Applying Signal Detection Models to Investigate the Effect of Sequential Item Presentation on the Police Lineup Task

Matthew Philip Kaesler

University of Adelaide

School of Psychology

August 2021

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Table of Contents

Abstract

Much research investigating the police lineup has argued that presenting items sequentially is superior to presenting them simultaneously, because sequential presentation reduces rates of innocent suspect identification, minimising the chance of false conviction. However, the research program that arrived at this conclusion was resolutely applied in focus, directing less attention to developing theories that might explain how sequential item presentation achieves this outcome.

Recent research has addressed this issue by applying signal detection theory to understanding the lineup task. This mathematical modelling framework characterises observed performance on a recognition memory test, such as the police lineup, as resulting from two latent variables; discriminability, the ability to distinguish target items (guilty suspects) from non-target items, and response bias, conceptualised as willingness to choose. Research employing this framework suggests that sequential item presentation achieve its reduction in innocent suspect identifications by encouraging witnesses to choose less readily than simultaneous presentation, rather than by increasing discriminability. Some studies also find that discriminability is greater for simultaneous presentation and that, on this basis, it should be preferred. However, this body of recent research has employed analysis techniques that fail to capture the unique constraints of sequentially presented lineup tasks. This may compromise the measurement of discriminability (and response bias), leading to incorrect conclusions when comparing sequentially presented lineup tasks to the simultaneous lineup.

This thesis addresses this limitation by developing signal detection models that capture the structural constraints of sequentially presented lineup tasks. These models are used in studies one and two to compare simultaneous lineup presentation to two sequentially presented lineup tasks, one on which identification of the current item terminates the task (stopping rule) and another on which two full laps of the items is completed before an

identification decision is made, as used in the United Kingdom (UK). Study three develops a model for examining changes in discriminability and response bias by serial position in the sequential stopping rule lineup. Each study involved the collection of new experimental data and studies one and three also analysed previously published datasets.

The results of studies one and two imply that sequential item presentation may have a small negative effect on discriminability compared to simultaneous presentation, but this effect was not consistently observed. Effects on response bias were larger and more reliable; the sequential stopping rule lineup was associated with the most conservative overall choosing, followed by the simultaneous lineup, then the UK sequential lineup. In study three, discriminability increased from serial position one to position two in the sequential stopping rule lineup, but not beyond. Changes in response bias by serial position differed depending on whether an identification was made before or after the presentation of the guilty suspect. Taken together, these results imply that there is no compelling reason for policymakers to prefer sequentially presented lineups to the simultaneous lineup. The insights generated from this thesis demonstrate the value of a formal modelling and approach and the need to consider carefully the match between model, task, and research question.

Declaration by Author

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.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also 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.

Matthew Kaesler

Acknowledgements

Thanks to my supervisors, Associate Professor Carolyn Semmler and Professor John Dunn, for supporting my research through their Australian Research Council Discovery Projects grant. This allowed me to conduct the large-sample studies that are necessary for investigating eyewitness memory phenomena and afforded me the opportunity to attend and present at numerous workshops and conferences.

Thanks to Associate Professor Carolyn Semmler for teaching me to think deeply about the assumptions underpinning all aspects of scientific research. I will always be grateful for the supervision meetings that extended far beyond the topic of eyewitness memory and will cherish the memories of our conference travels, particularly the whistle-stop tour of Cape Cod and Boston. Thank you for encouraging me during periods of self-doubt.

Thanks to Professor John Dunn for teaching me about modelling, mathematics and, Mandarin. I have learnt much from your ability to deconstruct a problem into its constituent parts and develop novel methods of investigating it from first principles. Thank you for your patience when you are already a few steps ahead and for always seeking to improve the quality of my work.

Thanks to Dr Kristi Urry and Dr Jessica O'Rielly for accompanying me on my candidature journey, even after you were finished! I am grateful that you listened to and empathised with the travails of the PhD process. Thanks for reassuring me that I was going to make it.

Thanks to those in the eyewitness memory field that willingly gave of their time and experience at various points and with whom I enjoyed some great conversations about psychological science; Professor John Wixted, Professor Laura Mickes. Dr Travis Seale-Carlisle and Dr Matthew Palmer.

Thanks to my wife, Gabrielle for steadfastly supporting me in taking on a project of indeterminate duration and scope and living all the highs and lows with me. Thanks to my parents for supporting me in so many ways. You allowed me to build the foundation on which this thesis rests.

Chapter 1

1.1 Introduction

In the seminal eyewitness memory text "On the Witness Stand", Hugo Munsterberg (1908) sketches a vision of a research program relating the rapidly-increasing knowledge base of experimental psychology to issues concerning the criminal justice system. Munsterberg (1908) proposes that the experimental method should be used to investigate applied problems in their original form, evaluating outcomes according to the aims of the end user. He cautions against addressing applied problems by generalising the results of "undigested" (p. 8) theoretical research. The field of eyewitness memory has taken this advice to heart. Much of the research conducted since a revival in the 1970s has been focused on minimising the possibility of one particular error; a witness to a crime misidentifying an innocent suspect (Clark, 2012a; Gronlund et al., 2015; Wells, 1978). This research program can lay claim to substantial real world impact. It has developed a number of reforms to the procedures for collecting eyewitness evidence that have been subsequently adopted as policy (Clark, 2012a). However, this focus on the "engineering" of lineup procedures to minimise a particular decision outcome has had unintended consequences. The development of most procedural reforms has proceeded without formal theoretical guidance, which has resulted in difficulties explaining how and why changes to lineup procedures affect decision behaviour. Until recently, the field has also neglected the measurement of psychologically relevant latent variables that are used to characterise decision performance in theoretical research on recognition memory (Bornstein & Penrod, 2008). As a result, the effects of the reforms already in use are unclear. In order to achieve its aim of improving the eyewitness evidence collection procedures used in the criminal justice system, the field must address these theoretical shortcomings. It is necessary to improve understanding of how currently adopted reforms affect decision behaviour and proceed with the development of new lineup

procedures from a solid theoretical base. This thesis will outline a measurement modelling approach based in signal detection theory that provides tools for building and testing theories of eyewitness memory. This approach is applied in three studies investigating one of the most contentious lineup reforms, the sequential presentation of lineup items. These studies offer theoretical insight in to what might drive observed differences in decision performance on sequentially-presented lineups compared to simultaneously presented lineups and act as a test of the recently proposed Diagnostic Feature Detection Theory of eyewitness memory (Wixted & Mickes, 2014).

