation of the phenomenon. Interestingly, models grounded in different sets of axioms may make remarkably similar, or even identical predictions and are thus considered to 'mimic' each other (Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). Such cases require more abstract methods of enquiry, such as conducting one or more thought experiments upon a formal hypothesis and then generating, through deductive processes, a novel set of predicted outcomes. Recent examples of the development of such critical tests include Kellen and Klauer (2014) critical test of discrete-state models within recognition memory experiments, and the battery of critical tests of random-scale[8]

[8] In which the momentarily perceived familiarity (value) of a lineup member (stimulus) varies due to random 'error'. As a result, the values of each member of a lineup is thus sampled from independent random distribution in that moment, such that their values may be ranked and never be tied (see Falmagne, 1978).

representation, latent variable independence, likelihood ratio monotonicity, and threshold representation, again in recognition memory, by Kellen, Winiger, Dunn, and Singmann (2021). In these cases, the theorists expanded the types of outcome data for assessment by extrapolating the models to different tasks, such as the 4-alternative forced-choice (4-AFC) task and a two stepped test which required participants to first perform a 5-alternate sub-setting task and then a 2-AFC reverse decision task.

Testing by fit

Where two or more models survive critical testing, or where no critical test can be found to separate them, theorists may choose to compare the predictive performance of the models by testing their fit with empirical data. After estimating the parameters of a model through maximum likelihood estimation (MLE) methods, various fitting statistics may then be used for the purpose of model selection. These include simple $\chi^2$and G^2 fitting statistics, or the more complex Akaike information criterion (AIC) and Bayesian information criterion (BIC), which account for such differences by giving weight to models containing fewer parameters. Where the models being compared contain the same number of parameters, simple fitting statistics may be applied. However, if the models are nested, or if they have differing levels of complexity (i.e., different number or parameters), then an evaluation using AIC and BIC statistics is a more appropriate method of model selection.

Testing by constraint

Models may also be tested by constraining their parameters in the face of the experimental manipulation of those parameters. For example, experimental manipulations of both discriminability (through memory strength manipulations) and decision criterion placement (through expected base-rate manipulations) might be used to answer the

following questions: “Are the parameters of a model linked to the psychological constructs that they explain?” and “do they behave in ways in which are predicted?” Under the constraint of its discriminability parameter, a model which provides a good representation of discrimination would fit empirical data collected from two different levels of base-rate expectation but would provide a poor fit to empirical data collected from two different levels of memory strength. Similarly, a model which provides a good representation of decision bias would fit, under the constraint of its decision criterion parameters, empirical data collected from two different levels of memory strength. However, it would not fit empirical data collected from two different levels of base-rate expectation.

Psychological plausibility

The psychological plausibility of a model requires judgements about both the assumptions underlying it, its systemic structure, and the nature of the outcomes predicted by it. For example, is it psychologically plausible for there to exist some subjective latent variable of familiarity generated within the cognitive system of an eyewitness when they are exposed to a lineup stimulus? Likewise, is it plausible for an eyewitness to recognize an innocent lineup member as being innocent, and thus discard them from their pool of possible choices? Such questions of plausibility are vital, as they affect the way in which a model integrates with both the conceptual and analytical frameworks applied to the phenomena which it aims to explain. Indeed, without some level of plausibility, such integration becomes untenable and the model obscure. Those models presented within this section are all, in one way or another, already accepted as being psychologically plausible representations of eyewitness decision behavior.

Section II: The program of research

This section presents a paper authored by Kym M. McCormick (the author of this thesis) and Carolyn Semmler—"Letting go of the Grail: Falsifying the theory of "true" eyewitness identifications"—which was submitted to the *Journal of Law and Human Behavior* in April 2022.

Authors' note

This work was supported by the Australian Research Council grant DP160101048. All experiments were conducted in accordance with the Australian Government's National Statement on Ethical Conduct in Human Research and approved by the University of Adelaide Human Research Ethics Subcommittee: School of Psychology (Ethics Approval Nos. 19/25, 18/74 and 19/01). Each experiment was also preregistered; see https://osf.io/b4pm6/ (Experiment 1), https://osf.io/vfwq5 (Experiment 2: Stage 1), and https://osf.io/dex8w (Experiment 2: Stage 2). Data, analytical code, and supplemental materials are available at https://osf.io/d9q28 . The authors declare that they have no conflicts of interest.

Statement of Authorship

Title of Paper	Letting go of the Grail: Falsifying the theory of "true" eyewitness identifications
Publication Status	Submitted for Publication
Publication Details	Submitted to the *Journal of Law and Human Behavior.* Awaiting review

Principal Author

Name of Principal Author (Candidate)	Kym Michelle McCormick		
Contribution to the Paper	Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization.		
Overall percentage (%)	85		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	13/7/2022

Co-Author Contribution

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above).
ii. permission is granted for the candidate in include the publication in the thesis; and
iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Carolyn Semmler		
Contribution to the Paper	Conceptualization, Methodology, Resources, Writing- Review & Editing, Supervision, Project administration, Funding acquisition.		
Signature		Date	13/7/2022

Letting go of the Grail: Falsifying the theory of "true" eyewitness identifications

Early discussions within both the judiciary and academia (see Sporer, 2008), coupled with later evidence collected by psychological researchers (for brief reviews see Police Executive Research Forum, 2013; Yates, 2017), have led to the gradual introduction of new eyewitness identification procedures and processes which aim to improve the probative value of eyewitness identification evidence (Devlin Committee, 1976; Higgs, 2011; Newirth, 2016). At the forefront of these improvements has been the development and implementation of novel identification tasks, including the introduction of the lineup task (British Home Department, 1929), the inclusion of unbiased identification instructions (Malpass & Devine, 1981), and the development and implementation of the sequential lineup (Lindsay & Wells, 1985), each of which aim to increase the diagnosticity of eyewitness identification evidence[9] by mitigating false identification risk. However, the introduction of such measures compels us to ask the question "at what cost?" (Clark, 2012); a question that seems to remain open for debate within the literature.

The earliest development in eyewitness identification processes involved the embedding of a suspect amongst five or more innocent persons, or fillers, commonly referred to as a lineup identification task. By increasing the eyewitness's choice-set, the lineup task introduces the possibility that an eyewitness will make a known, rather than an unknown, identification error[10] and thus be eliminated as reliable witness. The lineup's

[9] Diagnosticity is described as the likelihood that the suspect is guilty, given that they were identified by the eyewitness.

[10] Because we do not know the true nature of the suspect's guilt or innocence, their identification may be either correct or incorrect, and thus its accuracy is also unknown. However, since fillers are always known to be innocent, their identification is also known to be in error.

structure effectively reduces the risk of false identification by spreading such errors across the fillers within the lineup. This provides a level of protection to an innocent suspect. However, what if the suspect is instead guilty? By adding fillers to the choice-set, does this not simply increase the risk of erroneous filler selection and thus reduce the chance of making a correct identification? According to the theory of *filler siphoning* (Wells, 1993, 2001), because the processes governing target detection are independent to those governing guessing behavior, the presence of fillers does not significantly hinder the eyewitness's ability to detect a guilty suspect amongst them. Thus, filler siphoning provides an explanation for the superiority of lineups over showups (Yarmey & Yarmey, 1996), and by inference predicts a positive relationship between lineup size and probative value. In other words, the addition of a filler to a lineup will result in an increase in diagnosticity without incurring a corresponding increase to the rate of the missed identifications. However, this is not necessarily a view shared by all researchers (for example see Clark, 2012; Yang, Smalarz, Moody, Cabell, & Copp, 2019).

The second major modification to eyewitness identification procedures was the introduction of the unbiased lineup instruction. By providing additional information to eyewitnesses that "the lineup may or may not include the culprit", eyewitnesses are provided with the explicit option to reject a lineup. Under the theory of absolute and relative judgements (Lindsay & Wells, 1985; Wells, 1984), unbiased lineups allow eyewitnesses to shift away from making a high-risk *relative judgement*, in which the most familiar lineup member is identified, and towards the more accurate *absolute judgement*, in which the most familiar lineup member is selected if, and only if, they adequately match the eyewitness's memory of the culprit. As a result, the risk that an eyewitness will automatically identify an innocent suspect, simply because they are the lineup member who looks most like the culprit, is mitigated. However, while unbiased instructions

increase the diagnosticity of suspect identifications, the absolute judgement model predicts they will also come at a cost of missed identifications (Yang et al., 2019). Tempering this account, Wells, Steblay, and Dysart's (2012) theory of *partial memory* suggests instead that such costs weigh more heavily on accurate guesses than on "true" identifications, which are made only by eyewitness who accurately detect a guilty suspect. If proven, this dichotomous view of hit types—detections and random guesses—shifts the focus of performance improvement towards the preservation of good identifications (detections), and elimination of bad identifications (random guesses), thus making the issue of "cost" a moot point.

A third major modification occurred more recently with a considerable number of jurisdictions in the United States and Canada introducing sequential lineups in the hope that this new format would better eliminate non-credible eyewitnesses and thus improve the probative value of identification evidence brought before courts (Lindsay, 1999). However, since the widespread adoption of sequential lineups, there has been no true consensus regarding its superiority to traditional simultaneous presentations. For example, while there is some evidence that simultaneous lineups produce identifications with greater diagnosticity ratios (Lindsay & Wells, 1985; Steblay, Dysart, & Wells, 2011), other forms of analysis, such as ROC curve analysis (Mickes, 2015) and signal detection theory (SDT) modelling suggest they either reduce the eyewitness's overall ability to make a correct identification (Amendola & Wixted, 2015a; Wixted, Gronlund, & Mickes, 2014) or provide no benefit at all to overall decision accuracy (Kaesler, Dunn, Ransom, & Semmler, 2020). So, while the simultaneous presentation might increase the probability that an identified suspect is guilty, it may be doing so at the cost of incorrectly identifying an innocent one.

Despite these significant efforts towards maximizing the probative value of eyewitness identifications without enduring the cost of increased missed identifications, little research effort has been given towards gaining a better understanding of the processes that underpin eyewitness identification decisions. However, this does not mean that formalized models of eyewitness memory cannot be developed and applied successfully. Nor does it mean we must discard existing verbal models and theories. In fact, these may be utilized in the consideration of the development of formal model structures (e.g., Clark, 2003). For example, relative judgement theory suggests that eyewitnesses have access to continuously graded information, reflecting a *signal detection theory* (SDT) representation of eyewitness memory on which all current mathematical models of eyewitness memory are grounded. However, in contradiction to this proposition of continuous information, filler siphoning and partial memory theories each suggest that eyewitnesses only have access to a small number of discrete states of knowledge, i.e., accurate detection of guilty and innocent lineup members, and non-detection. They thus reflect instead a *high-threshold* (HT) representation of eyewitness memory. As a result of this contradiction, eyewitness researchers remain at a crossroads regarding our understanding of the nature of both identification accuracy and identification errors. Do they emanate from the employment of some continuously graded stochastic process[11] whereby the elimination of error becomes impossible? Or are they produced via a discrete-state process under which identification accuracy could be virtually guaranteed, if only eyewitnesses did not insist on identifying suspects by guessing?

[11] In this case, a stochastic (random) process is a mathematical model which assumes the presence of random error through the inclusion of one, or more, randomly distributed variables.

Such questions may only be satisfactorily answered by exploring the underlying structure of eyewitness memory and decision making. We therefore seek to identify the most appropriate theoretical and analytical frameworks required to support such explorations. This requires a dedicated scientific program of study that seeks to identify and describe potential models of eyewitness memory, critical test their predictions, and validate their suitability in representing the psychological constructs underlying eyewitness memory and identification decision making, namely the eyewitness's ability to discriminate between guilty and innocent lineup members—*discriminability*—and their propensity to identify someone from the lineup, or not—*decision bias*.

The Current Study

Our study has three distinct aims. Firstly, we aim to provide support for the mathematical modeling of eyewitness memory by critically testing the assumptions that underlie them—stochastic process and monotonic likelihood representation. By surviving these critical tests, we may gain confidence in our assessment of the appropriateness of mathematical models of eyewitness identification decision making. Secondly, we aim to critically test the two competing classes of mathematical models—HT and SDT models—such that only one model class might survive and thus succeeds the other. In achieving this, the surviving model will inform our answer to questions regarding the existence of "true" identifications. For example, if the HT model class is falsified, then we must concede that there are no "true" identifications and that a reduction in false identification rates cannot be without cost of increased missed identifications. Whereas, if the SDT model class is falsified, then support is provided for either the elimination of guessing behavior amongst eyewitnesses, or the development of tests which might more accurately diagnose a "true" identification. Finally, we aim to verify the surviving model class by

testing its ability to predict empirical data, by assessing the linkage between these models' parameters and key psychological constructs (e.g., discriminability and decision bias), as well their generalizability across various task structures.

Critical tests of stochastic process and monotonic likelihood representation.

A full exploration of critical tests of stochastic processes was initiated by (Block & Marschak, 1959), who demonstrated that forced-choice probabilities based on a stochastic process representation must satisfy a specific set of inequalities (see Appendix 3:A critical test for random scale representation hypothesis). However, it was not until a proof of the converse—that choice probabilities which satisfy the *Block-Marschak inequalities* (BMIs) must have a random utility (i.e., stochastic) representation (Falmagne, 1978)—that this set of choice probabilities could be used as a test of random-scale representation. Kellen, Winiger, Dunn, & Singmann (2021) further transitioned to the context of a *k*-alternative forced choice (*k*-AFC) task by using the resulting system of BMIs to critically test recognition memory for words. By aggregating choice data from 110 participants, which produced approximately 1,000 observations for each of seven *k*-AFC condition groups, where $k \in \{2,\ldots,8\}$ these researchers found that the resulting data provided near-perfect fits to data predicted by the BMI constraints. Furthermore, the data continued to conform to the BMIs constrained such that only monotonically decreasing likelihood solutions were included. For a detailed explanation of the BMI test, please refer to Appendix 3: A critical test for the monotonic likelihood hypothesis.

Critical test of high-threshold representation.

A critical test of discrete state representation was initially developed by Kellen and Klauer (2014), who showed that certain classes of discrete state models—*high-*

threshold (HT) models—may be distinguished from all other models in terms of first- and second -conditional choice probabilities[12], c_1 and c_2, respectively, such that HT models predict constant c_2, independent of changing c_1. This is in direct contrast to the predictions of non-HT models, such as the *signal detection theory* (SDT) model, which predict positive relationships between c_1 and c_2. Kellen and Klauer implemented their test using a 4-alternative ranking task for memory of a set of words. Discriminability was varied by manipulating memory strength for target words by presenting them for study either once (weak) or three times (strong). Consistent with the prediction of continuous models, they found that c_2 was correlated with c_1 in that they were both larger for participants in the strong condition than for those in the weak condition. Similar results were also found by McAdoo and Gronlund (2016), who extended the experiment to the recognition of faces.

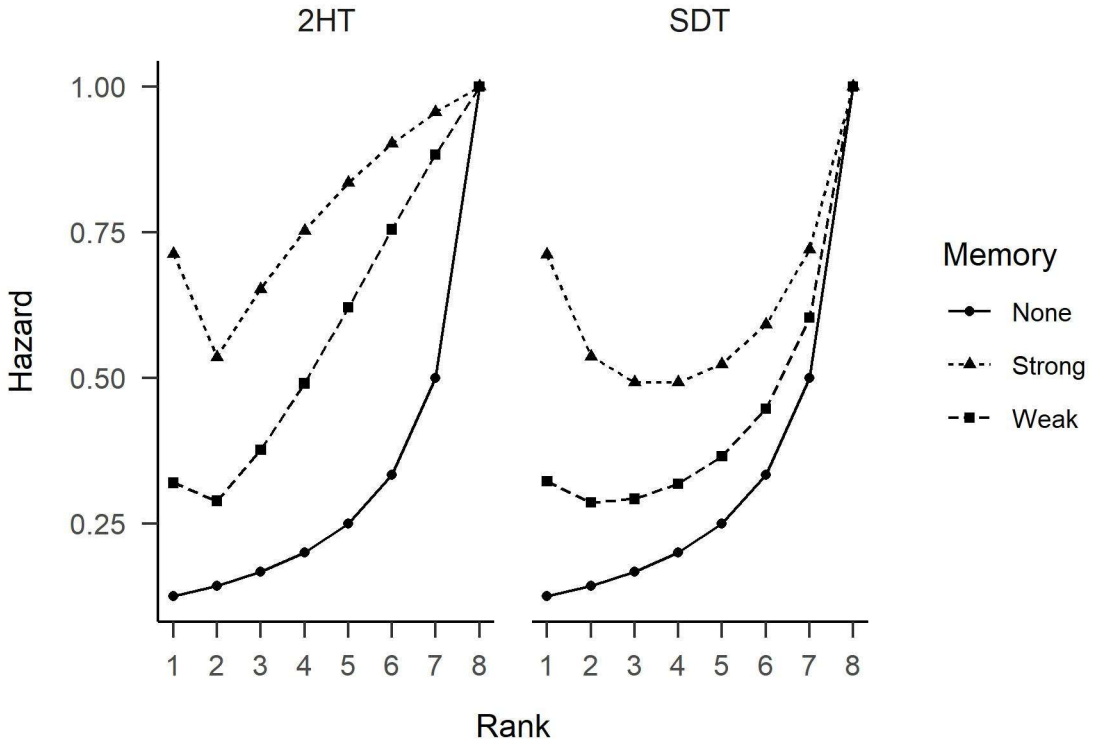
However, a shortcoming of this proposed test is the inherent assumption that memory strength only affects the probability of target detection. Yet, it is conceivable that when strengthening the representation of a target in memory, this may conjointly increase the chance that one or more fillers are also detected, thus increasing the probability of the rejection of target-absent lineups—a theoretical process called differentiation (Criss, 2006) that aligns well with the theory of partial memory. For example, partial memory of the culprit, such as face shape, could lead to the elimination of all lineup members with non-matching face shapes. In this case, the probability of detection for both the target *and* each filler is affected. As such, it is desirable to develop an alternative test that relaxes the c_2 test's assumption of fixed rates of filler detection. In doing so, it is important to move beyond first and second choices and instead consider the full range of conditional ranking

[12] Second choices conditional on incorrect first choices (see Kellen & Klauer, 2014).

probabilities, that is, $c_1,\ldots c_n$, also known as *hazard* distributions. The hazards predicted by HT and SDT models of eyewitness identification at three different memory strengths (no memory, weak memory, and strong memory) are illustrated in

Figure **2**, below. The results show that the two model-classes predict opposing patterns of hazard rates as discriminability increases. For example, the hazard function predicted by the HT model class becomes more concave for $x>1$ as memory increases, while that predicted by the SDT model becomes more convex. For formal descriptions of these predictions, please refer to the Appendix 3: A critical test for the monotonic likelihood hypothesis.