1.2 Measuring Memory

Measurement is the act of quantifying some property of an event or object. The mass of an object is one such property. Through our senses, we can establish that some objects are heavier than others. We can communicate the heaviness of objects by assigning numbers to represent heaviness according to a set of mathematical principles (Mitchell et al., 2017). A kitchen scale maps the observable property of heaviness to an underlying kilogram unit, allowing objects to be compared on this dimension despite spatial or temporal distance. In the case of mass, the property we wish to quantify is directly observable. In psychological science, the properties we wish to quantify are often not directly observable. However, the value of these so called "latent" variables may be inferred from some observable variable. Consider the property of memory strength. People encounter stimuli out in the world, or in the laboratory, and it is assumed that they encode these stimuli to a greater or lesser extent (Mickes et al., 2007; Spanton & Berry, 2020). It is known that the stored representations of these encoded stimuli grow weaker over time, i.e. "forgetting", although the exact mechanisms behind forgetting are still debated (Malmberg et al., 2019). At some point in time after encoding a stimulus, a person may need to draw on its representation in memory. How strong is their memory for the originally encoded stimulus?

It is intuitive to make the inference that greater accuracy in recalling or recognising previously encountered stimuli indicates greater memory strength. However, the accuracy of observable recognition decisions is affected by factors other than memory strength, such as willingness to respond or the format of the memory test. In order to measure memory strength independent of these factors, a measurement model is required that maps participants' observed responses on a particular test procedure to some underlying scale for memory strength. With the aid of such a psychological measurement instrument, it is possible to compare the effect of different memory test procedures on memory strength.

A challenge inherent to measuring latent variables like memory strength is the "problem of coordination" (Kellen et al., 2021a). It is assumed that there is a function that characterises the relationship between an observed variable and the latent variable of interest, e.g. between observed decision accuracy and memory strength. If this function can be derived, then values of the observed variable can be translated in to values of the latent variable. However, in order to derive this function, it is necessary to know all values of the observed variable for all corresponding values of the latent variable. This is not possible without first knowing the function that characterises their relationship, a problematic circularity. In order to proceed, it is necessary to employ a theory that makes sensible assumptions in proposing a mapping between the observed and latent variables. Fortunately, there is a long standing theoretical approach that is well suited to measuring eyewitness memory strength.

1.3 Signal Detection Theory

Signal detection theory (SDT) is a measurement modelling framework for evaluating the performance of a detector under conditions of uncertainty where to-be-detected signal can be easily confused with noise (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002). Recognition memory can be conceptualised as such a detection task (Egan,

1958) on which the to-be-detected "signal" is stimuli previously encountered in the world and the "noise" is stimuli not previously encountered. There is often uncertainty about whether a stimulus was previously encountered, such as when seeing a distant acquaintance across the room at a party. Researchers have long sought to understand the properties of memory by conducting laboratory experiments (Ebbinghaus, 1885/1965; Wundt & Judd, 1897). In a recognition memory experiment employing a "yes/no" test format, participants are first asked to learn a list of stimuli, often words (Egan, 1958). After some delay, they are presented with a series of test trials comprising either a single "old" item that appeared on the study list or a single "new" item that did not, and are asked to indicate whether the item appeared on the study list. That is, participants are tasked with detecting old items. Table 1.1 shows the possible decision outcomes for this test format. There are two possible correct responses, a hit or a correct rejection, and two possible incorrect responses, a miss and a false alarm.

Table 1.1

Decision Outcomes for a Yes/No Recognition Memory Task

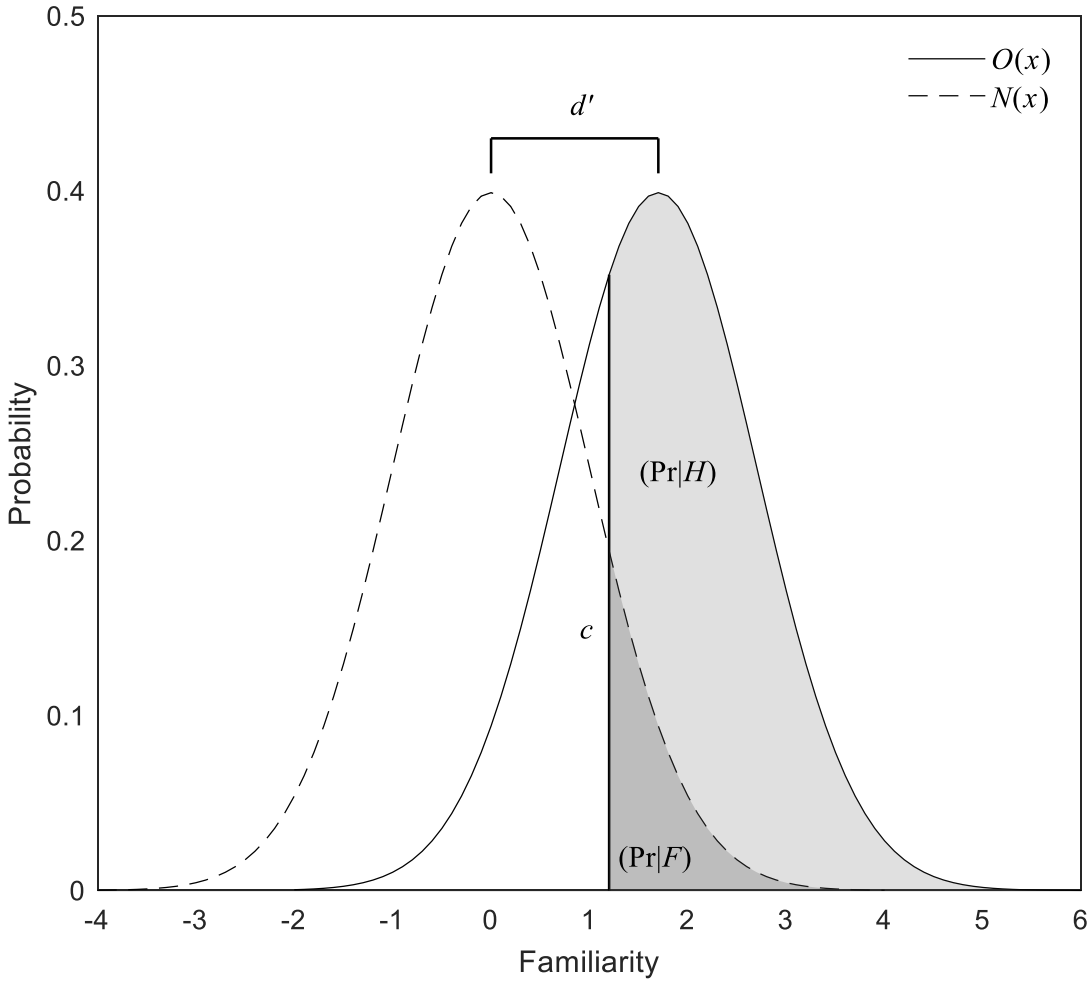
	Participant Decision	
Test Item	Item is Old	Item is New
Item is Old	Hit	Miss
Item is New	False Alarm	Correct Rejection