Figure 2: *Predicted Hazards for 8-item Target-Present Lineup Ranking Task*



Note. The figure on the left shows the hazard function predicted by the double high-threshold (2HT) model. Each of these hazard functions are monotonically increasing for all ranks >1. The figure on the right shows the hazard function predicted by the equal variant signal detection theory (EV-SDT) model. In contrast to the 2HT predictions, the SDT model only predicts monotonically increasing hazard rates for all ranks >1 for small rank 1 hazard values (e.g., in the case where the eyewitness has no memory of the perpetrator). However, as the rank 1 value increases, the SDT's predicted hazard rates for ranks >1 become non-monotonic.

Verification of the surviving model

The nature of theories of eyewitness memory requires that any formal model of eyewitness identification decision making includes parameters that independently represent eyewitness discriminability (i.e., the ability of the eyewitness to discriminate between their memory of a previously seen perpetrator, guilty suspects, and innocent lineup members) and decision bias (towards either identifying someone or rejecting the lineup). Thus, models of eyewitness identification that survive the critical testing of their assumptions and predictions should be further assessed to ensure both the predictive strength of the model and the psychological plausibility of their parameters. These tests require fitting the model as well as a set of constrained versions of that model (each of which constrain one or more parameters) against empirical eyewitness identification outcomes using maximum likelihood estimation (MLE) methods. Finally, the independence of the model's discriminability parameter, which is inherently linked to the latent variable of familiarity, will be assessed by comparing parameter values across two different task structures—ranking and identification—while keeping all other variables constant (e.g., encoding conditions, stimuli, lineup presentation, and prior expectation). If the discriminability parameter is shown to be indifferent to task structure, we may conclude that the latent variable is indeed independent and the model both generalizable and robust.

Method

The design of this study required the collection of eyewitness identification data from three distinct tasks—the *k*-AFC task, the simultaneous lineup task, and the ranking task. To maximize the number of observations collected, a program of two separate experiments was designed such that data for each required task was collected only once.

The first experiment primarily collected the k-AFC probabilities from a series lineup sizes $k \in \{2,\ldots,7\}$. The second experiment collected lineup identification data, confidence ratings, justification statements (required for another study), and ranking probabilities from an 8-item simultaneous lineup. This experiment manipulated target presence and attempted to manipulate both discriminability and decision bias (via memory strength and target-present base-rate information manipulations, respectively).

Transparency and openness

I describe the sampling plan, all data exclusions (if any), all manipulations, and all measures in the study. All analysis was conducted using R (R Core Team, 2016) using the following packages: readr (Wickham, Hester, & Bryan, 2022); psych (Revelle, 2022); tidyverse (Wickham et al., 2019); quadprogpp (Noorian, 2015); and MPTinR (Singmann & Kellen, 2013). The experiments within this study were preregistered separately: see https://osf.io/b4pm6/?view_only=972c68947a994b989e8627746b4e4ed4 (Experiment 1), https://osf.io/dex8w/?view_only=38edb38aa0f14f06b833ad9d65d968cc (Experiment 2: Stage 1), and https://osf.io/d9q28/?view_only=7461604216b04bb78f079baabfdcd284 (Experiment 2: Stage 2). Materials, raw data, and analytical code for this study can be found at *Computational models of eyewitness identification decisions* [online repository]; see https://osf.io/d9q28/?view_only=7461604216b04bb78f079baabfdcd284 .

Participants

The participants in this study were recruited online through Amazon Mechanical Turk (www.mturk.com) and paid $1.70 on completion of the assigned experimental task(s). Eligibility requirements for inclusion to the experiment were age ≥18 years, normal vision or corrected to normal vision, and English-language proficiency.

Participants self-selected their participation in one of three experiments, although they were not aware of the structure of these experiments. No participant was allowed to enter or complete more than one experiment. On entering the experimental task, each participant was randomly assigned to a manipulation group. They were then screened for English language proficiency and responded to a series of demographic questions, i.e., age, gender identity, and geographic location. While participation was not geographically restricted, almost all participants were found to be residing within the United States (98.87%). Experiment 1 included *N*=2,026 participants (Male = 1,032) aged 18-81 (*M* = 38.56, *SD* = 11.55) and Experiment 2 included *N* = 3,030 participants (Male = 1,420) aged 18-88 (*M* = 37.32, *SD* = 11.56).

Materials

All materials used within this study were sourced from the [masked information]. This database consists of portrait photographs and accompanying short videos of 99 male and 92 female actors, each of whom were students at the University [masked information]. All portraits were photographed from a set distance and under the same lighting, with each actor wearing identical clothing. While different portrait angles are included, e.g., profile and three-quarter views, only those photographs depicting full-face views with neutral expressions were included in the item selection process. A pool of nine photographs of female actors were selected from the database based on their shared distinguishing features, including gender, hair length and color, age, racial background, and eyewear. Of these, one was selected specifically as the *guilty suspect*, or *target*. These images are available within the supplementary materials.

Each video within the database (duration approximately 10*s*) was recorded using a simulated CCTV camera mounted on the ceiling of an office. The video begins by

showing, from behind, an actor working at a computer. The actor then rises, turns, and walks towards the camera. While passing under the camera, they briefly look up towards the camera and their face is shown from an above-right angle for approximately 3 seconds. A single video of the selected target was selected for presentation to participants during the encoding phase of each experiment conducted in this study. A pool of six videos of male actors were also randomly selected from the database. These were presented sequentially in pairs during the distraction phase of each experiment. The selection (without replacement) of paired videos from the pool was also randomized.

General procedure

During the *study phase* of each experiment, participants were asked to watch the encoding video closely and without interruption. Following this, they performed a *distractor task*. The details of this task are described in each experiment, below. Once the distractor task was completed, the participants entered the *testing phase* of the experiment. During this phase, each participant was presented with either a target-present or target-absent lineup. The target-present lineups consisted of the target and $k - 1$ fillers randomly drawn from the pool of eight fillers, while the target-absent lineup simply consisted of all eight fillers. All photographic lineups were presented horizontally across the screen in a single row (1 x k). To control for position effects, the position of each lineup photograph was randomized. Details regarding experimental manipulations and procedures are provided within the each of the experimental procedure descriptions below.

Experiment 1: *k*-Alternative forced-choice task

The aim of Experiment 1 was to collect *k*-AFC probabilities for $k \in \{2,\ldots,7\}$.[13] On entering the experiment, participants were randomly allocated to one of the six *k*-item lineup size manipulation groups, with final group sizes of 335, 353, 336, 328, 338, and 336, respectively. All participants first viewed the target video and then completed a visual search distractor task. Participants in manipulation groups 1, 2, …, 6 were then presented with a 2-, 3-, …, or 7-item target-present lineup, respectively. They then completed a forced-choice task by selecting "the photograph that most resembles" the person in the video.

Experiment 2: Identification and ranking tasks

The aim of Experiment 2 was to collect within-subject data for the 8-item simultaneous lineup identification and ranking tasks. The experiment used a 3 (strong, weak, very weak memory) x 2 (low vs high expectation) x 2 (target-present vs target-absent) between-subject experimental design. The data was collected in two stages. Stage one (Experiment 2a) focused on collecting data for the "strong" and "weak" memory groups. Stage two (Experiment 2b) focused on collecting data for the "very weak" memory group. This data was collected at a later date after preliminary results indicated

[13] Note that *k*-item ranking tasks produce 'first ranked' items which are considered mathematically equivalent to a *k*-AFC choice. Thus, 8-AFC data was proxied by first-ranked data collected during the Experiment 2a ranking task, with the aim to minimize both participant burden and experimental costs.

that the effect of the memory strength manipulation applied in experiment 2a required further strengthening.

The overall base rates for each memory manipulation were evenly distributed (strong = .35; weak = .32; and very weak = .33), as was the base rate for the target expectation (high = .5; low = .5). However, to maximize analytical power for the critical testing, which required target present data only, more samples from target present lineups were collected, with a base rate of .68.

All participants first viewed the target video and then completed either a short or long similarity rating distractor task, depending on whether they were allocated to the strong or weak memory manipulation group, respectively. The short similarity rating task required participants to view two videos, each of which had been randomly selected from the distractor video pool. After watching both videos, participants were then asked to consider the "faces of the two actors you have just seen" and provide a rating on a scale of 0 to 100 indicating "…how similar you think the two faces were." The long distractor task was expected to further weaken the participant's memory strength by requiring participants to complete a series of three short distractor tasks, thus viewing a total of six male faces. Expectation was manipulated by either informing participants that the target has an eight in ten (80%) chance of appearing within the lineups (high expectation), or a three in ten (30%) chance (low expectation).

The test phase required participants to complete a series of tasks, including: (1) a traditional simultaneous identification task, (2) a decision justification description, (3) a post-identification confidence rating, and (4) a ranking task.

Identification task.

All participants were presented with either a target-present, or a target-absent, 1 x 8 item simultaneous lineup. In the target-present lineup, the target was placed among

seven fillers that were drawn, at random, from a pool of eight similar faces. Participants were asked to "click on the picture of the thief to identify them", and if they could not see the thief, to "click on the silhouette image." They were again permitted to change their selection before submitting their final answer.

Justification of choice

To record participants' own thoughts about their identification choice, participants were asked to "briefly describe how you made your decision and what information you used to do so." This data was collected for a future metacognitive study.

Post-identification confidence rating.

Participants then rated their level of confidence in their identification decision using a 0-100 percentage scale.

Forced choice ranking task.

Participants were again shown the lineup presented to them in the identification task and were asked to indicate "which of the lineup members is the person you saw in the original video," and if they couldn't see the person, they selected "the photograph that you think most resembles the person you saw." Once the participant had confirmed their first choice, their selected item disappeared from the lineup. To collect second, third, and subsequent choices from the remaining lineup items, participants were asked to continue to select "the best resembling person from the remaining lineup members." Each selection was removed from the lineup before the participant selected their next best choice. The task proceeded until only one face remained.

Analysis, Results and Discussion

The analytical results and discussion addressing each of the study's three aims—critical testing of stochastic processes and monotonic likelihood representation; critical testing of high-threshold representation; and verification of the surviving model—are presented below.

Random-scale and monotonic-likelihood representation

To test the assumptions of random-scale representation and monotonic likelihood, data collected in from Experiment 1 (k-AFC probabilities, where $k \in \{2,\ldots,7\}$) was combined with Rank 1 probabilities (which are mathematically equivalent to 8-AFC probabilities, see Appendix 2: Signal detection theory model prediction) collected from the strong memory/target-present group in Experiment 2a. The resulting rates of target selections are presented in Table 1. The forced-choice accuracy for all lineup sizes was well above chance with performance decreasing as the lineup size increased, indicating positive memory for the target. The assumption of regularity, in which the rate of correct identification of the target decreases as the size of the lineup increases, also appears to have held. Following Kellen, Winiger, Dunn, and Singmann (2021), we adopted a frequentist model-fitting solution using a semi-parametric bootstrap procedure in which model fits were generated from 10,000 bootstrap samples of the empirical data. Each model fit was constrained by the requirement that its expected choice probabilities fulfilled the BMIs. Model predictions fit the observed data (G^2 = 3.14, p= .544). When monotonic-likelihood constraints were tested together with the BMIs, the fit of the resulting model was maintained (G^2 =3.14, p= .596; see Figure 3).

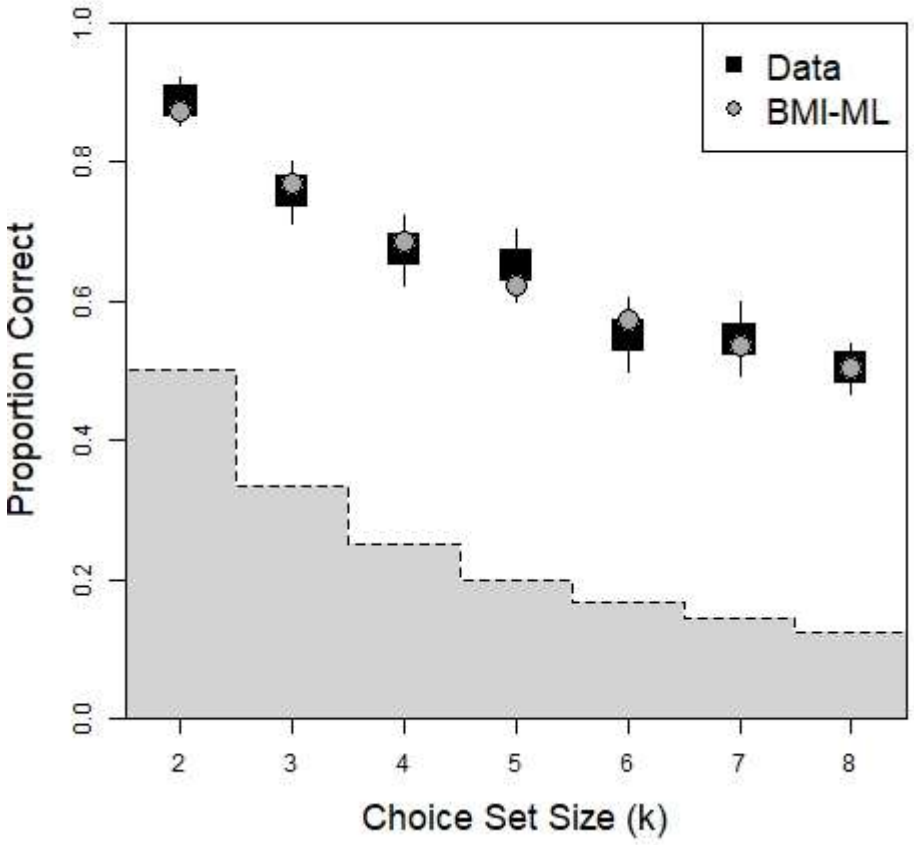
Our results suggest that forced-choice eyewitness memory is consistent with the BMIs and are therefore consistent with a random scale representation. The data is also

consistent under monotonic constraints, which further suggests that the distribution of latent strength values for previously seen targets are, on average, greater than the combined distribution of values for fillers. While it remains possible that eyewitness memory has a non-random scale representation, this is very unlikely. Indeed, increasing the power of the test through the collection of further data is likely to reduce the noise observed in the target identification rates across lineups. Further, considering the effect of the fillers are likely to have on eyewitness decision making, the fact that our empirical data significantly conformed to the BMI and ML constraints allows us to further investigate and develop mathematical models of eyewitness memory with greater confidence.

Table 1: *Counts of Target and Filler Forced Selections from Target-Present Lineups (Experiment 1)*

Lineup size	Target	Filler	Total
2	297	38	335
3	267	86	353
4	226	110	336
5	213	115	328
6	186	152	338
7	183	153	336
8	368	364	732

Figure 3: k-*AFC Performance (Experiment 1)*



Note. Chance, observed, and expected frequencies based on BMI and decreasing monotonic likelihood constraints for correct and incorrect responses for *k*-AFC eyewitness identification tasks, where $k = \in \{2, \ldots,8\}$.

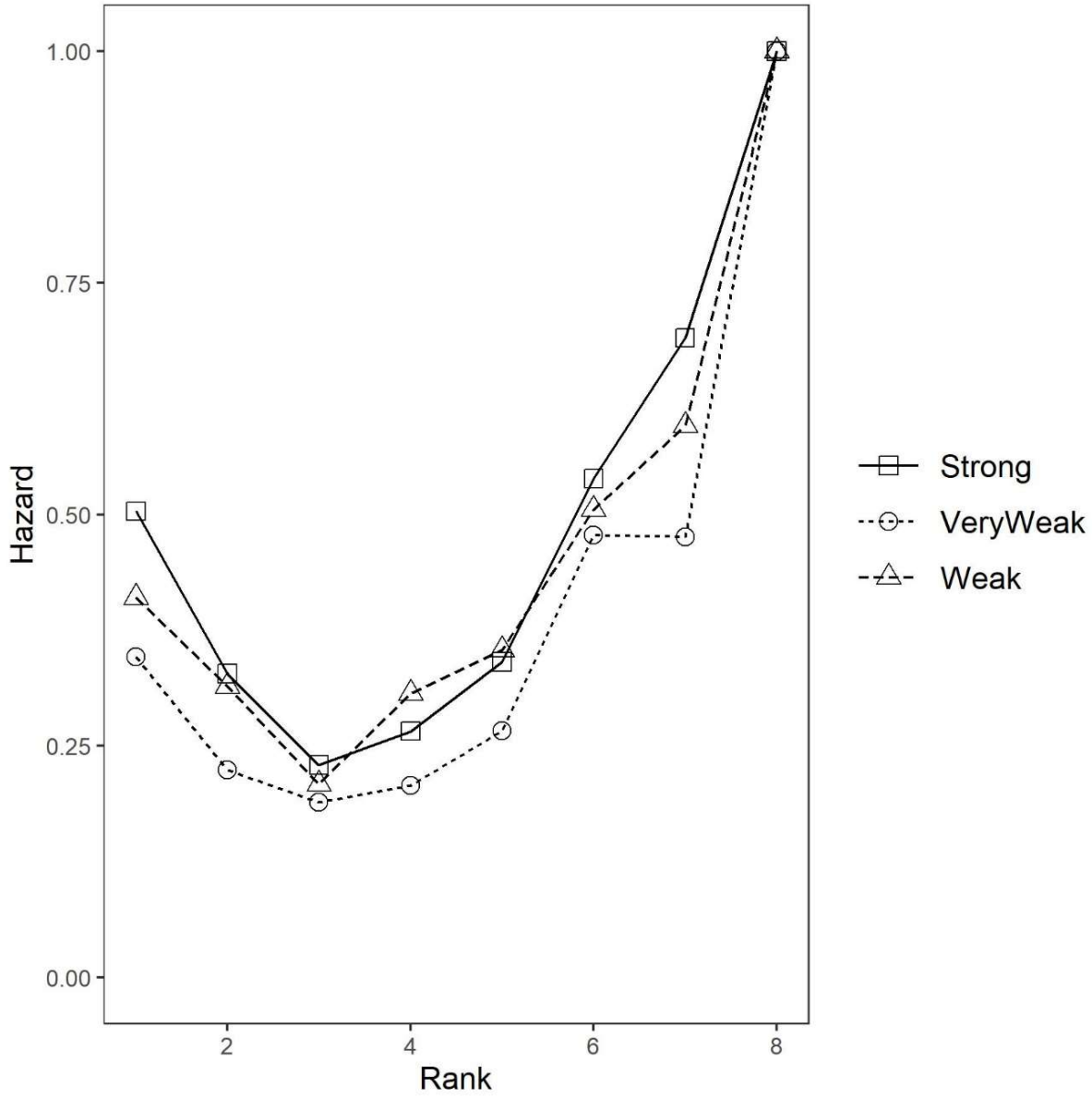
High-threshold representation

To test the assumption of high-threshold representation, target hazard rates were calculated from target-present lineup rankings of participants in each of the memory manipulation groups in Experiment 2 (see Table 2) and are illustrated in Figure 4. On inspection, the target hazards for $x \geq 2$ are all non-monotonic and conform to hazards predicted by the SDT model of eyewitness identification. The closest fitting constrained hazard model for each memory groups produced the following fitting statistics: $\chi^2_{\text{strong}} = 5.24$; $\chi^2_{\text{weak}} = 7.95$; and $\chi^2_{\text{very weak}} = 1.10$.

Table 2: *Counts of Ranked Target Selections and their Hazard Rates (Experiment 2)*

	Strong memory		Weak memory		Very weak memory	
Ranking	Count	(hazard)	Count	(hazard)	Count	(hazard)
1	368	(.503)	272	(.410)	227	(.346)
2	119	(.328)	123	(.315)	96	(.224)
3	56	(.223)	56	(.209)	63	(.189)
4	50	(.266)	65	(.307)	56	(.207)
5	47	(.341)	52	(.354)	57	(.266)
6	49	(.539)	48	(.505)	75	(.478)
7	29	(.691)	28	(.596)	39	(.476)
8	13	(1.00)	19	(1.00)	43	(1.00)

Figure 4: *Observed Ranking Hazards by Memory Strength (Experiment 2).*



To determine if the observed data was generated by a non-HT ranking process a bootstrap hypothesis test was conducted. To achieve this, n = 1,000 target ranking data sets were independently generated from the multinomial distribution of the empirical target rankings. The closest fitting constrained hazard function (as predicted by high threshold theory) was then found for each simulated observed data set. A second (paired) set of predicted rankings was then generated. This was achieved by first deriving a multinomial distribution of predicted target rankings from the constrained hazard function generated by the fitting process, and then finding the closest fitting constrained hazard function. These hypothesized ranking distributions represent the underlying binomial distribution of target rankings under the hypothesis that their hazard rates are monotonically increasing for $x \geq 2$. The probability that our fit values for the three memory groups were less than any value within their respective hypothesized distributions were: $p_{\text{strong}} = .043$; $p_{\text{weak}} = .013$; and $p_{\text{very weak}} = .383$. We may therefore be confident in our conclusion that the study's observed target hazard for x = >1 becomes non-monotonic as memory strength increases, thus falsifying the high-threshold model of eyewitness identification processes and providing support for the SDT model's account.

This rejection of the 2HT model essentially means that false identifications *do not* exclusively occur through guessing behavior, nor are correct identifications simply a mixture of rare "true" identifications and "lucky/educated" accurate guesses. Instead, identification errors will occur because there exists inherent ambiguity within the information available to the eyewitness, and it is the strength of this ambiguity that dictates maximal decision accuracy. Thus, without changing discriminability parameters, the minimization of false identification errors will always come at the cost of missed identifications.

Verification of the SDT models

To investigate how well the SDT models predict eyewitness identification outcomes, the *equal variant signal detection theory* (EV-SDT) model and the *unequal-variant signal detection theory* (UV-SDT) models were tested by fit to the empirical data collected from the identification task in Experiment 2 under the assumption of a MAX decision rule[14] and tested by constraining the parameters. This allows us to connect the models with reality as we expect that discriminability will increase under strong memory conditions and criteria will change with the expectations that our participants had of target presence.

To reduce the effects of minor observation error, confidence ratings were blocked into five categories, such that across all the conditions approximately equal numbers appear in each category. The resulting categories, from the least confident (category 1) to the most confident (category 5) were: 0-38%, 39-58%, 59-69%, 70-79%, and 80-100%. Each confidence category is separated by a lower decision criterion *c*, such that the five categories are bounded by five decision criteria. For example, category 1 is bound on the left by criterion c_1 and on the right by criterion c_2, category 2 is bound on the left by c_2 and on the right by c_3, et cetera, with category 5 only having the upper bound c_5. Frequency counts for responses to target-present and target-absent arrays within each confidence rating category for each memory group are given in Table 3 (below).

[14] In an eyewitness identification model, the MAX decision rule states that the lineup member generating the maximum value of familiarity will be identified *iff* that value exceeds the critical value set by the eyewitness (i.e., the eyewitness's decision criterion).

Table 3: *Identification Outcomes by Memory Strength, Prior Expectation, and Confidence Group (Experiment 2)*

Expectation	Confidence	Correct ID	Correct Rejection	Filler (TA)	Filler (TP)	Miss
			Strong Memory			
High	1	12	10	17	22	6
	2	31	12	30	24	10
	3	39	10	40	35	8
	4	45	11	34	44	5
	5	61	4	26	27	6
Low	1	10	6	14	20	11
	2	22	12	22	30	20
	3	39	15	22	40	19
	4	30	12	20	24	14
	5	49	7	12	18	11
			Weak Memory			
High	1	12	5	16	26	12
	2	17	7	29	32	18
	3	23	10	30	35	10
	4	21	4	29	32	13
	5	50	9	19	16	11
Low	1	14	8	19	23	16
	2	15	10	12	25	16
	3	28	15	21	37	20
	4	25	15	17	29	15
	5	35	13	21	14	23
			Very Weak Memory			
High	1	16	5	28	32	10
	2	15	8	34	27	15
	3	19	12	22	48	13
	4	24	6	25	24	9
	5	19	8	14	22	9
Low	1	13	13	27	31	19
	2	16	13	17	49	30
	3	16	27	23	39	38
	4	17	9	20	22	16
	5	21	16	7	10	17

Note. Filler (TP) = Filler identification from target present lineup; Filler (TA) = Filler identification from target absent lineup; Confidence: 1 = 80-100%; 2 = 70-79%; 3 = 59-69%; 4 = 39-58%; 5 = 0-38%.

Table 4

SDT Model Parameters and Fits for Observed Identification Data (Experiment 2)

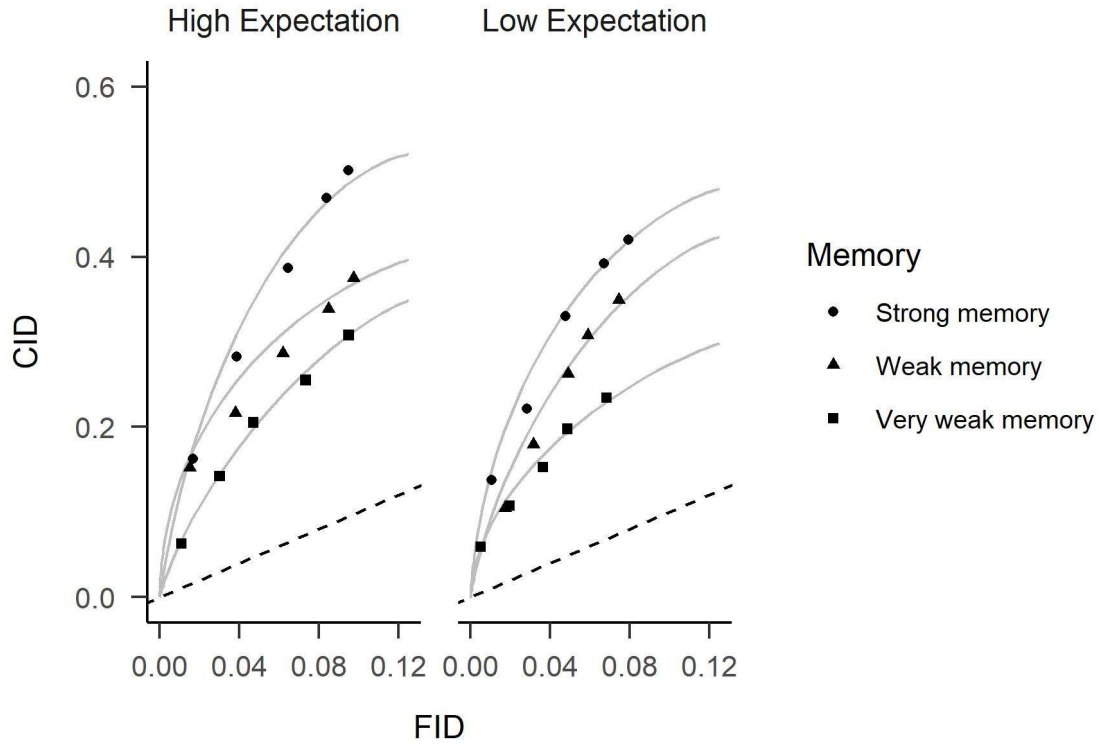
Memory	Expectation	Model	Model Parameters							Fit	
			c_5	c_4	c_3	c_2	c_1	d'	σ_{target}	$\chi^2(df)$	p
Strong	Low	EV-SDT	2.358	1.978	1.614	1.368	1.211	1.276	1.000	4.357 (12)	.993
		UV-SDT	2.174	1.760	1.449	1.196	0.999	1.390	0.798	10.903 (13)	.619
	High	EV-SDT	2.231	1.782	1.450	1.177	0.974	1.339	1.000	17.811 (12)	.216
		UV-SDT	2.354	1.977	1.614	1.369	1.212	1.281	0.987	4.341 (13)	.987
Weak	Low	EV-SDT	2.281	1.926	1.614	1.442	1.255	1.074	1.000	11.738 (12)	.627
		UV-SDT	2.255	1.840	1.524	1.260	1.052	0.984	1.222	14.051 (13)	.370
	High	EV-SDT	2.205	1.820	1.521	1.269	1.069	1.093	1.000	17.685 (12)	.221
		UV-SDT	2.258	1.915	1.613	1.446	1.263	1.128	0.904	10.889 (13)	.620
Very Weak	Low	EV-SDT	2.511	2.090	1.773	1.508	1.301	0.825	1.000	12.924 (12)	.533
		UV-SDT	2.297	1.885	1.552	1.298	1.023	0.908	0.925	11.132 (13)	.600
	High	EV-SDT	2.313	1.892	1.553	1.297	1.017	0.858	1.000	11.666 (12)	.633
		UV-SDT	2.585	2.119	1.782	1.507	1.291	0.607	1.249	9.284 (13)	.751

Note. Model fits assume the MAX decision rule. $c_5, \cdots, c_1$ = confidence criterion from low confidence to high confidence.

Testing by fit

The ROC-space illustrating the observed correct identification rate (CID)—false identification rate (FID) pairs for each confidence rating category and the theoretical ROC curve predicted by the closest fitting MAX UV-SDT model, for the strong and very weak memory manipulation groups are shown in Figure 5, below. Within the space, CID-FID confidence rating pairs falling closer to the origin represent decisions made at higher levels of confidence, and curves falling further above the chance line represent decisions made with higher levels of sensitivity. The χ^2 goodness-of-fit between the predicted and observed data were subsequently minimized through the adjustment of seven model parameters ($c_1, c_2, c_3, c_4, c_5, d', \sigma_{\text{target}}$) using a maximum likelihood method (Dunn, 2010). The model parameters and fits, assuming a MAX decision rule, for each of the four manipulation groups are reported in Table 4. While both the EV- and UV-SDT models provided a good fit across all manipulation groups, the UV-SDT model provides additional information relating to relative distribution of latent familiarity values. This information may provide new opportunities to discover significant differences in performance between various lineup procedures, especially in their predicted outcomes for highly conservative identification decisions, i.e., identifications made with a high level of confidence.

Figure 5: *ROC-space of 8-item simultaneous lineup outcomes for each experimental manipulation group (Experiment 2)*



Note. Empirical CID-FID pairs are shown as a set of five solid points and their corresponding best-fitting theoretical MAX UV-SDT ROC curve is shown as a solid curve drawn through these points. The dash line represents chance (no predictive value).

Testing by constraint

The independent effects of memory strength manipulation on eyewitness discriminability and expectation were tested independently on collapsed data. Parameters μ_{target} and σ_{target} of the UV-SDT model were first assessed by constraining them to be equivalent when fitting the UV-SDT to empirical ROC curves, while allowing the criterion parameters $c_1, \ldots, c_5$ to vary freely. The resulting constrained model was found to be a poor fit to identification data collapsed by memory ($\chi^2(5) = 26.88$, $p \leq .001$), but a good fit to identification data collapsed by expectation (thus confirming the link between the UV-SDT model's μ_{target} and σ_{target} parameters and the psychological construct of discriminability, as influenced by memory strength.

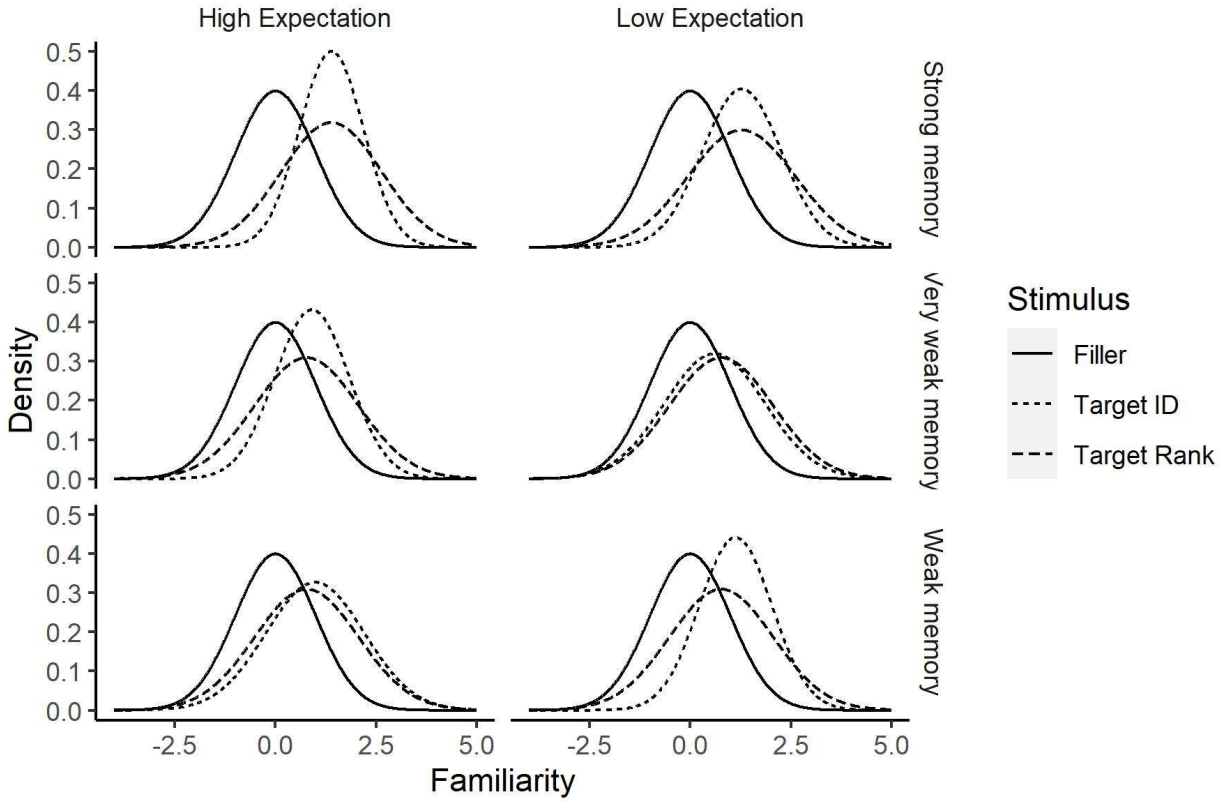
To test if there was any independent effect of prior expectation ($\chi^2(5) = 7.40$, $p = .192$) of a target present lineup on criterion placement, the model parameters $c_1, \ldots, c_5$ were constrained such they were equivalent, while allowing parameters μ_{target} and σ_{target} to vary freely. Again, the constrained model fit poorly to identification data collapsed by expectation ($\chi^2(2) = 28.21$, $p \leq .001$). However, there was also a poor fit with the identification data collapsed by memory ($\chi^2(2) = 18.91$, $p \leq .001$). So, while there was no effect found between the eyewitness's expectation of target presence and the discriminability parameters, the strength of the eyewitness's memory seems to have had a significant effect on both discriminability and decision criterion parameters. For example, eyewitnesses with stronger memories had both a larger μ_{target} parameter value *and* larger corresponding values for each of the c_i parameters.