SDT characterises recognition performance, the correctness of the observed decision outcomes in Table 1.1, as resulting from two unobservable latent variables, discriminability and response bias. Discriminability is the ability of the detector to distinguish signal (old items presented at test) from noise (new items presented at test). In a recognition memory context, discriminability is thought to map on to memory strength. Response bias is the tendency for the detector to indicate the presence of a signal (decide an item is old). The SDT formalism that maps observed responses on a recognition memory test to these latent

variables begins with the assumption that the presentation of a test item elicits some level of familiarity or "signal strength" in participants, modelled as a random draw from a probability distribution. These distributions are usually assumed to be Gaussian in form (Macmillan & Creelman, 2005), which is the case for the models used in this thesis, but other distributions may be used (e.g. DeCarlo, 2010). Figure 1.1 shows an equal variance signal detection model of a yes/no recognition memory task with two Gaussian probability distributions for familiarity.

Figure 1.1

Equal Variance Signal Detection Model with a Single Decision Criterion



The Old distribution, $O(x)$, gives familiarity values for the presentation of to-be-detected old items, and the New distribution, $N(x)$, gives familiarity values for the

presentation of new items. Note that $O(x)$ is positively displaced along the x-axis compared to $N(x)$ because, on average, the presentation of an old item that appeared in the study phase should elicit greater familiarity than a new item.

On each trial, the familiarity for the presented item is drawn from the relevant distribution shown in Figure 1.1. This familiarity value is compared to the bold vertical line labelled "*c*", the decision criterion. The placement of this criterion is the measure of the latent variable response bias. On a single trial of a yes/no recognition memory test, an item that elicits familiarity > c will be classified by participant as having appeared in the study phase ("yes"), while an item that elicits familiarity < c will be classified as not having appeared in the study phase ("no"). The criterion is often measured relative to the midpoint between the means of the signal and noise distributions (the ideal placement). It may also be measured relative to the mean of $N(x)$, which is conventionally fixed to zero (Macmillan & Creelman, 2005). In this case, $c = \Phi^{-1}(1 - F)$, where Φ^{-1} is the inverse normal cumulative distribution function. The probability of a hit (H) is given by the proportion of $O(x)$ that is to the right of c, while the probability of a miss (M) is the proportion of $O(x)$ distribution to the left, that is $(\Pr|M) = 1 - (\Pr|H)$. Similarly, the probability of a false alarm (F) is given by the proportion of $N(x)$ distribution to the right of c and the probability of a correct rejection (R) is the proportion to the left of c, that is $(\Pr|R) = 1 - (\Pr|F)$. It is clear from Figure 1.1 that the placement of the decision criterion involves a trade-off between two different kinds of errors, false alarms and misses. Holding all else equal, false alarms can be avoided by adopting a bias toward conservative responding, moving c rightwards along the x-axis, but this necessarily leads to a decrease in hits. Similarly, misses can be avoid by setting a more lenient response bias, moving c leftwards along the x-axis, which leads to more false alarms.

In the equal variance signal detection model, the latent variable discriminability, labelled *d'* on Figure 1.1, is measured as the distance between the mean of $O(x)$, μ_{old}, and the mean of $N(x)$, μ_{new}, in units of their common standard deviation. It can be calculated by the formula $d' = \Phi^{-1}(H) - \Phi^{-1}(F)$, (Macmillan & Creelman, 2005). This equation makes use of the relationship between H and F and values of μ_{old} and μ_{new} in relation to c. When measuring discriminability with respect to c, i.e. $d' = (\mu_{old} - c) - (\mu_{new} - c)$, the proportion of "yes" responses, i.e. H and F, is related to values of $\mu - c$ via the inverse normal cumulative distribution function. That is, when $H = .5$ and $F = .5$, $\mu_{old} - c = 0$ and $\mu_{old} - c = 0$. When H and F are $> .5$, $\mu_{old} - c$ and $\mu_{new} - c$ are positive, when H and F are $< .5$, $\mu_{old} - c$ and $\mu_{new} - c$ are negative. Holding all else equal, an increase in discriminability, i.e. an increase in the distance between μ_{old} and μ_{new}, will result in an increase in correct responses (hits and correct rejections) and decrease incorrect response (false alarms and misses). For the model shown in Figure 1.1, $d' = 1.7$ and $c = 1.2$, both measured relative to the mean of $N(x)$, which is set to zero, in units of the standard deviation, which is set to one.

SDT was first applied to recognition memory by Egan (1958) using a list learning paradigm and a rating task. After learning a word list, participants with individual test items, some old and some new, and asked to rate them on either a five- or seven-point Likert scale, where one was sure that the word is "old", i.e. was part of the test list, and five/seven was sure that the word is "new". Egan (1958) showed that the pattern of participants' data conformed to those predicted by signal detection theory with underlying Gaussian distributions for signal and noise. A large body of subsequent research has shown that signal detection theory provides a parsimonious explanation of recognition memory phenomena (Kellen et al., 2021b; Wixted, 2007), leading to its continued use in present day recognition memory research.

1.4 Applying Signal Detection Theory to Eyewitness Identification Data

There are two main procedures used by police to collect identification evidence, the showup and the lineup. On a showup, a witness to a crime is presented with a single suspect and asked to indicate whether the suspect is the perpetrator of the crime (Neuschatz et al., 2016). This procedure is sometimes used when police have both an eyewitness and have apprehended a suspect, who may or may not be the perpetrator of the crime under investigation, shortly after a crime has been committed (Behrman & Davey, 2001). The showup task is directly analogous to a yes/no recognition memory test. The suspect may either be the perpetrator of the crime (an "old" item) or innocent of the crime (a "new" item). The witness's task is to detect the perpetrator and they may either indicate that the suspect is the perpetrator ("yes") or that they are not ("no"). This means that the equal variance signal detection model shown in Figure 1.1 can be directly applied to measure discriminability and response bias on the showup task.

The police lineup, however, differs in important ways from the yes/no task and showup procedure that are characterised by the equal variance signal detection model. First, multiple items are presented on a police lineup, as opposed to a single item on the yes/no test and show up. A suspect appears in conjunction with a number of known-innocent foils, who are most often chosen to resemble the witness's description of the suspect (Luus & Wells, 1991; Navon, 1992) but may also be selected to resemble the suspect's appearance (Clark et al., 2015b). Police lineups on which the suspect is the perpetrator are referred to here as "target present" and lineups on which the suspect is not the perpetrator are referred to as "target absent". The witness's task is to determine whether the perpetrator – hereafter referred to as the "target" – is present among the test items and, if so, identify the target (Duncan, 2006). Witnesses on the showup task judge whether only a single item is the perpetrator. A witness presented with a lineup may either identify an item or indicate that the no item

matches their memory of the target, termed a rejection. The possible decision outcomes given the guilt of the suspect and the witness's identification decision are displayed in Table 1.2.

Table 1.2

Possible Decision Outcomes for the Police Lineup Task

	Witness Identification Decision		
State of the World	Suspect	Foil	Rejection
Suspect is Guilty	Target Identification	Foil Identification	Miss
Suspect is Innocent	False Identification	Foil Identification	Correct Rejection

When the suspect is guilty, i.e. the suspect is the target, the witness can correctly identify the target (target identification), incorrectly identify a known-innocent foil (foil identification) or incorrectly reject the lineup (miss). If the suspect is innocent, the witness can incorrectly identify the innocent suspect (false identification), incorrectly identify a known innocent foil or correctly reject the lineup (correct rejection). It is evident from Table 1.2 that the police lineup results in a 2 x 3 decision matrix with six possible decision outcomes, in contrast to the 2 x 2 structure of the yes/no task. This means that the equal variance signal detection model cannot be directly applied to measure discriminability and response bias on the police lineup because its assumptions do not reflect the structure of the task. However, signal detection models can be modified or extended to capture the structure of many decision tasks (Macmillan & Creelman, 2005) and discriminability can be compared across tasks that differ in structure when a suitable model is applied to data from each task. Measurement and comparison of discriminability on the yes/no and two-alternative forced choice (2AFC) tasks provides an instructive example.

On a 2AFC recognition task, participants are presented with two items and asked to indicate which one appeared on a previously studied list. Because an old item is always present on 2AFC trials, only two decision outcomes are possible; correctly selecting the old

item, analogous to a hit, or incorrectly selecting the new item, analogous to a false alarm. In contrast, four decision outcomes are possible for the yes/no task. Thus, observed decision performance is reliably greater for the 2AFC task compared to the yes/no task, i.e. it is easier for participants to make the correct response. If the equal variance model calculation $d' = \Phi^{-1}(H) - \Phi^{-1}(F)$ is applied to data from both tasks, it will appear as though discriminability is greater for the 2AFC task. When calculating 2AFC discriminability, it is necessary to account for the fact that participants select from one of two items on each trial as compared to judging a single item on each yes/no trial. When an appropriate signal detection model is applied, 2AFC discriminability is often very similar to yes/no discriminability despite superior observed decision performance for the 2AFC task (Jang et al., 2009; Macmillan & Creelman, 2005; Wixted & Mickes, 2018).

In much the same way, measuring discriminability on a police lineup task therefore requires modifications to the commonly-employed equal variance signal detection model. Additionally, there are important structural differences between the police lineup tasks currently used in real police investigations. This means that models of police lineup tasks are necessary in order to compare them discriminability (and response bias) between these tasks, a primary aim of this thesis. The following section describes the police lineup tasks investigated in this thesis and a general signal detection framework for modelling them. This framework is used to develop models that account for the unique structure of particular police lineup tasks, allowing them to be compared in terms of discriminability (and response bias).

1.5 Police Lineup Tasks

In real police lineups, the true guilt or innocence of the suspect is unknown. This means that it is difficult to estimate discriminability and response bias from field data (Wixted et al., 2016), because the decision outcomes shown in Table 1.2 are not able to be determined. Due to this inherent limitation of field data, the lineup is most often investigated

using laboratory experiments (Wells, 1978), on which the guilt or innocence of the suspect is manipulated as a between subjects factor, in addition to other between subjects conditions. In the experimental version of the lineup task, participants encode an event, usually a simulated crime. After some delay or a distraction task they are presented with a either a target present lineup that is composed of the target from the simulated crime and the appropriate number of foils, or a target absent lineup containing one designated innocent suspect and foils. When there is no designated innocent suspect, the false identification rate can be estimated by dividing the target absent foil identification rate by the lineup size. This is based on the assumption that the identification of each foil on a fair target absent lineup is equiprobable. In experimental research, it is typical for the target absent condition to have no designated innocent suspect. The practice of designating an innocent suspect, particularly one that resembles the target to a greater degree than the other foils, has become less common based on the assumption that innocent suspects in real police lineups should resemble the perpetrator to the same degree as the foils (McQuiston-Surrett et al., 2006).

There are many dimensions along which police lineups can vary. Researchers have investigated the effect of pre-lineup instructions, the similarity of the known-innocent foils to the target and whether or not the lineup administrator knows which item is the suspect (Clark, 2012a), among others. This thesis focuses on the presentation format of the lineup items. Current police lineup procedures present items to the witness either simultaneously or sequentially.

1.5.1 Simultaneous Lineup Presentation

Simultaneous item presentation is the typical method of presenting a lineup. All items are presented to the witness in one array and the witness may either identify an item from the array or reject it. While no jurisdiction for which data is available explicitly recommends simultaneous presentation, procedural guidelines for the majority of jurisdictions describe or

imply conducting a simultaneous lineup (Fitzgerald et al., 2021). The origins of the simultaneous police lineup are difficult to determine, but numerous documented instances of lineups conducted in English case records have been found dating back to the 1850s and 1860s (Bentley, 2003). Around this period, awareness increased in the legal profession that poor quality identifications, such as those made by a witness in the dock at trial or from a face glimpsed in the muzzle flash of a pistol, often led to miscarriages of justice (Wills, 1838/1850). The practice of placing the suspect in an "identification parade" with similar looking fillers was enacted to improve the reliability of identification evidence, protecting the suspect. The beginning of the modern period of lineup research highlighted similar issues, ironically, with the simultaneous lineup itself (Lindsay & Wells, 1985). Documented miscarriages of justice resulting from false identifications prompted researchers investigate alterations to the lineup procedure that would reduce possibility that innocent suspects would be identified (Wells, 1978).

1.5.2 The Sequential Lineup with a Stopping Rule

One procedural alteration proposed to reduce the false identification rate observed on the simultaneous lineup is the sequential presentation of lineup items (Lindsay & Wells, 1985). The rationale for why sequential item presentation would reduce false identifications can be found in the distinction between absolute and relative judgement strategies (Wells, 1984). Sequential item presentation was proposed to encourage witnesses to make absolute comparisons between their memory of the perpetrator and the lineup items, rather than selecting the lineup item that most resembled their memory of the perpetrator relative to the other lineup items. This rationale is expanded in Section 1.8.1.2.