Assessing independence of the latent variables

To assess the independence of the latent variable μ_{target} across experimental tasks the MAX UV-SDT model was directly fitted to the TP observations only. This allows for

greater parity between the lineup ranking and TP identification task model parameters. The resulting models for each identification task's (i.e., the traditional identification, TP only identification, and TP ranking tasks) decision space were compared (see Figure 6). While there were no significant differences in μ_{target} across the different task's model fits (M_Δ = -0.14, 95% CI = [-0.34,0.06], $t(7) = -1.79$, $p = .134$), it was noted that there were significant differences in σ_{target} parameter values (M_Δ = 0.24, 95% CI = [0.06,0.43], $t(7) = 3.40$, $p = .019$), with the variance of the distribution of target values being narrower within the UV-SDT model for identification tasks than for ranking tasks. These results indicate that the parameter μ_{target} remains independent and thus stable across different task structures. However, this does extend to equivalence in discriminability within the decision space. For example, changes in relative variance σ_{target}, from narrow to wide, between the decision spaces of two differing tasks may result in greater levels of accuracy for highly conservative identification decisions at the expense of the accuracy of more liberal decisions.

Figure 6: *Comparison of identification and ranking task UV-SDT models for each manipulation group (Experiment 2)*



Note. Each figure depicts the decision space for each manipulation group. While the μ_{target} parameter remains constant across the two tasks, variance of the target distributions (σ_{target}) are smaller for the identification task, than for the ranking task.

General Discussion

By providing evidence that eyewitness identification processes are stochastic in nature, this study has confirmed the appropriate use of mathematical models of eyewitness memory, such as those applied by Wixted and Mickes (2014), Lee and Penrod (2019), and Dunn, Kaesler, and Semmler (2022), to the analysis of empirical outcomes. By doing so, we can confidently move towards a transformation of the eyewitness memory theory-set: from one containing primarily verbally described models of behavior, to one including mathematical representations. As a result, our study provides three important benefits. Firstly, by allowing for the inclusion of mathematical models within the eyewitness memory theory-set, the underlying assumptions of the theory become more salient, thus allowing for a greater level of scrutiny in relation to the theory's overall psychological plausibility. Secondly, mathematical models of eyewitness memory provide more precise predictions of identification behavior. This makes it possible to critically test a model's hypothesized predictions. Essentially, by becoming more susceptible to falsification, and yet surviving such tests, a theory-set containing *good* mathematical models of behavior may not only be considered a strong theory but may also be regarded as the best available conceptualization of truth (Bjork, 1973; Clark, 2008; Hintzman, 1991). Thirdly, mathematical models provide more appropriate analytical methodologies to the evaluation and comparison of diverse eyewitness task structures, including different lineup sizes, filler similarity ratings, and presentation styles. For example, empirical ROC curve analysis, while useful when "comparing apples with apples", cannot account for different lineup sizes. This is because the manipulation of filler numbers will have a negative impact on outcome accuracy (and thus the area under the ROC curve) without any real impact to eyewitness discriminability performance, as measured by the model's parameters. As such, even the use of pAUC measures do not circumvent this problem.

While the development and introduction of mathematical models of eyewitness identification decision processes and tasks has been an essential step towards a more scientific approach to the analysis of eyewitness identification data, the axioms and assumptions underpinning these models have, until now, remained untested. As such, a second aim of this work was to discover and apply a critical test to the HT model class of eyewitness identification, so that they may either be accepted as the sole representation of eyewitness memory or be falsified and rejected. We identified a critical test of HT representation—monotonic hazards for second and subsequent choices—which was subsequently proven by Chechile and Dunn (2021). When applied to the empirical ranking data collected in experiment 2, the HT model was shown to be false and was thus rejected. By falsifying the high threshold account, the surviving SDT model not only supersedes it, but remains the only viable mathematical model of eyewitness memory.

The final aim of our study was to confirm the viability of this model by testing its predictive ability, the independence of its parameters and their linkage to psychological constructs of discriminability and identification decision bias, and its ability to be generalized across different task structures. By applying theoretical ROC curve analysis, we found the UV-SDT model both accurately predicted eyewitness identification outcomes and provided an additional parameter for exploration—relative variance σ_{target}. Furthermore, by constraining the UV-SDT model, we were also able to demonstrate that its parameters μ_{target} , σ_{target} align very well with the psychological constructs of eyewitness discriminability (as manipulated by memory strength). Following from these results is the disappointing reality that we must also reject the assumption that there exists a set of "true" identifications. Instead, we are forced to accept the fact that eyewitness identifications will always include the risk of error. Indeed, no matter how convincing the eyewitness or strict the identification task, eyewitness identifications may never be considered infallible and thus

should never be relied upon as the only source of evidence in a prosecution (Wells, 1984). Lacking a source of "true" identifications, it is impossible to create any gain in identification accuracy without bearing a cost of increases to missed identifications.

Based on the evidence provided by this study, we are confident that SDT models of eyewitness memory provide precise, accurate predictions of eyewitness identification decision behavior, using a set of parameters that link to the psychological constructs underlying that behavior, namely discriminability and decision bias. However, it must be remembered that mathematical models of human behavior, such as the SDT models of eyewitness memory are more than mere measurement devices. While it is obvious that they provide an analytical framework within which to interpret empirical observations, mathematical models also provide a conceptual framework within which researchers might ask novel questions about the cognitive processes involved in eyewitness identification decision making. But most of all, together with existing verbal models they form a unified theoretical framework of eyewitness memory, describing the relationships between lineup stimuli, presentation formats, performance expectations, the impact of a-priori information, post-identification confidence mapping, and eyewitness identification accuracy. As a result, eyewitness researchers need no longer work within a field devoid of a strong, unified theory. Indeed, great opportunities abound, not only for the advancement in our understanding of eyewitness identification decision making, but also in the advancement of procuring eyewitness identifications of consistently high probative value to support courts in their administration of justice.

Section III: Comments on the use of SDT models in eyewitness identification research

As discussed in Section I, mathematical models of eyewitness identification behavior provide a wide range of benefits. They make assumptions explicit, facilitate the evaluation of hypotheses, and provide a framework for thinking. In addition, mathematical models enhance our understanding of the richness of our data (Bjork, 1973), prompting theorists to systematically explore various assumptions and constraints. While Section II provided validation evidence for the use of SDT models of eyewitness identification, Section III will present the ways in which SDT models have transformed the field via the dissemination and uptake of new measurement tools and the introduction and testing of novel model adaptations to task specific structures. It will also consider assumptions which have not yet been fully explored, but which suggest the investigation of novel approaches for the maximization of eyewitness reliability.

However, before we begin, it must first be noted that SDT models of eyewitness identification are contextualized as being models of *population behavior*, rather than models of individual cognitive processes. That is, SDT models of eyewitness identification behavior predict the judgment outcomes of many eyewitnesses who have been presented with a single choice set. As such, there are several benefits and limitations to the use of SDT models within the field. For example, because eyewitness identification experiments almost always operate at the population level, all distributions within SDT models can be assumed normal (See Case II of Thurstone, 1927/1994). This assumption mediates issues raised by Wixted, Mickes, Wetmore, Gronlund, and Neuschatz (2017), who argued that a lack of knowledge regarding distributional form seriously inhibited the application of mathematical models to eyewitness

memory. Unfortunately, by conducting experiments at the population level, researchers can only gain insights to the effects of estimator and system variables at that level. They therefore cannot provide predictions of individual performance, nor can they provide precise information regarding the reliability of any *single* eyewitness identification decision.

There is much that may still be gained by continued exploration of eyewitness identification behavior at the population level through the testing of theories and hypotheses using mathematical models. This section provides a brief overview of the ways in which the introduction of the SDT model of eyewitness identification has supported researchers in their attempts to improve the measurement of outcome variables, access the full richness of eyewitness memory data, including metacognitive information, and develop stronger theories of eyewitness memory. Lastly, a conclusion is given with suggestions for the future focus of eyewitness studies based on the findings of this thesis.

A tool for measurement

Within eyewitness memory research, there have emerged two methodologically distinct approaches to the analysis of empirical eyewitness identification data—ROC analysis and mathematical model fitting—each of which have significantly impacted our understanding of eyewitness memory and the variables affecting identification performance. The analysis of ROC decision spaces provide an atheoretical method for the measurement of discriminability and there exists a range of open-source software available to researchers who wish to analyze them, including ROCR for R (Sing, Sander, Beerenwinkel, & Lengauer, 2005) and pROC for R (Robin et al., 2011). In contrast, model fitting is a methodology based in theory, as it requires a formal model of behavior from which to simulate predicted outcomes. By adjusting the parameters of the predictive model, the fit between observed and expected outcomes may be minimized and the parameters of the latent variables estimated.

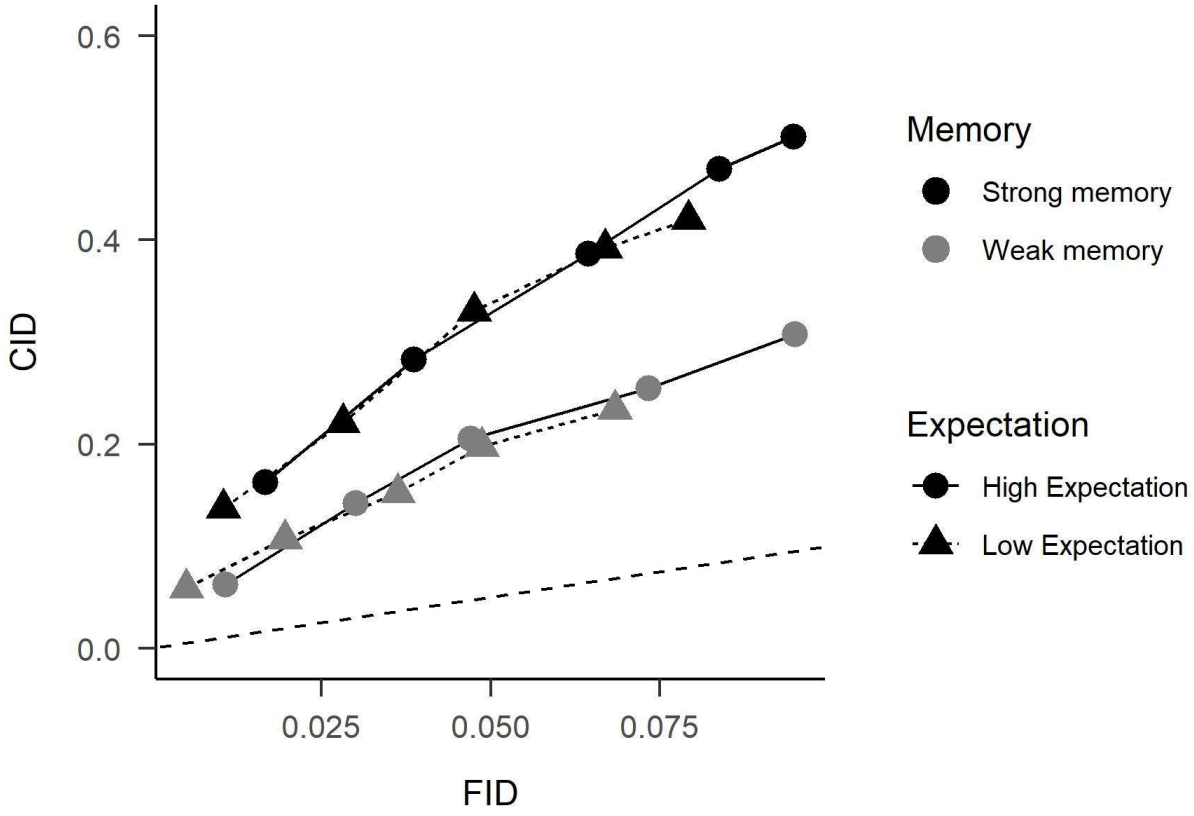
While model fitting generally requires a deep understanding of mathematical and/or computational modelling, this approach has become much more accessible to eyewitness researchers through the recent release of new and comprehensive open-source software, such as pyWitness for Python (Mickes, Seale-Carlisle, Chen, & Boogert, 2022, March 9), which provides for the fitting of empirical data to SDT models of eyewitness identification, data simulation, and power analysis, in addition to the production and analysis of ROC curves. This chapter provides a brief commentary of these various approaches, focusing on how SDT models inform the analytical frameworks for the measurement of eyewitness identification performance, and emphasize the richness inherent in the data.

Estimating information metrics: ROC curve analysis

An initial catalyst to the broader introduction of mathematical models to eyewitness identification data was the popularization of receiver operating characteristic (ROC) curve analysis (Egan, 1958) by Mickes, Flowe, and Wixted (2012), which was later confirmed as a valuable tool for the independent measurement of performance and response bias in empirical eyewitness identification studies (Gronlund & Neuschatz, 2014). Grounded in mathematical models of signal detection, ROC curves provide a visual means to evaluate changes in discriminability and/or bias across experimental manipulations and are constructed using pairs of cumulative summary statistics—the rate of correct identifications from target-present trials (CID) and the rate of false (suspect) identifications from target-absent trials (FID)—across different levels of decision bias. The position of each CID-FID pair within the ROC decision space is governed both by discriminability performance and the eyewitness's bias towards identifying someone from the lineup, or not. For example, if a curve is drawn from the origin through each CID-FID pair within the ROC decision space—the empirical ROC curve—and its position within that space will provide information regarding discriminability

performance (distance from the diagonal) separately from information about the eyewitness's decision bias (distance from the origin; see Figure 7, below).

Figure 7: *Eyewitness identification ROC decision space for experimental manipulations of memory strength (strong vs weak) and target expectation (high vs low).*



Note. As memory strength increases, the ROC curve shifts upwards, and the partial area beneath the curve (pAUC) increases. As identification decision becomes more conservative, the ROC curve shifts towards the origin and the curve becomes more truncated.

As they make no assumptions regarding the distributional form of sensory information, ROC curves are considered a non-parametric index of discriminability (Wixted et al., 2014), with their analysis simply requiring the estimation of their integral, commonly referred to as the 'area under the curve' (AUC). This measure represents a summary of performance that is directly linked to the SDT model of the 2-AFC task.[15] ROC curves which lie upon the diagonal, from the origin to unity, have an AUC measure of 0.5, and indicate random guessing (i.e., no memorial information). In contrast, an ROC curve with an AUC measure of 1.0 never actually enters the decision space, and thus indicates perfect discriminability.

The usual application of AUC analysis to binary decisions outcomes, however, does not translate perfectly to the outcomes of the lineup task. This is because lineup tasks provide an additional identification choice to the eyewitness—filler identification. Thus, as a direct result of the *k*-item design of eyewitness identification tasks, the eyewitness identification ROC curve becomes naturally truncated, with the maximum false alarm rate for a lineup of size k, being $1/k$. Furthermore, where a dedicated 'innocent suspect' has not been selected for target absent lineups, FID must be calculated by dividing the rate of filler identifications from target-absent lineups by the lineup size *k*. However, this version of FID represents the conditional probability that the suspect is selected given that the eyewitness identified someone from a *fair* target-absent lineup, an assumption that is rarely tested.