For each sequentially presented item, the witness makes a decision to identify or reject. If the witness identifies the current item, the procedure terminates; they are not shown the remaining lineup members. This is known as the "stopping rule". If the witness rejects the

current item, they are shown the next item. If the witness rejects all items, this is considered an overall lineup rejection. A large body of experimental work shows that the sequential stopping rule lineup procedure reliably achieves its aim of reducing rates of false identification (Steblay et al., 2011b). On this basis, researchers advocated for the adoption of the procedure by police jurisdictions. It is now used in approximately 30% of United States jurisdictions, although it is generally conducted without the stopping rule (Wells et al., 2015a). Sequential item presentation is also recommended in the procedural guidelines for Canada, Germany, Denmark, Norway, the United Kingdom (UK) and Sweden (Fitzgerald et al., 2021).

1.5.3 The Sequential Lineup in the United Kingdom

While the lineup task used in the UK is sequentially presented, it differs substantially from the task described by Lindsay and Wells (1985). The procedure was developed in the mid nineteen-nineties by West Yorkshire Police and was in regular use in that jurisdiction from approximately 1997 (Pike et al., 2002) before being subsequently adopted across the UK. It is unclear whether its development was influenced by eyewitness memory research.

On the UK lineup, the complete sequence of items must be viewed twice before the witness can choose to either identify an item from the lineup or to reject it. Additionally, each item is a 15-second long video of a head and shoulders rotating profile, in contrast to the photographs typically used in simultaneous and stopping-rule sequential lineups. It seems that video-based lineups were advanced largely on the basis that they are easier to construct and administer than live lineups (British Broadcasting Corporation, 2003). The adoption of this video-based procedure required changes to legislation at this time, as live lineups were previously treated as superior to video lineups from a legal standpoint (Pike et al., 2002). This procedure is of theoretical interest because the items are sequentially presented but the single

identification decision is deferred, in contrast to the yes/no decision required for each item on the sequential stopping rule lineup.

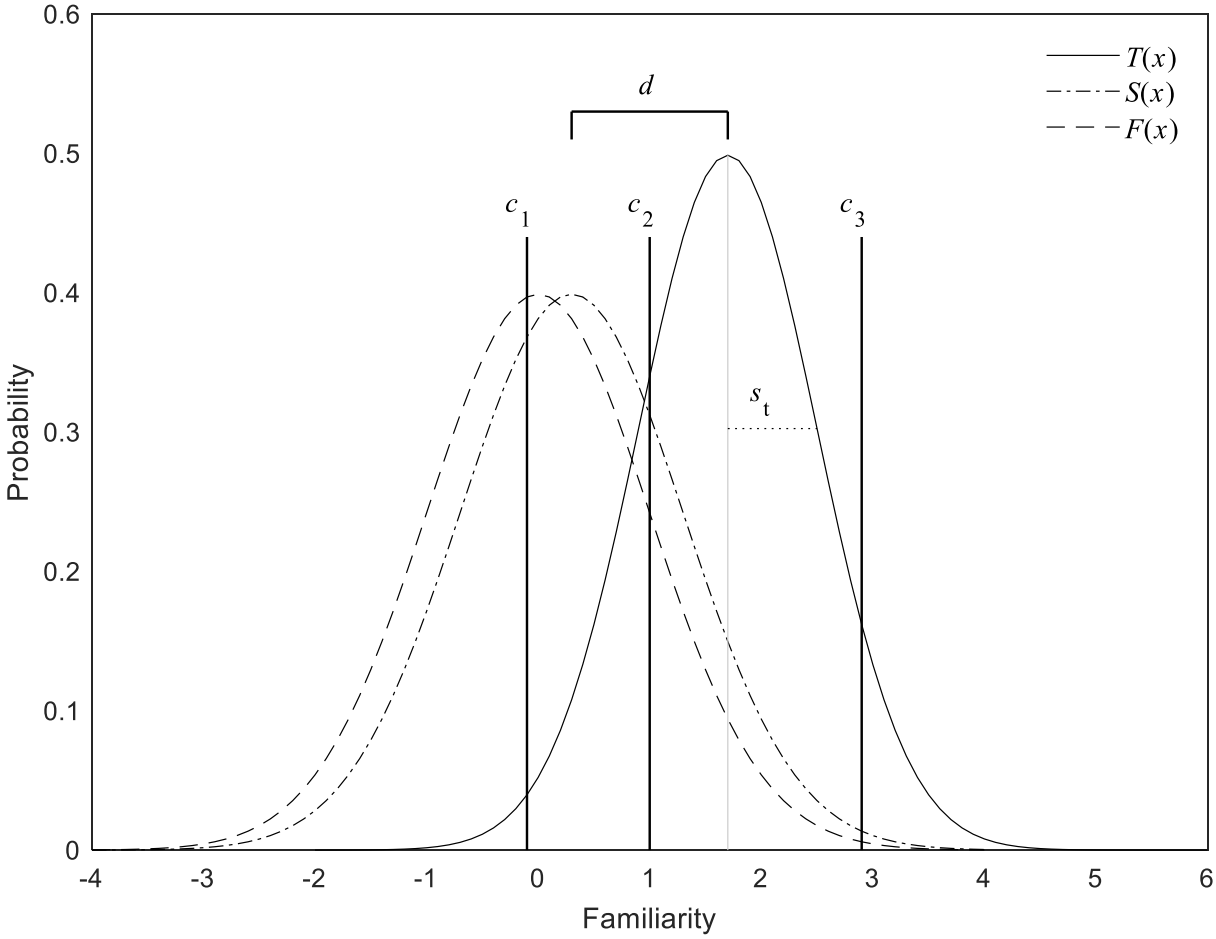
1.6 A Signal Detection Model of the Lineup Task

The starting point for developing signal detection models of the police lineup tasks described in the previous section is to assume that each trial involves the presentation of an array of n items. A target present (TP) lineup contains the perpetrator (target) of a crime and $n - 1$ known-innocent foils. A target absent lineup (TA) is composed of one designated innocent suspect and $n - 1$ foils or, more commonly, n foils. Familiarity for targets is a random draw from a Gaussian distribution, $T(x)$, with mean d_t and standard deviation s_t, analogous to the signal distribution in Figure 1.1. Familiarity for innocent suspects is a random draw from a Gaussian distribution, $S(x)$, with mean d_s and standard deviation s_s. Familiarity for foils is a random drawn from a Gaussian distribution, $F(x)$, with mean d_f and standard deviation s_f. The distributions $F(x)$ and $S(x)$ are analogous to the noise distribution of the equal variance model in Figure 1.1; both model the signal strength elicited by the presentation of non-target items. For experimental designs without a designated innocent suspect on TA lineups, all foils are assumed to be drawn from the same distribution. In this case, $d_s = d_f$ and the model reduces to two distributions, $T(x)$ and $F(x)$. It is expected that $d_t > d_f$ and $d_t > d_s$ because targets encoded when witnessing the event should, on average, elicit greater memory strength for witnesses than foils or innocent suspects. It is assumed, without loss of generality, that $d_f = 0$ and $s_f = 1$. Values for the free model parameters d_t, s_t and $c_1 \ldots c_n$ are estimated relative to d_f and s_f. If there is a designated innocent suspect on TA lineups, d_s is estimated relative to d_f and s_f.