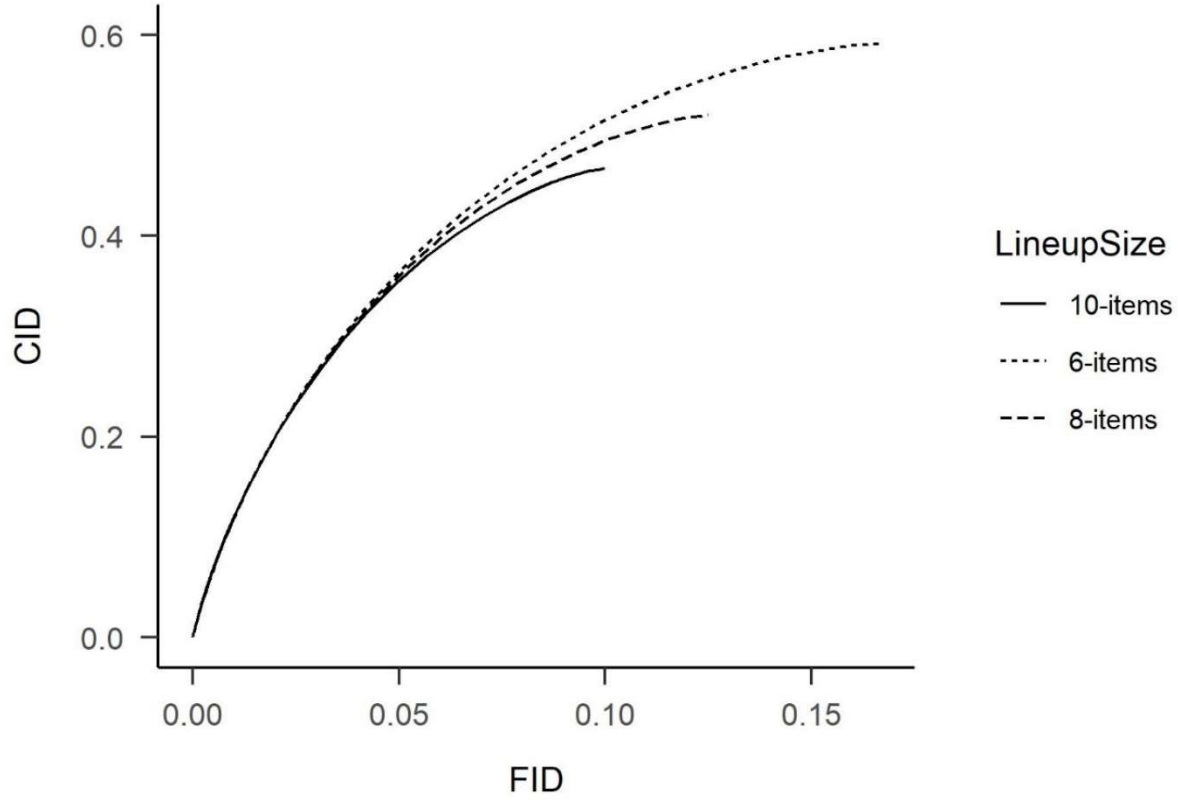
The inclusion of fillers in the identification task also places structural limitations on rates of correct identification. For example, when presented with a target-present lineup, unless the eyewitness can perfectly discriminate between the guilty suspect and the fillers, there is always a chance that they will erroneously identify a filler. Thus, even at the most liberal decision bias, where an eyewitness will always choose someone, the rate of correct

[15] For a mathematical proof, see Egan (1975).

identification can never reach the maximum value of 1. Indeed, as decision bias becomes more liberal, the effect of such truncation becomes more pronounced, and the ROC curve begins to 'flatten' (see Figure 8). In consequence, unlike standard ROC curves constructed from 2-item tasks, which extend across the whole decision space, the analysis of ROC curves drawn using empirical lineup data is limited to the area under the partial curve (pAUC).

While the introduction of ROC curve analysis to the field of eyewitness memory was instrumental to furthering our understanding of eyewitness identification data, this atheoretical methodology has two inherent weaknesses. Firstly, while the calculation of pAUC for a single ROC curve seems quite straight forward, the comparison of two or more curves produced under differing experimental manipulations is not. This is because decision bias under one manipulation may be more conservative than under another, causing each curve to terminate at different values of *FID*. As a result, the calculation of pAUC requires the analyst to make a decision regarding the most appropriate cut-off point to use when comparing the level of discriminability between two or more test groups. If it is decided that this point is to be made at the maximum *FID* of the shorter curve, then not all the available data informs the analysis. If it is instead decided to use the maximum *FID* of the longer curve, then the shorter curve will need to be extrapolated using a smoothing function. Although it has been argued that this limitation of pAUC analysis are of no particular consequence for policy makers, since its analysis is conducted within the `area of interest', that is, where the eyewitnesses have made their identifications (Wixted, Vul, Mickes, & Wilson, 2018), they should not be ignored by researchers.

Figure 8: *Decision space depicting ROC curves predicted by SDT models of eyewitness lineups of the same level of discriminability, but with differing lineup sizes: 6-items; 8-items; and 10-items.*



Note. Under the MAX decision rule, as lineup size increases, the ROC curve becomes more truncated, with CID values flattening out as FID increases. ROC curves of larger lineups are predicted to have a smaller area under the curve than ROC curves of smaller lineups, despite having equivalent discriminability parameters (d', σ_{target}).

Secondly, the drawing of a traditional ROC decision space using eyewitness identification data does not require rates of filler identifications from a target-present lineup. Indeed, it completely ignores this empirical data under the assumption that such identifications provide no information to the estimation of discriminability. However, the loss of such information, simply because it does not 'fit' the suggested measurement tool inherently reduces the statistical power of pAUC measures of discriminability.

Estimating theoretical parameters: SDT model fitting

The fitting of SDT models to empirical data allows researchers to estimate the model's parameters representing estimated discriminability ($\widehat{d'}$) and bias ($\hat{c}$). This is achieved through a process of maximum likelihood estimation, in which the parameters of the assumed model are optimized, such that predicted outcomes provide as close a match to the observed outcomes as possible.[16] Not only does this method link the analysis of empirical data to the prevailing theory (and the assumptions of that theory), it also requires the full set of identification outcome information, information that is often ignored in the analysis of ROC curves (Wells et al., 2015), and thus improves the statistical power of the measure. However, this is not to say that ROC decision spaces are not relevant, indeed they provide an accessible communication tool for researchers in the dissemination of their results. Indeed, recently published eyewitness identification studies report the results of both ROC curve analysis and SDT model fitting, with the ROC decision space presenting both empirical CID-FID pairs and ROC curves predicted by the fitted model, alongside estimated model parameters (for example, see Colloff, Flowe, & Smith, 2021; Colloff et al., 2022; Colloff, Seale-Carlisle, Karoğlu, & Rockey, 2021; Winsor et al., 2021). This use of methodological triangulation

[16] The predictive ability of a model is measured by some fitting statistic (e.g., χ^2, G^2, et cetera), which is maximized such that under the assumed model, the observed data is most likely.

serves to increase our understanding relating to the precision and power of our analysis, it also allows for a greater level of transparency in relation to the assumptions underlying the analysis of empirical data. This is because studies which report estimations of SDT model parameters must also provide details regarding the assumed model, including the constraints placed on the model's parameters and the decision rule applied to it.

A tool for developing theory

The introduction of mathematical models of measurement to eyewitness identification research has also facilitated an increased awareness of the available richness within eyewitness identification data, allowing researchers additional avenues for the development and testing of new hypotheses and existing theories. While this thesis has focused on testing existing theories and the selection of the most appropriate class of mathematical model to studies of eyewitness memory, other researchers have already begun to explore SDT models for the purpose of further developing our understanding of eyewitness identification behavior. This chapter provides an overview of this work, with a focus on how the SDT model allows for the facilitation of such investigations, namely the relaxation of model constraints and the introduction and exploration of parameters and their proxy's.

However, before there can be a discussion on how the SDT model's parameters have been thus far explored by eyewitness theorists, it is important to first explore the seminal work of Thurstone, whose Law of Comparative Judgment (Thurstone, 1927/1994) provides the grounding to SDT. Thurstone made explicit six assumptions, each of which apply to SDT: (1) that different stimuli will elicit different stimulus values through some unknowable cognitive process; (2) that these values exist along a single psychometric continuum; (3) that the values elicited by stimuli presented simultaneously are correlated; (4) that values elicited on successive exposures to a stimuli form modal distributions; (5) that the position held by

each distribution on the continuum is unique; and (6) that the variance of such distributions are unequal. Each of these assumptions, with exception of the first, provides opportunities for eyewitness memory theorists to explore new models of eyewitness identification behavior. An obvious example of such an exploration has been the focus of this thesis, which critically tested assumptions two (2) and four (4), and validated assumption six (6). This leaves us with the theoretical explorations of assumptions five (5) and six (6), that is, the consideration of non-zero correlation between filler and target stimulus values on each presentation, and of non-equal distributions of filler stimulus values within the population dataset, each of which forms the discussion within the following sections of this chapter.

Relaxing the constraint of zero-correlation

While excluded from the original SDT models (Swets et al., 1961), as well as from SDT models of memory (Egan, 1958), Thurstone's (1927/1994) description of a Theory of Relative Judgements included a parameter which allowed correlation between stimulus values by including a correlation coefficient ρ within the formal model description. Where *simultaneous contrast* existed within a choice set, their attributed values would become negatively correlated, forcing the exaggeration of the discriminal difference between them (i.e., the extent of distributional overlap between stimuli values). Thurstone suggested that it was "…probable that this effect tends to be a minimum when the specimens have other perceivable attributes, and that it is a maximum when other distracting stimulus differences are removed" (p. 268). That is, when speaking of eyewitness identification lineups, the effect of simultaneous contrast is at its maximum when a lineup is *fair* such that each of its members closely resemble the description of the culprit, such as sex, hair and skin color, age, build, distinctive features etc., and at its minimum when only the suspect resembles the culprit. Thus, by removing distracting stimulus differences, such as those which make the

lineup unfair, and by presenting stimuli simultaneously, the overall decision performance may be maximized.

Thurstone's simultaneous contrast hypothesis was first introduced to eyewitness identification research by (Wixted & Mickes, 2014) to explain ROC curve superiority of simultaneous vs sequential identification task presentations (see also Carlson, 2019; Colloff & Wixted, 2020; Terrell, Baggett, Dasse, & Malavanti, 2017; Wilson, Donnelly, Christenfeld, & Wixted, 2019). By adapting Thurstone's original *Case I* of the *2*-AFC task model (in which non-zero ρ was assumed to remain fixed across trials) to *k*-item simultaneous lineups, Wixted et al. (2018) developed a SDT model of eyewitness identification—the *ensemble* model. This novel model was found to successfully account for the superior discriminability performance of simultaneous lineups, when compared sequential lineup presentations, within the ROC decision space. Akan, Robinson, Mickes, Wixted, and Benjamin (2020) later introduced a second, but similar model of simultaneous lineup presentation in which a correlation coefficient parameter was added to an existing SDT model of eyewitness identification—the *independent observations* model. Their resulting *dependent observations* model provided near identical predictions to that of the ensemble model, providing further support to Thurstone's hypothesis of simultaneous contrast.

Relaxing the constraint of filler equivalence

While it is important for formal models to be parsimonious, that is to include as few a number of parameters as possible whilst maintaining satisfactory level of predictive power, much information about a phenomenon can be lost through the process of model reduction. The development of SDT models of eyewitness behavior is not only a good example of such a loss, but also an example of how such losses may be regained. The most poignant example is that of the assumption that all familiarity values attributed to innocent lineup members are

drawn randomly from the same distribution. This assumption is made for the purpose of mathematical simplicity and is generally explained away by the fact that the eyewitness has not been exposed to innocent lineup stimuli. However, this need not be the case. For example, the modification of the SDT model to *unfair* lineups—in which the suspect resembles the description of the culprit more than do any of the fillers—sees the distribution of both the target and the innocent suspect distributions lie above those of the filler distributions on the continuum and the SDT decision space includes three, rather than two distributions (see Figure 9; Lee & Penrod, 2019; Wickens, 2002, p. 126).

Expanding upon the notion that stimuli within lineup tasks might have non-equivalent distributions, Lee and Penrod (2019) sought to discover how fillers within both target-absent and target-present lineups might influence identification performance. Their Multi-*d'* model allowed for the relaxation of the assumption of distributional equivalence across innocent filler and suspect lineup members. By allowing distributional means to freely vary while keeping variance equal, the authors were able to provide a novel approach to the investigation of lineup bias (fairness), filler selection, confidence ratings, and presentation formats. By computing $\widehat{d}'$ from z-transformed outcome pairs, these authors' relaxation of SDT model parameters illustrates the way in which mathematical models support various 'thought experiments', opening us to new ways of thinking about the interaction between eyewitness memory, lineup stimuli, presentation structures, and metacognition.

Relaxing the constraint of fixed criterion placement

SDT models also make assumptions about how decision criteria are positioned within that space. While this thesis presented only one decision rule—the MAX decision rule—which is the decision rule governing the *independent observations* model, many alternatives are available. For example, it may also be required that the sum of the stimuli values also

exceed some minima, transforming the independent observations model into the *integration* model (Duncan, 2006), or it may be required that the stimulus with the maximum value exceeds the mean value of the remaining items by some minima, transforming it instead into the *ensemble* model (Wixted et al., 2018).

These, and other, decision rules are usually applied to the whole lineup stimulus set, with the criterion being constrained to a single value. However, Dunn et al. (2022) suggested that such a constraint may not be appropriate across the different positions in a sequential lineup. By allowing the decision criterion to vary across lineup positions, their proposed *independent sequential lineup* model revealed that response criterion is both sensitive to the underlying discriminability of the eyewitness (in the presentation of the first two items) and changes depending on whether the eyewitness has been exposed to the target during the test phase. More specifically, decision criteria become progressively conservative before target presentation, very conservative directly after target presentation, and then more liberal as the presentation progresses further. This finding is quite astonishing, as it says much, not only to our view of eyewitness memory as an abstract construct, but also to the relationship between the eyewitness and their own memory trace, as well as to their understanding of the demands of the task at hand. Finally, these researchers concluded that their model was both broadly similar and consistent with the ensemble model (Wixted et al., 2018) and that it was supportive of the broadly accepted theories of relative judgment (Thurstone, 1927/1994; Wells, 1984). Thus, by investigating one parameter of the SDT model, these theorists were able to triangulate their findings with those of investigations of another SDT model parameter, resulting in the further strengthening of SDT model-based account of eyewitness memory.

Exploring the mapping of certainty to identification decisions

While mathematical models allow us to formalize measurement in ways which facilitate both inductive and deductive reasoning in the development of theory within a single field of study (Robinaugh et al., 2020) and as across multiple domains in the development of overarching theoretical frameworks (Muthukrishna & Henrich, 2019), within the field of eyewitness identification, the most profound effect of SDT models has been the increased awareness of the richness inherent in eyewitness identification data, particularly in relation to confidence. Previously only investigated through a statistical lens (e.g., correlational studies and calibration analysis), our conceptualization of confidence ratings and their relationship to eyewitness memory and identification decision making was transformed by the introduction of confidence ROC curve analysis. Eyewitness identification ROC curves are constructed using confidence ratings[17] in proxy for decision bias, with ratings of high-confidence inferring a highly conservative decision bias and ratings of low-confidence inferring a comparatively liberal decision bias. Under the assumption of the SDT model, the mapping of confidence ratings seems, at first, relatively straightforward, with each ordered class of confidence being separated by a decision criterion in a way which adheres to the order constraints of decision bias (see Figure 9). When mapping confidence ratings directly onto the eyewitness identification SDT model decision space, it is easy to see that as confidence increases, decision bias becomes more conservative, resulting in higher post-predictive values (PPV) of identifications. Thus, as a measure of accuracy, PPV is directly tied to the decision criterion and providing an index of response performance, rather than an index of overall performance potential, such as pAUC or $\widehat{d}$' (Semmler, Dunn, Mickes, & Wixted, 2018).

[17] In eyewitness identification experiments, confidence ratings are collected immediately after the identification decision. Eyewitnesses either provide a percentage level of confidence (0-100% scale) or may select from a defined Likert scale labelled with either numbers or words representing level of confidence (e.g., 1,2,3… or low, medium, high et cetera). Where eyewitnesses provide a percentage of confidence, these may be grouped into equally sized ordered groups (e.g., deciles).

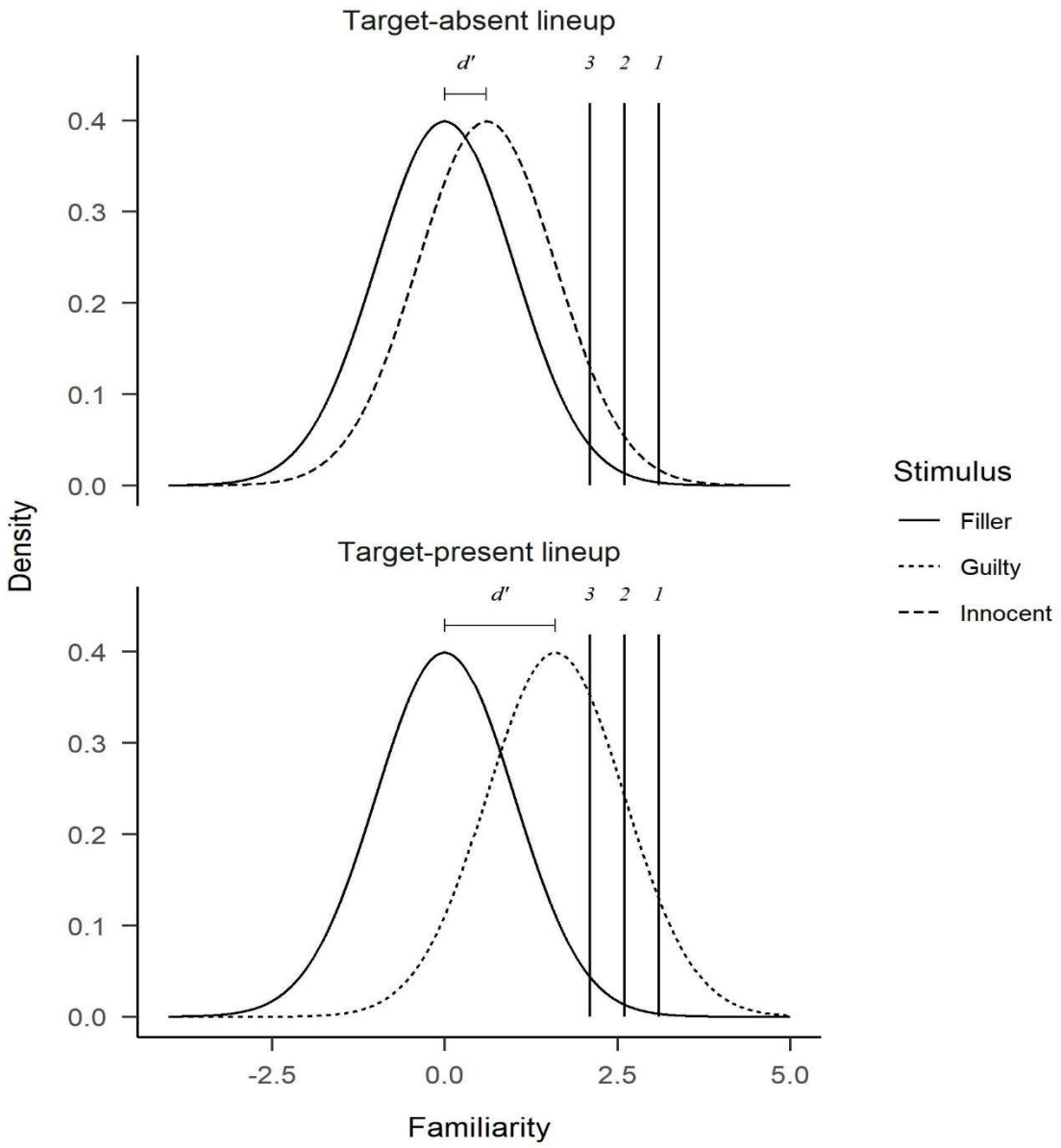
By plotting the PPV-confidence pairs onto a *confidence-accuracy characteristic* (CAC) decision space (Mickes, 2015), analysts may investigate how confidence might be affected by different experimental manipulations. Where the positions of confidence criteria are spaced close together, changes in PPV are small are the CAC curve is flat. As a result, confidence is neither correlated with, nor calibrated to identification accuracy. This effect is commonly seen in children (Brackmann, Sauerland, & Otgaar, 2019), who have difficulty expressing certainty using standard ratings of confidence (e.g., percentage ratings), and as a result, require different methods of indicating expected accuracy (Winsor et al., 2021). Where the positions of confidence are 'fanned out' across the SDT decision space, changes in PPV are high and the CAC curve is both positive and steep. While calibration is not necessarily achieved, correlation is strong. This effect is commonly seen in eyewitness experiments manipulating estimator variables, where very high ratings of confidence consistently predict PPV $> 90\%$, assuming *pristine* experimental conditions[18] are maintained (Wixted & Wells, 2017). And yet not all high-confidence ratings predict high levels of accuracy, especially where so-called pristine conditions are not met, such as where the lineup is unfair or biased towards the selection of an innocent suspect (Sauer, Palmer, & Brewer, 2019).

These findings reveal that the mapping of confidence criteria to the SDT model decision space is extraordinarily complex, with decision criterion placement being sensitive to both underlying discriminability (Wixted, Read, & Lindsay, 2016), expected base-rates (this thesis), and the lineup size (Wooten et al., 2020). This begs the question: How do eyewitnesses modify their decision criterion (and by proxy, their confidence) such that high-confidence decisions are consistently accurate? An answer was offered by Semmler et al. (2018), who suggested a *constant likelihood ratio* model, whereby eyewitnesses will only

[18] Pristine testing conditions are defined by Wixted & Wells (2017) include "initial, uncontaminated memory tests using fair lineups, with no lineup administrator influence, and with an immediate confidence statement (p. 1)

indicate a high confidence rating if, and only if, a consistently high level of accuracy for that particular level of certainty is attained. In direct opposition to the *optimality hypothesis* (Deffenbacher, 1980), the constant likelihood ratio model suggests that eyewitnesses take into account both underlying discriminability and base rates when placing their identification decision criterion, with decisions made under different levels of certainty fanning out accordingly. While such fanning out is often seen with different patterns of placement between groups, it is the anchoring of high-confidence decisions, both within and between groups, that has gained the most attention. Indeed, it seems remarkable that eyewitnesses can behave as Bayesian observers when making their identification decisions.

Figure 9: *Multi-d' decision space of target-absent and target-present lineups depicting a biased (unfair) target-absent lineup in which the innocent suspect stands out from the lineup fillers.*



Note. The upper figure shows the decision space for an unfair target-absent lineup in which the density function of familiarity values for the innocent suspect has shifted to the right of the density function of familiarity values for the fillers, which all share the same distribution set to standard normal. The lower figure shows the decision space for the target-present lineup, in which the filler values also share the same standard normal distribution while the

familiarity values for the guilty suspect has shifted a larger distance to the right, such that $d'_{innocent} < d'_{guilty}$. The mapping of confidence criteria (vertical lines labelled one, two, and three) have been constrained to be both ordered (from low confidence to high confidence as x increases) and equivalent (across the target-present and target-absent lineups decision spaces).

Unfortunately, the constant likelihood ratio model is, as yet, unable to account for identification decisions made under less pristine conditions, such as biased or unfair lineup presentations. For example, where an innocent suspect stands out in a lineup, eyewitnesses tend to be overconfident and as a result, high-confidence identifications become less accurate and leaving the usage of high-confidence ratings as an indicator of probative value within courts premature (Berkowitz, Garrett, Fenn, & Loftus, 2022). Saraiva et al. (2020) suggests that eyewitnesses may use a potentially misleading heuristic—perceived task difficulty—as a proxy for underlying discriminability, especially where fillers become implausible options or where memory self-efficacy is rated as very high or very low. Thus, it seems that, as true Bayesian observers, eyewitnesses apply all available information during their eyewitness identification decisions, including metacognitive information not currently included in existing SDT models of eyewitness memory.

Conclusion and future focus

The development and verification of mathematical models of behaviour within applied fields of study is neither a simple nor an easy process. Yet, the potential rewards are great. Within the field of eyewitness memory, mathematical models of identification decisions have combined with verbal models, creating a succinct theory of identification performance. This newly emerging theory-set provides precise predictions of the full range of identification outcomes, as well as comprehensible explanations of complex eyewitness decision making behaviour. This thesis has not only confirmed the fact that empirical eyewitness identification data can be represented by a mathematical model, but it has also given strength to the overarching theory of eyewitness memory by providing a set of models—SDT models—which have survived a hierarchy of critical tests of model properties (Kellen et al., 2021), including high-threshold representation and latent variable

independence. It has also verified the psychological plausibility of the model's parameters of discriminability and bias, as well as its ability to accurately predict empirical outcomes. Finally, a brief overview was provided on the numerous ways in which SDT models of eyewitness memory have made a positive impact to eyewitness identification research, both as a tool for measurement, and as a tool for theory development.

While much has been achieved in recent years, it has been noted that until the time of writing, the uptake of analytical frameworks grounded in SDT, or indeed any mathematical model of eyewitness identification, has been slow. Many researchers within the field continue to apply statistical methodologies in the interpretation of their empirical data, often in ways which disregard the richness of the data. However, this is not surprising due to the difficulty in making shifts towards theory-based data analysis, especially where those theories are being actively developed. Although the recent introduction of new open-source software, such as pyWitness (Mickes et al., 2022) is promising, it should be noted that the accessibility of these analytical tools, particularly to those researchers who rely on proprietary software such as SPSS, is probably rather low. Indeed, without access to tools that provide a simple graphic user interface, such as a Shiny app or other interactive web-based programs, researchers are unlikely to adopt any new methodology, simply because it would require too great an investment in retraining. Time that most academics do not have to spare.

Considering that there is so much yet to discover, a discussion of the future focus within the field of eyewitness identification is potentially limitless, even if such a discussion was constrained to the development, testing, and verification of novel SDT models of eyewitness identification. This discussion will therefore be limited to the exploration of two figures presented within this thesis. To begin, let us consider Figure 9 from the previous chapter. This figure demonstrates the EV-SDT multi-*d*' model of eyewitness identification in which the innocent suspect 'stands out' from the fillers within the target-absent lineup. In this

model, it is assumed that the confidence criteria are equivalent across both target-absent and target-present lineup groups. However, neither equal variance nor criterion equivalence are necessary constraints. Under the structure of a SDT model of lineup tasks, it is possible to fit the model to the target-absent dataset, separately from the target-present dataset, since each of these decision spaces include the minimum structural requirements: two lineup member 'familiarity' value distributions (filler and suspect); and a set of decision criteria given at various levels of bias. By doing so, not only it would be possible to calculate the extent to which the innocent suspect stands out from the fillers $d'_{innocent}$ separately from the eyewitness' ability to discriminate the guilty suspect from the same fillers d'_{guilty}, it would also be possible to evaluate differences in distributional variance (i.e., between $\sigma_{innocent}$ and σ_{guilty}) and the distance of the criteria shift depicting overconfident responses to the unfair/biased lineup.

Another consideration for further discovery is the increase of performance within the identification 'area of interest' (i.e., where identifications are made by eyewitnesses performing lineup tasks). While it is imperative that the lineup task does not limit access to the memory trace (Dunn et al., 2022), it is also important that the relative variance of the target distribution is maximised, so as to maximise the likelihood ratio. This is evident in Figure 6 from Section II, which compares the UV-SDT model decision spaces for lineup and ranking tasks. The standard deviation of the estimated target distributions in the ranking task were greater than one under all conditions, with no significant difference in d'. When compared to the (within group) estimated target distributions of the identification task, the ranking task has potentially higher likelihood ratios within the area of identification decision making, that is, for high values of x. However, because participants do not make identification decision during a ranking task, such conservatively biased criteria cannot be measured without also collecting confidence ratings for rankings. Not only could confidence

ratings be collected for first ranked items, but they could also be collected for all rank levels. This would allow for the collection of additional metacognitive information regarding the relative differences between first and second ranked items, not unlike the confidence-based identification task suggested by Brewer, Weber, and Guerin (2019). By collecting such ratings after conducting a formal lineup task, investigators may potentially gain vital additional information regarding the guilt—or indeed innocence—of the suspect, without 'contaminating' the collection of primary evidence.

In conclusion, the introduction of mathematical models to the measurement of empirical eyewitness identification data has facilitated a shift towards the building of a body of *scientific* evidence. That is, evidence collected through empirical studies that apply theoretical and analytical frameworks grounded in strong theory, theory which is constructed from models that are not only falsifiable and psychologically plausible, but which survive critical testing.

Appendix 1: Formal models of eyewitness memory

This appendix provides the formal descriptions of the single high threshold (1HT), the double high threshold (2HT), and the signal detection theory (SDT) models of eyewitness identification for both showup and lineup identification tasks.

Continuous models of eyewitness memory

Let $f(\cdot)$ and $F(\cdot)$ be the Gaussian probability and cumulative density functions (respectively) of subjective non-target values, and $t(\cdot)$ and $T(\cdot)$the Gaussian probability and cumulative density functions (respectively) of subjective target values. Under the assumptions of the EV-SDT, discriminability, represented by a single parameter d', may be simply indexed by the distance between the means of the target and non-target sampling distributions. Parameter c is the decision criterion, which is positioned at some point along the continuum of familiarity. Following this, parameters d' and c for the showup task may be calculated as follows

$$d' = z(\mathrm{TID}) - z(\mathrm{FID}),$$
$$c = -\left[z(\mathrm{TID}) + z(\mathrm{FID})\right]/2$$

with the outcome probabilities of a showup identification task simply being $\mathrm{TID} = \int_c^{\infty} t(x)\, dt$ and $\mathrm{FID} = \int_c^{\infty} f(x)\, dt$, with their simple complements being MISS = 1 - TID and CR = 1 - FID. The calculation of TID and FID under the UV-SDT uses the same function, however, introduces the parameter s, being the standard deviation of the target distribution expressed in terms of filler SD units, in the calculation of $t(x)$.

For a lineup of size *k*, the addition of fillers complicates the calculation. The outcome probabilities are

$$P(TID) = \int_c^\infty t(x)F(x)^{k-1}\,dx,$$

$$P(FID_{TP}) = (k-1)\int_c^\infty f(x)T(x)F(x)^{k-2}\,dx,$$

$$P(MISS) = T(c)F(c)^{k-1},$$

$$P(SID) = 1 - F(c)^k \cdot 1/k$$

$$P(FID_{TA}) = 1 - F(c)^k \cdot (k-1)/k\text{, and}$$

$$P(CR) = F(c)^k.$$

Single high threshold (1HT) model

Let *t* be the probability that a guilty suspect is detected and *g* be the probability that the conditional probability that the eyewitness will identify someone from the lineup when in a non-detect state.

Showup task

The conditional outcome probabilities for a showup are

$$P(\text{guilty suspect ID}) = t + (1-t)g,$$

$$P(\text{missed ID}) = (1-t)(1-g),$$

$$P(\text{innocent suspect ID}) = g$$

$$P(\text{correct rejection}) = 1-g.$$

Lineup task

The conditional outcome probabilities for a target present lineup of size k are

$$P(\text{guilty suspect ID}) = P(\text{TID}) = t + \frac{(1-t)g}{k},$$

$$P(\text{missed ID}) = P(\text{MISS}) = (1-t)(1-g),$$

$$P(\text{filler ID}|\text{TP}) = P(\text{FID}_{\text{TP}}) = (1-t)g(1-1/k),$$

$$P(\text{correct rejection}) = P(\text{CR}) = 1-g,$$

$$P(\text{filler ID}|\text{TA}) = P(\text{FID}_{\text{TA}}) = g(1-1/k), \text{ and}$$

$$P(\text{innocent suspect ID}) = P(\text{SID}) = g/k.$$

Double high threshold model

Let *t* be the probability that a guilty suspect is detected, *f* be the probability that an innocent lineup member is detected, and *g* the conditional probability that the eyewitness will choose someone from the lineup when in a non-detect state.

Showup task

The conditional outcome probabilities for a showup are therefore

$$P(\text{TID}) = t + (1-t)g,$$

$$P(\text{MISS}) = (1-t)(1-g),$$

$$P(\text{SID}) = (1-f)g, \text{ and}$$

$$P(\text{CR}) = f + (1-f)(1-g).$$

Lineup task

For a target present (TP) lineup, given the lineup size *k* and *f*, the probability of $a \in \{1, 2, \ldots, k\}$ lineup members remaining *available* for selection is

$$P(\text{a}|\text{TP}; k, f) = s(a) = \binom{n-1}{a-1}(1-f)^{a-1}f^{n-a}. \qquad \textit{Equation 1}$$

Note that the total number of fillers in the lineup who have been detected by the eyewitness is equal to $k - a \geq 0$. Thus a may be interpreted as the effective lineup size (refer to Section I: Double high-threshold model). Following from this, if the target is not detected by the eyewitness, the probability that the target will be randomly identified from a target-present lineup of size k may be represented by

$$r = \sum_{a=1}^{k} \frac{s(a)}{a}$$

and the eyewitness identification outcome probabilities are

$$P(\text{TID}) = t + (1-t)gr,$$

$$P(\text{FID}_{\text{TP}}) = g(1-t)(1-r), \text{and}$$

$$P(\text{MISS}) = (1-t)(1-g).$$

The outcome probabilities for TA lineups are:

$$P(\mathrm{SID}) = (1 - f^k)g(1/k),$$

$$P(\mathrm{FID_{TA}}) = (1 - f^k)g(1 - 1/k), \text{and}$$

$$P(\mathrm{CR}) = f^k + (1 - f^k)(1 - g).$$

Appendix 2: Critical tests

A critical test for random scale representation hypothesis

When an eyewitness performs an identification task, they are presented with a set of images, or a *lineup*, depicting one or more persons who fit the description of a previously seen perpetrator. It is postulated that eyewitness then collects evidence-information from each member of the lineup, generating unique subjective values of evidence for each image using a random process. The stronger the relative value generated by an image, the more likely it emanated from a guilty suspect, and the more likely the eyewitness will correctly identify them. Under this assumption of a *random-scale representation* of the lineup member's subjective values, decision accuracy will depend on (1) the relative strengths of these values, (2) the number of images presented within the lineup, and (3) the eyewitness's bias towards identifying someone, or not.