Figure 1.2 shows a signal detection model for the lineup task with a designated innocent suspect and three decision criteria.

Figure 1.2

Unequal Variance Signal Detection Model of the Lineup Task



Unlike the equal variance model shown in Figure 1.1, for this model, $s_t \neq s_f$. This is known as the unequal variance signal detection model and is used in recognition memory research because the assumption that $s_t = s_f$ is known to be violated on recognition memory tasks. For list learning paradigms in basic recognition memory research, it is usual that $s_t > s_f$ (Mickes et al., 2007), which is presumed to reflect variable encoding quality across participants for the items in the study phase. For the lineup task, it has been found that $s_t < s_f$ (Seale-Carlisle et al., 2019; Wixted et al., 2018). The calculation of discriminability, labelled

d in Figure 1.2, differs depending on the experimental design. When there is a designated innocent suspect, it is the ability to discriminate between targets and innocent suspects, i.e. $d = d_t - d_s$. When there is no designated innocent suspect, it is the ability to discriminate between targets and foils, i.e. $d = d_t - d_f = d_t - 0 = d_t$. The other parameter that summarises discriminability for an unequal variance model is s_t. Both d_t and s_t affect the amount of overlap between the target distribution and the innocent suspect or foil distribution. Holding all else equal, an increase in d_t or a decrease in s_t will decrease the amount of overlap between the distributions, leading to fewer decision errors and increasing discriminability.

In contrast to the equal variance model depicted in Figure 1.1, the model shown in Figure 1.2 has multiple decision criteria. The single decision criterion in Figure 1.1 that separates "yes" responses from "no" is analogous to the criterion labelled c_1 in Figure 1.2. This is the threshold that separates identifying an item from rejecting the lineup. Criteria c_2 and c_3 are increasingly strict criteria for classifying an identification. The placement of the decision criteria can be altered by varying the payoff for hits compared to false alarms or by changing the base rate of trials on which a target is presented. This shifts the optimal placement of the choose/no choose threshold. A low base rate of target present trials or large penalty for false identifications requires a more conservative criterion to maximise returns, while a high base rate of target present trials or a large bonus for hits requires a more lenient criterion. In this case, the criteria shown on Figure 1.2 would represent response bias calculated from separate experimental conditions. In much recognition memory research, the decision criteria reflect post-decision confidence ratings, based on the assumption that confidence judgments are scaled directly from the memory strength or familiarity generated by a test item (Stretch & Wixted, 1998). Post-decision confidence judgements are often collected for lineup identifications on the basis that they provide insight into the reliability of witness decisions. A large body of research shows that confidence and accuracy are

positively related for identification decisions, particularly at high levels of confidence (Wixted & Wells, 2017). This relationship does not appear to hold for rejection decisions (Brewer, 2006) but this is a less pressing applied problem because it is less likely that the suspect in a rejected lineup to be charged with a crime. When assessing the reliability of identification evidence, it is advisable to rely on the post-decision confidence estimates collected immediately after the initial identification decision, because subsequent estimates elicited from witnesses are liable to inflation (Wixted & Wells, 2017).

The lowest confidence rating, labelled c_1 in Figure 1.2, is the choose/no choose threshold. This boundary separates the response of identifying an item from rejecting the lineup. This is analogous to the single criterion, c, in Figure 1.1. Criteria c_2 and c_3 represent thresholds for classifying an identification with greater levels of confidence. The number of criteria in a model is arbitrary. In some studies, the experimental design constrains the number of criteria, such as when confidence is collected on a Likert-type scale (e.g. Mickes et al., 2012). In other instances, confidence is collected on a fine grained scale, perhaps 0 – 100, and the number of criteria and the width of each confidence category is set such that cell counts are as even as possible between categories (e.g. Seale-Carlisle et al., 2019). This method is employed where possible in this thesis. In general, adding criteria to a given model will improve its ability to account for observed data. However, there is a point beyond which adding additional criteria will not benefit the model because the number of criteria exceeds the resolution of participants' ability to use the confidence scale.

1.6.1 Decision Rules for Lineup Models

As discussed in Section 1.4, the lineup task involves both a detection component (is the target present in the test array?) and an identification component (if so, which member is the target?) because multiple items are presented on each trial (Duncan, 2006). As a result, models of the lineup must include a decision rule that specifies how the familiarity values of

the lineup items are compared to the decision criteria to determine whether an item will be identified or the lineup rejected. A number of different decision rules have been used in lineup research, all instantiated within the unequal variance Gaussian framework shown in Figure 1.2. It is important to note that these decision rules do not aim describe the process by which individual witnesses complete the lineup task. Rather, they aim to account for the constraints of a given lineup task in order to provide an accurate mapping between observed decision performance and discriminability and response bias at the population level. The following sections describe the decision rules for models currently employed in lineup research.

1.6.1.1 The Integration Model (SDT-INT)

Originally developed by Duncan (2006), the aim of this model is to represent the lineup as a "compound decision" task that includes both detection and identification components. Within this framework, detection and identification may be associated with different decision rules. For the "integration" model (SDT-INT), the decision rule is to identify the maximally familiar lineup item if the summed familiarity of the lineup items exceeds the lowest decision criterion, otherwise, reject the lineup. The detection decision is based on the summed familiarity of the lineup items and the identification decision is based on the maximally familiar item in the lineup.

This was the first signal detection model applied to lineup research (Palmer et al., 2010) and has been used in a recent study (Smalarz et al., 2019). Its continued use is of some concern given that is has been shown to provide a poor fit to the data compared to competing models that specify alternative decision rules, possible because it makes an incorrect prediction about the impact of correlated memory strengths, i.e. the similarity between lineup members (Wixted et al., 2018). As a result, it may provide inaccurate estimates of discriminability and response bias for a given lineup procedure. Additionally, the model is

complex to implement relative to the better-performing maximum familiarity model (Wixted et al., 2018). The mathematical functions for the probability of each lineup decision outcome according to this model are available in Section 2.4.5.2, with more detail in Appendix A.

1.6.1.2 The Maximum Familiarity Model (SDT-MAX)

The maximum familiarity model, also known as the independent observations model, has a long history in perceptual research (Macmillan & Creelman, 2005; Wixted et al., 2018). The rule is to identify the maximally familiar item in the lineup if it exceeds the lowest decision criteria, otherwise, reject the lineup. In the context of a compound decision task, both the detection and identification components are based on the same decision rule. This model has provided an adequate fit to lineup data from a number of studies, although the more complex Ensemble model has a performance advantage (Wixted et al., 2018). The mathematical functions for the probability of each lineup decision according to this model are also available in Section 2.4.5.1 and Appendix A.