Axiomatic testing for random-scale and monotonic likelihood representation of eyewitness identification data requires the collection of forced-choice probabilities from a series of target-present lineups. These forced-choice lineup identification tasks may be defined by the joint distribution of subjective values of familiarity, as generated by each individual lineup member. However, because this definition is impractical for use in general eyewitness identification studies, where only target discriminability is of interest, it is assumed that, because fillers are innocent and have never been seen by the eyewitness, they

will be equally unfamiliar. That is all fillers within a lineup are assumed to be identically distributed and thus *exchangeable*.[19] This means that when an eyewitness generates a latent-value for each filler $\{y_1, \ldots, y_{n-1}\}$ within a lineup of size n, it is assumed they will do so from identical 'filler' distributions. This very much simplifies any mathematical model representing eyewitness memory, as well as the testing of the core assumptions that underlie them.

In addition to the equivalence constraint, there exists a number of core assumptions, or axioms, underpinning random-scale distributions, each of which impose a different level of constraint on the outcome space—*above-chance target selection*, *regularity*, and the *Block-Marschak inequalities*—and are discussed in turn below.

Above-chance target selection

The first core assumption of random-scale representation is the axiom of *above-chance target selection*, which states that, given that an eyewitness has only has memory of the perpetrator, when forced to choose someone from a target-present lineup (i.e., a lineup that includes the perpetrator), it is expected that the eyewitness will prefer the target, and will thus select the target at a rate greater than chance. The assumption of above-chance target selection states that, given a target-present lineup L, the probability that an eyewitness correctly selects the target t from a lineup L, denoted $P^{\langle |L| \rangle}(t)$, is

$$P^{\langle |L| \rangle}(t) \geq \frac{1}{|L|},$$

[19] Exchangeability of fillers may be observed in lineups for whose functional size is calculated to equal its physical size. This is a general expectation for all well-constructed lineups consisting of fillers who each match the eyewitness's description of the culprit.

where |L| is the cardinality (i.e., lineup size) of *L*. For example, if we present an eyewitness with a 6-item target-present lineup. If forced to choose someone, the probability that the eyewitness will correctly select the target will be

$$P^{\langle 6\rangle}(t) \geq \frac{1}{6}.$$

Now, this provides a relatively weak test of random-scale representation. This is simply because the constraint it imposes leaves a relatively large outcome space within which the constraint is predicted. It is possible to strengthen the test by assessing a family of target-present lineup subsets of L that contain at least one filler, denoted $\mathrm{F}\left(L,i\right)$, where $i \in \left\{2,3,\ldots,|L|\right\}$. For example, target-selection rates from five target-present lineups of ascending sizes—i.e., from 2- to 6-item lineups sizes—may be tested against the assumption of above-chance selection, such that

$$P^{\langle 2\rangle}(t) \geq \frac{1}{2},$$
$$P^{\langle 3\rangle}(t) \geq \frac{1}{3},$$
$$P^{\langle 4\rangle}(t) \geq \frac{1}{4},$$
$$P^{\langle 5\rangle}(t) \geq \frac{1}{5}, \text{and}$$
$$P^{\langle 6\rangle}(t) \geq \frac{1}{6}.$$

However, while this does improve things this test of random-scale representation remains a weak one.

Regularity

The second, and stronger axiom is that of regularity. This simply postulates that the rate at which an eyewitness will correctly select a previously seen perpetrator from a lineup will decrease as the size of the lineup increases. For example, if an eyewitness was presented with

the lineup $|L|=6$, then as per the assumption of above-chance selection, the expected rate of choosing the target is

$$P^{\langle 6\rangle}(t) \geq \frac{1}{6}.$$

However, if the eyewitness was instead presented with the same lineup from which one of the fillers has been removed, then the new expected rate of target selection will increase to become

$$P^{\langle 5\rangle}(t) \geq \frac{1}{5}.$$

Thus, the axiom of regularity asserts that

$$P^{\langle |L|-1\rangle}(t) \geq P^{\langle |L|\rangle}(t)$$
$$P^{|L|-1}(t) - P^{|L|}(t) \geq 0$$

is true.

As per the case with above-chance target selection, we may strengthen this test of random-scale representation by assessing each subset dyad within the family $\mathrm{F}(L,i)$, such that, and in our example case of $|L|=6$,

$$P^{\langle 2\rangle}(t) - P^{\langle 3\rangle}(t) \geq 0,$$
$$P^{\langle 3\rangle}(t) - P^{\langle 4\rangle}(t) \geq 0,$$
$$P^{\langle 4\rangle}(t) - P^{\langle 5\rangle}(t) \geq 0, \text{and}$$
$$P^{\langle 5\rangle}(t) - P^{\langle 6\rangle}(t) \geq 0.$$

Unfortunately, while this is a stronger than the one provided by the axiom of above-chance target selection, it is not a strict one.

Block-Marschak inequalities

A full exploration of critical tests of random-scale representation was initiated by (Block & Marschak, 1959), who demonstrated that forced-choice probabilities based on a random-scale representation must satisfy a set of inequalities. However, it was not until a proof for the converse—choice probabilities which satisfy the Block-Marschak inequalities (BMIs) are based on a random utility representation (Falmagne, 1978)—that this set of choice probabilities could be used to test the assumption of random-scale representation. This work was further transitioned to the context of a k – alternative forced choice (*k*-AFC) task by Kellen, Winiger, Dunn, and Singmann (2021), who used the resulting system of BMIs to critically test recognition memory for words. By aggregating choice data from 110 participants, which produced approximately $1{,}000$ observations for each of seven *k*-AFC condition groups, where $k \in \{2, 3, 4, 5, 6, 7, 8\}$ these researchers found that the resulting data provided near-perfect fits to data predicted by the BMI constraints.

The reason the BMIs require data that has been collected from lineups of differing cardinalities (i.e., number of lineup members) is because the BMI test considers regularity beyond simple probability dyads. Instead, BMI also includes the comparison of ordinal probability triads, tetrads, pentads, … etc. by considering all the target-present subsets

$$D \subseteq S_i \subseteq \mathrm{F}\ (L, k)$$

such that, for each subset S_i, there exists a BMI defined by

$$\sum\nolimits_{\mathrm{S}:D\subseteq S_i} (-1)^{|S_i,\ D|}\, P^{|D|}(t) \geq 0,$$

where $\mathrm{S}: D \subseteq S_i$ is the set of target-present subsets of S_i, $\left|S_i \setminus D\right|$ is the cardinality of the subset S_i minus subset D, and $P^{\langle |D| \rangle}(t)$ is the probability of selecting t from subset D.[20]

Let us again imagine we have a 6-item target-present lineup

$$L = \{t, f_1, f_2, f_3, f_4, f_5\}$$

Firstly, there are $2^5 - 1 = 31$ subsets $S \subseteq L$ that include the target and *at least one filler*. If we then consider one of these target-present lineup subsets

$$L = \{t, f_1, f_2\},$$

then there is a set of target-present subsets

$$\mathrm{S}: D \subseteq S_i = \{D_1, D_2, D_3, D_4\},$$

where

$$\begin{aligned} D_1 &= \{t\}, \\ D_2 &= \{t, f_1\}, \\ D_3 &= \{t, f_2\}, \text{and} \\ D_4 &= \{t, f_1, f_2\}. \end{aligned}$$

According to the definition of the BMI for random-scale representation to hold, the following inequality must be met:

$$\begin{aligned} (-1)^{4-2} P^{\langle 2 \rangle}(t) + (-1)^{4-3} P^{\langle 3 \rangle}(t) + (-1)^{4-3} P^{\langle 3 \rangle}(t) + (-1)^{4-4} P^{\langle 4 \rangle}(t) &\geq 0, \\ P^{\langle 2 \rangle}(t) - P^{\langle 3 \rangle}(t) - P^{\langle 3 \rangle}(t) + P^{\langle 4 \rangle}(t) &\geq 0, \text{and} \\ P^{\langle 2 \rangle}(t) - 2P^{\langle 3 \rangle}(t) + P^{\langle 4 \rangle}(t) &\geq 0. \end{aligned}$$

Of course, for random-scale representation to hold across the whole data set, it is required that all the BMIs for all $S \subseteq L$ that include the target and at least one filler, are held

[20] Note that while subset S_i contains the target and at least one filler, subsets *D* must contain the target and may contain any number of fillers between zero and $\left|S_i\right| - 1$.

true. Thus, there exists a set of inequalities which arrange choice-probabilities such that they 'cancel each other out'. Expanding this to all triads across the lineup, the set of BMIs include

$$P^{\langle 2\rangle}(t) - 2P^{\langle 3\rangle}(t) + P^{\langle 4\rangle}(t) \geq 0,$$
$$P^{\langle 3\rangle}(t) - 2P^{\langle 4\rangle}(t) + P^{\langle 5\rangle}(t) \geq 0, \text{ and}$$
$$P^{\langle 4\rangle}(t) - 2P^{\langle 5\rangle}(t) + P^{\langle 6\rangle}(t) \geq 0.$$

Subsequently, if we then consider all tetrads, the expected inequalities are:

$$P^{\langle 2\rangle}(t) - 3P^{\langle 3\rangle}(t) + 3P^{\langle 4\rangle}(t) - P^{\langle 5\rangle}(t) \geq 0, \text{and}$$
$$P^{\langle 3\rangle}(t) - 3P^{\langle 4\rangle}(t) + 3P^{\langle 5\rangle}(t) - P^{\langle 6\rangle}(t) \geq 0.$$

And finally, comparing the distribution of all subsets of L:

$$P^{\langle 2\rangle}(t) - 4P^{\langle 3\rangle}(t) + 6P^{\langle 4\rangle}(t) - 4P^{\langle \infty\rangle}(t) + P^{\langle 6\rangle}(t) \geq 0$$

That is a total of fifteen different inequalities which the observed data is expected to meet if it is to be considered to have random-scale representation, and results in a strong test of random scale representation.

A critical test for the monotonic likelihood hypothesis

Imagine now that the eyewitness is instead required to rank all lineup members from the most likely to be the target to the least likely. The regularity assumption will also translate to subsequent target ranking probabilities, in that the target is expected to be ranked in the first position at a greater rate than they are ranked second. Likewise, the target is also expected to be ranked second more often than they are expected to be ranked third, and so on. Thus, if presented with a lineup L, the probability that the target will be ranked in position i, denoted as $R_i^{\langle |L|\rangle}(t)$, where i exists within the set $\{1, 2, \ldots, |L|\}$. The subsequent target ranking probabilities will be ordered as follows:

$$R_1^{\langle |L|\rangle}(t) \geq R_2^{\langle |L|\rangle}(t) \geq \ldots \geq R_{|L|}^{\langle |L|\rangle}(t)$$

and are thus expected to be *monotonically decreasing*.

Such ranking probabilities are simple to calculate if we define accurate ranking choices as those in which the target is ranked first and all the fillers within the lineup being ranked in the second, third, … , and last positions, and then assume a binomial distribution. Let *f* be the probability of accurately ranking a filler in a position below that given to the target and *i* be the ranking position of the target. The probability that the target is ranked *i* among |*L*| alternatives, denoted $R_i^{\langle |L| \rangle}$, is the probability that $|L|-i$ fillers are *accurately ranked* below i and *i* - 1 fillers are *erroneously ranked* higher than *i*. Therefore.,

$$R_i^{\langle |L| \rangle}(t) = \binom{|L|-1}{i-1}(1-f)^{i-1}(f)^{|L|-1}.$$

To demonstrate how this binomial distribution may be expressed in terms of forced-choice probabilities, we first consider the probability that an eyewitness selects the target from a 6-item target-present lineup. This is equivalent to ranking the target first:

$$\begin{aligned} P^{\langle 6 \rangle}(t) &\equiv R_1^{\langle 6 \rangle}(t) \\ &= \binom{6-1}{1-1}(1-f)^{1-1}(f)^{6-1} \\ &= \binom{5}{0}(1-f)^{0}(f)^{5} \\ &= f^5. \end{aligned}$$

Following from this, we note the following relationship between the probability target selection in a forced choice task and the probability of accurately ranking all fillers within a lineup in positions below the target:

$$P^{\langle |L| \rangle}(t) \equiv f^{|L|-1}$$

Thus, the probability a target is ranked second may be expressed in forced-choice probabilities in the following way:

$$
\begin{aligned}
R_2^{\langle 6\rangle}(t) &= \binom{6-1}{2-1}(1-f)^{2-1}(f)^{6-2} \\
&= \binom{5}{1}(1-f)^{1}(f)^{4} \\
&= 5\left(f^4 - f^5\right) \\
&= 5\left(P^{\langle 5\rangle}(t) - P^{\langle 6\rangle}(t)\right).
\end{aligned}
$$

Interestingly, this directly corresponds to one of the BMI dyads, scaled by the number of ways in which fillers may be placed in ranks above the target, which in this case is five. Subsequently, the probability that the target is ranked third, given by

$$
\begin{aligned}
R_3^{\langle 6\rangle}(t) &= \binom{6-1}{3-1}(1-f)^{3-1}3 \\
&= 10\left(f^3 - 2f^4 + f^5\right) \\
&= 10\left(P^{\langle 4\rangle}(t) - 2P^{\langle 5\rangle}(t) + P^{\langle \infty\rangle}(t)\right),
\end{aligned}
$$

which corresponds to one of the BMI triads, scaled by ten. Thus, in addition to the BMIs, a system of monotonic likelihood inequalities, as given by regularity of rank positions

$$
\begin{aligned}
&R_1^{\langle 6\rangle}(t) - R_2^{\langle 6\rangle}(t) \geq 0, \\
&R_2^{\langle 6\rangle}(t) - R_3^{\langle 6\rangle}(t) \geq 0, \\
&R_3^{\langle 6\rangle}(t) - R_4^{\langle 6\rangle}(t) \geq 0, \ldots etc.,
\end{aligned}
$$

may be expressed these in terms of their corresponding forced-choice inequalities, including

$$
\begin{aligned}
P^{\langle 6\rangle}(t) - 5\left(P^{\langle 5\rangle}(t) - P^{\langle 6\rangle}(t)\right) &\geq 0, \\
5\left(P^{\langle 5\rangle}(t) - P^{\langle 6\rangle}(t)\right) - 10\left(P^{\langle 4\rangle}(t) - 2P^{\langle 5\rangle}(t) + P^{\langle \infty\rangle}(t)\right) &\geq 0, \\
10\left(P^{\langle 4\rangle}(t) - 2P^{\langle 5\rangle}(t) + P^{\langle 6\rangle}(t)\right) - 10\left(P^{\langle 3\rangle}(t) - 3P^{\langle 4\rangle}(t) + 3P^{\langle 5\rangle}(t) - P^{\langle 6\rangle}(t)\right) &\geq 0, \ldots etc.,
\end{aligned}
$$

allowing for a further constraint on forced-choice data, such that only monotonically decreasing likelihood solutions will be accorded.

Predicted hazards functions[21]

Given that the ranking probability distribution is a discrete distribution

$$R(x) \in \{1, \ldots, k\},$$

where k is the lineup size, its hazard function is a discrete distribution defined by

$$h(x) = \frac{R(x)}{\sum_{i=x}^{k} R(i)}$$

High-threshold model prediction

Under the high-threshold models of eyewitness memory, a guilty suspect (the target) will always be assigned to rank 1 if it is detected by the eyewitness. If it is not detected, then it will be randomly assigned a rank between 1 and $m + 1$, where m is the number of non-rejected fillers, that is, with equal probability. Because the ranking task does not allow for target rejections, $m + 1$ may be viewed as the effective lineup size a, and hence

$$R(x) = \begin{cases} t + (1-t)\,p(x), & x = 1 \\ (1-t)\,p(x), & x > 1 \end{cases},$$

where $p(x)$ is the probability that the target is assigned rank x given that it is not detected.

To calculate $p(x)$, it is necessary to first calculate the set of effective lineup size probabilities $s(a)$ for each $a \in \{x, \ldots, k\}$, where k is the total number of lineup members (see Equation 1 on page 83). It then follows that, for each $a \in \{x, \ldots, k\}$, the probability the target will be selected from the remaining available lineup members is $s(a)/a$, and thus

$$p(x) = \sum_{a=x}^{k} \frac{s(a)}{a}.$$

[21] Equations given within this section are based on those provided to the researchers by J. C. Dunn, personal communication, (2020).

Signal detection theory model prediction

In contrast, signal detection theory (SDT) models suggest that the target will only be assigned rank $x \in \{1, \ldots, k\}$ if its perceived strength of familiarity is weaker than k - 1 fillers and stronger than the remaining k - x fillers. Let $F(\cdot)$ be the standard normal cumulative density functions (respectively) of subjective non-target stimulus values , and $t(\cdot)$ be the normal probability distribution of subjective target stimulus values. Considering all possible combinations of filler strengths, the probability that the target is assigned to rank $x \in \{1, \ldots, k\}$ is

$$R(x) = \binom{k-1}{x-1} \int_{-\infty}^{+\infty} t(z) \left(1 - F(z)\right)^{x-1} F(z)^{k-1} dz.$$

Appendix 3: References

Akan, M., Robinson, M. M., Mickes, L., Wixted, J. T., & Benjamin, A. S. (2020). The effect of lineup size on eyewitness identification. *Journal of Experimental Psychology: Applied, 27*(2), 369–392. https://doi.org/10.1037/xap0000340

Amendola, K., & Wixted, J. T. (2015). Comparing the diagnostic accuracy of suspect identifications made by actual eyewitnesses from simultaneous and sequential lineups in a randomized field trial. *Journal of Experimental Criminology, 11*(2), 263-284. https://doi.org/10.1007/s11292-014-9219-2

Amendola, K., & Wixted, J. T. (2015). No possibility of a selection bias, but direct evidence of a simultaneous superiority effect: a reply to Wells et al. *Journal of Experimental Criminology, 11*(2), 291-294. https://doi.org/10.1007/s11292-015-9227-x

Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(1), 197-215. https://doi.org/10.1037/0278-7393.22.1.197

Berkowitz, S. R., Garrett, B. L., Fenn, K. M., & Loftus, E. F. (2022). Eyewitness confidence may not be ready for the courts: A reply to Wixted et al. *Memory, 30*(1), 75-76. https://doi.org/10.1080/09658211.2021.1952271

Bjork, R. A. (1973). Why mathematical models? *American Psychologist, 28*(5), 426-433. https://doi.org/10.1037/h0034623

Blackwell, H. R. (1953). Psychophysical thresholds: Experimental studies of methods of measurement. *University of Michigan Engineering Research Institute Bulletin, 36*, 227. Retrieved from https://babel.hathitrust.org/cgi/pt?id=mdp.49015000638156&view=1up&seq=10

Block, H. D., & Marschak, J. (1959). Random orderings and stochastic theories of response. *Cowles Foundation Discussion Paper, 66*. Retrieved from https://elischolar.library.yale.edu/cgi/viewcontent.cgi?article=1288&context=cowles-discussion-paper-series

Brackmann, N., Sauerland, M., & Otgaar, H. (2019). Developmental trends in lineup performance: Adolescents are more prone to innocent bystander misidentifications than children and adults. *Memory & Cognition, 47*(3), 428–440. https://doi.org/10.3758/s13421-018-0877-6

Behrman, B., & Davey, S. (2001). Eyewitness Identification in Actual Criminal Cases: An Archival Analysis. *Law and Human Behavior*, *25*(5), 475-491. https://doi.org/10.1023/A:1012840831846

Brewer, N., Weber, N., & Guerin, N. (2019). Police lineups of the future? *American Psychologist, 75*(1), 76–91. https://doi.org/10.1037/amp0000465

British Home Department. (1929). *The report of the Royal Commission on police powers and procedure*. London, UK: HathiTrust Digital Library Retrieved from https://babel.hathitrust.org/cgi/pt?id=uc1.b3113320&view=1up&seq=59&q1=identification

Carlson, C. A. (2019). Lineup fairness: Propitious heterogeneity and the diagnostic feature-detection hypothesis. *Cognitive Research: Principles and Implications., 4*(1), 2. https://doi.org/10.1186/s41235-019-0172-5

Chechile, R. A., & Dunn, J. C. (2021). Critical tests of the two high-threshold model of recognition via analyses of hazard functions. *Journal of Mathematical Psychology, 105*, 102600. https://doi.org/10.1016/j.jmp.2021.102600

Clark, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology, 17*(6), 629-654. https://doi.org/10.1002/acp.891

Clark, S. E. (2008). The importance (necessity) of computational modelling for eyewitness identification research. *Applied Cognitive Psychology, 22*(6), 803-813. https://doi.org/10.1002/acp.1484

Clark, S. E. (2012). Costs and benefits of eyewitness identification reform. *Perspectives on Psychological Science, 7*(3), 238-259. https://doi.org/10.1177/1745691612439584

Clark, S. E. (2012). Eyewitness identification reform: Data, theory, and due process. *Perspectives on Psychological Science, 7*(3), 279-283. https://doi.org/10.1177/1745691612444136

Clark, S. E., Erickson, M., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior, 35*(5), 364-380. https://doi.org/10.1007/s10979-010-9245-1

Colloff, M. F., Flowe, H. D., & Smith, H. M. J. (2021). Active exploration of faces in police lineups increases discrimination accuracy. *American Psychologist, 77*(2), 196–220. https://doi.org/10.1037/amp0000832

Colloff, M. F., Flowe, H. D., Smith, H. M. J., Seale-Carlisle, T. M., Meissner, C. A., Rockey, J. C., Pande, B., Kujur, P., Parveen, N., Chandel, P., Singh, M. M., Pradhan, S., & Parganiha, A. (2022). Active exploration of faces in police lineups increases discrimination accuracy. *American Psychologist, 77*(2), 196-220. https://doi.org/10.1037/amp0000832

Colloff, M. F., Seale-Carlisle, T., Karoğlu, N., & Rockey, J. C. (2021). Perpetrator pose reinstatement during a lineup test increases discrimination accuracy. *Scientific reports, 11*, 13830. https://doi.org/10.1038/s41598-021-92509-0

Colloff, M. F., & Wixted, J. T. (2020). Why are lineups better than showups? A test of the filler siphoning and enhanced discriminability accounts. *Journal of Experimental Psychology, 26*(1), 124-143. http://dx.doi.org/10.1037/xap0000218

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language, 55*(4), 461-478. https://doi.org/10.1016/j.jml.2006.08.003

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior, 4*(4), 243-260. https://doi.org/10.1007/BF01040617

Davis, J. P., Valentine, T., Memon, A., & Roberts, A. J. (2015). Identification on the street: a field comparison of police street identifications and video line-ups in England. *Psychology, Crime & Law*, *21*(1), 9-27. https://doi.org/10.1080/1068316X.2014.915322

Devlin Committee. (1976). Report to the secretary of state for the Home Department of the Departmental Committee on evidence of identification in criminal cases. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/228523/0338.pdf

Duncan, M. (2006). *A signal detection model of compound decision tasks*. Defense Technical Information Centre. Retrieved from: https://apps.dtic.mil/sti/citations/ADA473015

Dunn, J. C., Kaesler, M., & Semmler, C. (2022). A model of position effects in the sequential lineup. *Journal of Memory and Language, 122*, 104297. https://doi.org/10.1016/j.jml.2021.104297

Egan, J. P. (1958). Recognition memory and the operating characteristic. *USAF Operational Applications Laboratory Technical Note, 58-51,* ii, 32.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. New York: Academic Press.

Falmagne, J.-C. (1978). A representation theorem for finite random scale systems. *Journal of Mathematical Psychology, 18*(1), 52-72. https://doi.org/10.1016/0022-

2496(78)90048-2

Fechner, G. T. (1860/1966). *Elements of psychophysics*. New York: New York : Holt, Rinehart and Winston.

Fife, D., Perry, C., & Gronlund, S. D. (2014). Revisiting absolute and relative judgments in the WITNESS model. *Psychonomic Bulletin & Review, 21*(2), 479. https://doi.org/10.3758/s13423-013-0493-1

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry, 31*(4), 271-288. https://doi.org/10.1080/1047840X.2020.1853461

Gronlund, S. D., & Neuschatz, J. S. (2014). Eyewitness identification discriminability: ROC analysis versus logistic regression. *Journal of Applied Research in Memory and Cognition, 3*(2), 54-57. https://doi.org/10.1016/j.jarmac.2014.04.008

Higgs, E. (2011). *Identifying the English*. London: Continuum International.

Hintzman, D. L. (1991). Why are formal models useful in psychology? In W.E. Hockley and S. Lewandowsky (Eds) *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock*, pp. 39-56. Lawrence Erlbaum Associates, Inc.

Kaesler, M., Dunn, J. C., Ransom, K., & Semmler, C. (2020). Do sequential lineups impair underlying discriminability? *Cognitive Research: Principles and Implications, 5*(1), 35-35. https://doi.org/10.1186/s41235-020-00234-5

Kellen, D., & Klauer, K. C. (2014). Discrete-state and continuous models of recognition memory: Testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1795-1804. https://doi.org/10.1037/xlm0000016

Kellen, D., Winiger, S., Dunn, J. C., & Singmann, H. (2021). Testing the foundations of signal detection theory in recognition memory. *Psychological Review, 128*(6), 1022-

1050. https://doi.org/10.1037/rev0000288

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Personality and social psychology review, 2*(3), 196-217. https://doi.org/10.1207/s15327957pspr0203_4

Lakatos, I. (1980). *The methodology of scientific research programmes: Volume 1: Philosophical papers*. Cambridge university press.

Lee, J., & Penrod, S. D. (2019). New signal detection theory-based framework for eyewitness performance in lineups. *Law and Human Behavior, 43*(5), 436-454. https://doi.org/10.1037/lhb0000343

Lindsay, R. C. L. (1999). Applying applied research: selling the sequential line- up. *Applied Cognitive Psychology, 13*(3), 219-225. https://doi.org/10.1002/(SICI)1099-0720(199906)13:3<219::AID-ACP562>3.0.CO;2-H

Lindsay, R. C. L., & Wells, G. L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law and Human Behavior, 4*(4), 303-313. https://doi.org/10.1007/BF01040622

Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70*(3), 556-564. https://doi.org/10.1037/0021-9010.70.3.556

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

Malpass, R. S. (1981). Effective Size and Defendant Bias in Eyewitness Identification Lineups. *Law and Human Behavior, 5*(4), 299-309. https://doi.org/10.1007/BF01044945

Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology, 66*(4), 482-489.

https://doi.org/10.1037/0021-9010.66.4.482

McAdoo, R., & Gronlund, S. (2016). Relative judgment theory and the mediation of facial recognition: Implications for theories of eyewitness identification. *Cognitive Research, 1*(1). https://doi.org/10.1186/s41235-016-0014-7

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition, 4*(2), 93-102. https://doi.org/10.1016/j.jarmac.2015.01.003

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied, 18*(4), 361-376. https://doi.org/10.1037/a0030609

Mickes, L., Seale-Carlisle, T. M., Chen, X., & Boogert, S. (2022, March 9). *pyWitness 1.0: A Python eyewitness identification analysis toolkit*: psyarxiv.com.

Moran, R. (2016). Thou shalt identify! The identifiability of two high-threshold models in confidence-rating recognition (and super-recognition) paradigms. *Journal of Mathematical Psychology*, 73, 1-11. https://doi.org/org/10.1016/j.jmp.2016.03.002

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour, 3*(3), 221-229. https://doi.org/10.1038/s41562-018-0522-1

Newirth, K. A. (2016). An eye for the science: Evolving judicial treatment of eyewitness identification evidence. *Journal of Applied Research in Memory and Cognition, 5*(3), 314-317. https://doi.org/10.1016/j.jarmac.2016.06.009

Noorian, F. (2015). quadprogpp-package: Quadratic Programming++ for R (Version 1.1-0) [R package]. https://rdrr.io/github/fnoorian/quadprogpp/man/quadprogpp-package.html

Penrod, S. (2006). *Eyewitness guessing and choosing*. Paper presented at the American

Psychology-Law Society, St. Petersburg, FL.

Penrod, S., Garcia, L., & Robertson, R. (2005). *Assessing the Impact of Eyewitness Guessing and Lineup Bias on Eyewitness Performance.* Paper presented at the 29th International Congress on Law and Mental Health, Paris, France.

Police Executive Research Forum. (2013). *A national survey of eyewitness identification procedures for law enforcement agencies*. Retrieved from www.policeforum.org

R Core Team. (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Revelle, W. (2022). psych: Procedures for Psychological, Psychometric, and Personality Research (Version 2.1.9) [R package]. Retrieved from https://CRAN.R-project.org/package=psych

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Muller, M. (2011). pROC: An opensource package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*(77), 1-8. https://doi.org/10.1186/1471-2105-12-77

Robinaugh, D., Haslbeck, J., Ryan, O., Fried, E. I., & Waldorp, L. (2020). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science, 16*(4), 725-743. https://doi.org/10.1177/1745691620974697

Saraiva, R. B., van Boeijen, I., Hope, L., Horselenberg, R., Sauerland, M., & van Koppen, P. J. (2020). Eyewitness metamemory predicts identification performance in biased and unbiased line-ups. *Legal and Criminological Psychology, 25*(2), 111-132. https://doi.org/10.1111/lcrp.12166

Sauer, J. D., Palmer, M. A., & Brewer, N. (2019). Pitfalls in using eyewitness confidence to diagnose the accuracy of an individual identification decision. *Psychology, Public Policy, and Law, 25*(3), 147-165. https://doi.org/10.1037/law0000203

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

Sen, A. K. (1986). Prediction and economic theory. Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences, 407(1832), 3-23. https://www.jstor.org/stable/pdf/2397778.pdf

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics, 21*(20). Retrieved from http://rocr.bioinf.mpi-sb.mpg.de

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of Multinomial Processing Tree models with R (Version 1.4.1) [R package]. *Behaviour Research Methods, 45*(2), 560-575. https://doi.org/10.3758/s13428-012-0259-0

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society open science, 3*(9), 160384. https://doi.org/10.1098/rsos.160384

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*(1), 34. https://doi.org/10.1037//0096-3445.117.1.34

Sporer, S. L. (2008). Lessons from the origins of eyewitness testimony research in Europe. *Applied Cognitive Psychology, 22*(6), 737-757. https://doi.org/10.1002/acp.1479

Steblay, N., 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. https://doi.org/10.1037/a0021650

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68*(5), 301-340. https://doi.org/10.1037/h0040547

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection.

Psychological Review, 61(6), 401-409. https://doi.org/10.1037/h0058700

Terrell, J., Baggett, A., Dasse, M., & Malavanti, K. (2017). A hybridization of simultaneous and sequential lineups reveals diagnostic features of both traditional procedures. *Applied Psychology in Criminal Justice, 13*(1), 96-108. http://dev.cjcenter.org/_files/apcj/APCJ%20SPRING%202017-terrell.pdf_1495141277.pdf

Thurstone, L. L. (1927/1994). A law of comparative judgment. *The American Journal of Psychology, 100*(3/4), 587-609. https://doi.org/10.2307/1422696

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48*(1), 28-50. https://doi.org/10.1016/j.jmp.2003.11.004

Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology, 14*(2), 89-103. https://doi.org/10.1111/j.1559-1816.1984.tb02223.x

Wells, G. L. (1993). What Do We Know About Eyewitness Identification? *American Psychologist, 48*(5), 553-571. https://doi.org/10.1037/0003-066X.48.5.553

Wells, G. L. (2001). Eyewitness lineups: Data, theory, and policy. *Psychology, Public Policy, and Law, 7*, 791-801. https://doi.org/10.1037/1076-8971.7.4.791

Wells, G. L., Smalarz, L., & Smith, A. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition, 4*(4), 313-317. https://doi.org/10.1016/j.jarmac.2015.08.008

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2012). Eyewitness identification reforms: Are suggestiveness- induced hits and guesses true hits? *Perspectives on Psychological Science, 7*(3), 264-271. https://doi.org/10.1177/1745691612443368

Wetmore, S. A., McAdoo, R. M., Gronlund, S. D., & Neuschatz, J. S. (2017). The impact of fillers on lineup performance. *Cognitive Research: Principles and Implications, 2*(1),

48. https://doi.org/10.1186/s41235-017-0084-1

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. New York: Oxford University Press.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., . . . Yutani, H. (2019). Welcome to the {tidyverse}. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., Hester, J., & Bryan, J. (2022). readr: Read Rectangular Text Data (Version 2.1.2) [R package]. Retrieved from https://CRAN.R-project.org/package=readr

Wilson, B. M., Donnelly, K., Christenfeld, N., & Wixted, J. T. (2019). Making sense of sequential lineups: An experimental and theoretical analysis of position effects. *Journal of Memory and Language, 104*, 108. https://doi.org/10.1016/j.jml.2018.10.002

Winsor, A. A., Flowe, H. D., Seale-Carlisle, T. M., Killeen, I. M., Hett, D., Jores, T., . . . Colloff, M. F. (2021). Child witness expressions of certainty are informative. *Journal of Experimental Psychology: General, 150*(11), 2387-2407. https://doi.org/10.1037/xge0001049

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

Wixted, J. T., Gronlund, S. D., & Mickes, L. (2014). Policy regarding the sequential lineup is not informed by probative value but is informed by receiver operating characteristic analysis. *Current Directions in Psychological Science, 23*(1), 17-18. https://doi.org/10.1177/0963721413510934

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection

model of eyewitness identification. *Psychological Review, 121*(2), 262-276. https://doi.org/10.1037/a0035940

Wixted, J. T., Mickes, L., Wetmore, S. A., Gronlund, S. D., & Neuschatz, J. S. (2017). ROC analysis in theory and practice. *Journal of Applied Research in Memory and Cognition, 6*(3), 343-351. https://doi.org/10.1016/j.jarmac.2016.12.002

Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. M. (2018). Models of lineup memory. *Cognitive Psychology, 105*, 81–114. https://doi.org/10.1016/j.cogpsych.2018.06.001

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

Wooten, A. R., Carlson, C. A., Lockamyeir, R. F., Carlson, M. A., Jones, A. R., Dias, J. L., & Hemby, J. A. (2020). The number of fillers may not matter as long as they all match the description: The effect of simultaneous lineup size on eyewitness identification. *Applied Cognitive Psychology, 34*(3), 590-604. https://doi.org/10.1002/acp.3644

Yang, Y., Smalarz, L., Moody, S. A., Cabell, J. J., & Copp, C. J. (2019). An expected cost model of eyewitness identification. *Law and Human Behavior, 43*(3), 205. https://doi.org/10.1037/lhb0000331

Yarmey, A., & Yarmey, M. (1996). Accuracy of eyewitness identifications in showups and lineups. *Law and Human Behavior, 20*(4), 459-477. https://doi.org/10.1007/BF01498981

Yates, S. Q. (2017). Eyewitness identification: Procedures for conducting photo arrays. Retrieved from https://www.justice.gov/file/923201/download