1.6.1.3 The Ensemble Model

This model has been developed recently as part of an effort to aid theory building in lineup research. For the Ensemble model, the decision rule is to identify the maximally familiar lineup member if the difference between the maximally familiar lineup item and the average familiarity of the lineup items exceeds the lowest decision criterion, otherwise, reject the lineup. The Ensemble model has been found to provide a better fit to lineup data than SDT-INT and SDT-MAX when fit to lineup data and is now widely used (e.g. Colloff & Wixted, 2019; Seale-Carlisle et al., 2019). The mathematical functions for the probability of each lineup decision according to this model are available in (Wixted et al., 2018).

1.6.2 The Unique Constraints of the Sequential Stopping Rule Lineup

SDT-MAX, SDT-INT and the Ensemble model are all plausible models of the simultaneous lineup task and the UK lineup task. However, these decision rules are not

compatible with the sequential stopping rule lineup task. Recall that the decision to identify an item on the sequential lineup terminates the procedure, remaining items are not shown (Lindsay & Wells, 1985). Some proportion of participants will identify prior to the end of the sequential lineup so the number of items viewed before identifying differs from trial to trial. On the simultaneous and UK lineups, all items are shown on every trial. This stopping rule is therefore incompatible with the detection decision rules for SDT-INT and the Ensemble model, which are a function of the familiarity of all lineup items, the sum and arithmetic mean, respectively. Additionally, it is not possible to determine which lineup item is maximally familiar lineup item unless all items are presented, which affects the identification decision rule for SDT-INT and the Ensemble model and both the detection and identification decision rules for SDT-MAX. These models cannot capture the underlying data generating process of the sequential stopping rule lineup.

This model misspecification is of theoretical importance because fitting these models to data from a sequential stopping rule lineup may lead to inaccurate parameter estimates, in turn leading to incorrect conclusions about the data. A number of studies have fit SDT-MAX and SDT-INT to data from the sequential stopping rule lineup. Palmer and Brewer (2012) fit SDT-INT to a corpus of earlier studies comparing the simultaneous lineup and the sequential lineup. They found that the two procedures did not differ in terms of discriminability, but that sequential stopping rule presentation led to more conservative responding. This important study was one of the first to find results consistent with the notion that the apparent superiority of the sequential stopping rule lineup, particularly in reducing false alarms, was due to a conservative criterion shift (Ebbesen & Flowe, 2002; Meissner et al., 2005). Horry et al. (2012b) used SDT-INT to investigate the practice of concealing the number of items in a sequential stopping rule lineup from the witness, known as “backloading”. This study also investigated position effects, analysing decisions at serial positions two and six in order to

understand how backloading might affect witness's responding over the course of the lineup. They found that the more items participants were told to expect, the more conservative their responding and that backloading prevented responding becoming more lenient over the course of the sequential lineup. They also found greater discriminability for position two compared to position six on a non-backloaded lineup. Carlson et al. (2016) also investigated backloading, fitting Duncan's (2006) version of SDT-MAX to the data. Similar to Horry et al. (2012b), they found that more backloading led to more conservative responding, although they did not find any difference in discriminability by serial position. Horry et al. (2015) used SDT-INT to explore the effects of a allowing a second lap through the sequential lineup items. This allowed participants to change their identification decision on the second lap. They found that electing to see a second lap was associated with poorer discriminability and that responding was more lenient on the second lap compared to the first.

It is possible that the results of these studies are affected by the misspecification of the SDT-MAX and SDT-INT models when fit to sequential lineup data. In order to measure discriminability and response bias on the sequential lineup and compare it to performance on other tasks, a signal detection model is required that can account for the stopping rule constraint.

1.6.3 A Model of the Sequential Lineup with a Stopping Rule

The starting point for deriving a decision rule is to assume that the *n* items on each sequential stopping rule lineup trial are presented one at a time in a random order, with an index from 1 to *n*. A decision rule that captures the stopping rule constraint is to identify the first lineup item of the items 1 – *n* that exceeds the lowest decision criterion. If no item exceeds the decision criterion, then reject the lineup. In contrast to the decision rules for SDT-MAX, SDT-INT and the Ensemble model, the SDT-SEQ decision rule allows for the possibility that the maximally familiar item in a lineup may not be identified if a less familiar

item that appears before it in exceeds the decision criterion. The mathematical functions for the probability of each lineup decision according to this model are available in Section 2.4.5.3 and Appendix A.

The SDT-SEQ model was introduced in a conference proceedings paper (Kaesler et al., 2017), followed by an expanded journal article that forms the first study of this thesis (Kaesler et al., 2020). It has since been employed in simulations by Wixted and Mickes (2018) and fit to empirical data by Wilson et al. (2019) to examine position effects in a six-item sequential lineup. Wilson et al. (2019) found that discriminability increased from position one to position two but did not significantly increase beyond position two. AUC showed the opposite pattern, decreasing with serial position. This is the first empirical demonstration of a dissociation between AUC and discriminability for the sequential lineup of the type shown in simulations by Rotello and Chen (2016). Wilson et al. (2019) also found that response bias varied by serial position but did not systematically increase or decrease.

The use of SDT-SEQ to examine position effects reinforces the need to consider the match between task and model. In order to estimate discriminability and response bias at each position, Wilson et al. (2019) constructed subsets of the data when the target appeared in position *j* and fit the model to the data, specifying that the target appeared in a particular position. While SDT-SEQ accounts for the position of the target in the lineup and can therefore account for the position at which target identifications are made, it does not account for the position at which foil identifications are made. For example, a foil identification made at position two is indistinguishable in the data structure from a foil identification made at position five. Because data from a subset of all trials when the target appeared in position *j* contains foil identifications made at all positions $\neq j$, SDT-SEQ may not accurately estimate discriminability and response bias at each serial position. This shortcoming of SDT-SEQ is addressed in study three, which describes a probabilistic model called the ISL model that

accounts for both the position of the target and the position at which identifications are made. This model can be recast as a signal detection model, SDT-ISL in order to measure discriminability and response bias at each serial position. In study three, the ISL model is used to reanalyse the Wilson et al. (2019) data and to analyse data from a new sequential lineup experiment.

1.6.4 Estimating Lineup Discriminability and Response Bias

Unlike the model of the yes/no task described in Section 1.3, the relative complexity of the lineup task means that the lineup models based on the framework in Figure 1.2 do not have closed form solutions for estimating discriminability and response bias. As a result, it is necessary to fit the models to data in order to estimate the parameter values parameters. This thesis employs the maximum likelihood estimation technique (Myung, 2003). In order to conduct maximum likelihood estimation, it is necessary to derive equations for the probability of each lineup decision outcome for a particular model and given set of parameter values. These equations are known as likelihood functions. From a set of starting parameter values, a computational optimisation routine searches the parameter space iteratively to arrive at the set of parameters are most likely to have generated the observed data. At each iteration, the optimiser calculates predicted data by passing the current set of parameter values to the likelihood functions, compares the predicted data against the observed data, and calculates a goodness-of-fit index. The optimiser then adjusts the parameter values for the next iteration, according to a particular algorithm, in order to improve the goodness-of-fit index. When no further improvement is possible, or another condition has been reached such as a maximum number of allowable iterations, the optimiser terminates and returns the best-fitting set of parameter values, a final goodness-of-fit value, and the predicted data. A supplementary document that steps through R code for fitting the SDT-MAX model to lineup data can be found in Appendix D.

1.6.5 Measurement at the Task Level

Psychological science is typically focused on providing explanations at the level of the individual. As discussed in Section 1.7.1, the eyewitness memory research conducted in the early 20th century displayed this tendency, focusing largely on assessing the credibility of individual witnesses. However, the modern period of lineup research has been characterised by the measurement of performance at the group level. This maps on to an important applied aim; to understand how changes to lineup procedures affect observed decision outcomes for a given police jurisdiction. The main question posed in this thesis is aligned with this aim; what is the effect on discriminability and response bias, and the resulting rates of target, foil and false identification, for a hypothetical jurisdiction if lineups are presented sequentially rather than simultaneously? To this end, the models introduced in the previous sections are used to measure discriminability and response bias for a population of witnesses rather than for individual witnesses. To understand why this is the case, it is necessary to review the structure of real world lineup data and its relationship to experimental research.

As alluded to in Section 1.5, each witness in a real criminal case makes a single lineup decision for a single suspect[1]. As the true guilt of the suspect is not known, these decisions can only be classified as either suspect identifications, foil identifications or rejections. Analysis of field data has focused on how changes to lineup procedure affect rates of these three possible responses (Horry et al., 2012a; Mecklenburg et al., 2008; Wells et al., 2011). It is not possible to assess decision performance directly from this data, as there is no indication of which suspect identifications and lineup rejections are correct or incorrect. If the guilt of the suspect could be conclusively established, or a certain base rate of target present lineups is assumed (Wixted et al., 2016), each identification decision made by a witness can be

[1]In rare cases, multiple witnesses may be asked to identify the same suspect (Horry et al., 2012a)

classified as one of the decision outcomes in Table 1.2. These identification decisions can be aggregated to calculate the decision outcome rates for a jurisdiction. The models described in this section can also be fit to data in this format in order to measure discriminability and response bias for a jurisdiction (Wixted et al., 2016). In order to maintain ecological validity, participants in experimental lineup studies also typically provide only a single lineup decision (cf. Mansour et al., 2017). These decisions are aggregated in each between-subjects condition before the data is analysed. This is analogous to comparing data from two jurisdictions that differ in their administration of the lineup task.

Measurement of latent variables at the group level with single trial experiments ensures that research outcomes are relevant to policy. It is possible to draw inferences about lineup tasks from "single trial" experiments that may be translated to policy with minimal adjustment. This may not be the case in an experiment employing many test trials for each participant, as is common in cognitive psychology, because this design does not match as closely the constraints of the task as applied in real police jurisdictions.

1.7 Investigating the Lineup Task

The preceding sections describe a measurement modelling framework based on signal detection theory for comparing lineup procedures on two latent variables that determine observed decision performance, discriminability and response bias. In order to explain how this framework can increase understanding in lineup research, it is necessary to establish the context in which it will be applied. The following sections review the origins of the field's general approach to investigating the lineup task and examine existing approaches to comparing lineup task performance.

1.7.1 The Origins of Modern Lineup Research

The possibility that psychological science might inform the legal system has been recognised since the beginnings of experimental psychology (Wundt & Judd, 1897). Much

early eyewitness memory research focused on the environmental factors that might affect witness' encoding of an event and the personal qualities of witnesses that may indicate their reliability (McCarty, 1929; Munsterberg, 1908; Robinson, 1935; Stern, 1903; Whipple, 1917). Wells (1978) shifted the focus of eyewitness to the procedures used to collect evidence, proposing the binary classification of factors affecting eyewitness memory as either estimator variables, which the justice system cannot control, or system variables, which the justice system can control. Understanding how changes to system variables affect witness decision behaviour is likely to be useful when developing evidence collection procedures that aim to reflect the priorities of stakeholders in the criminal justice system. The lineup procedure is one such system variable and is amenable to experimental investigation, as there are many possible methods for administering the task and multiple points at which a witness's ability to recognise might be affected or their post-decision confidence manipulated. Wells (1978) suggests that the field focus largely on the system variable research because the applied benefits are potentially greater than for estimator variable research.

Wells (1978) distilled the broad aims of early eyewitness research into a directive focused on the outcomes of the criminal justice system; to produce knowledge that would maximize guilty convictions while minimizing false convictions. In defining the scope of the system vs. estimator variable distinction, he distinguished between eyewitness memory research that addresses the applied concerns of the criminal justice system, prioritising ecological validity, and eyewitness memory research that is relevant to theoretical questions about memory, providing the example of false memory research (Loftus & Palmer, 1974). Like Munsterberg (1908), Wells (1978) conceived of applied eyewitness memory research as requiring direct applied outcomes that could be translated into practice. This applied focus

influenced the choice of methods for comparing lineup task performance in a long program of subsequent system variable research (Bornstein & Penrod, 2008).

1.7.2 Approaches to Comparing Lineup Task Performance

The central applied problem of eyewitness memory is that eyewitnesses are mistaken some proportion of the time. The most obvious consequence of this error is the many cases of false conviction that have resulted from the identification of an innocent suspect from a lineup (Garrett, 2012). There is also a largely unknown harm from perpetrators in lineups going free, who may commit additional crimes. Additionally, jurors find eyewitness evidence particularly persuasive (Semmler et al., 2012). Thus, the accuracy of witness lineup decisions is important to achieve the desired outcomes of the criminal justice system. On this basis, a sensible starting point for addressing the applied aim of maximising guilty convictions and minimising false convictions is to examine the effect of changes to lineup procedure on target identifications and false identifications (see Table 1.2). These outcomes are most often reported as rates, which can be calculated by dividing the target present decision frequencies by total number of target present lineups and the target absent decision outcome frequencies by the number of target absent lineups.

1.7.2.1 Analysing Choosing Rates

Some early lineup research directly compared rates of target identification and false identification between experimental conditions (e.g. Brown et al., 1977; Leippe et al., 1978). The limitation of this approach is that, unless one condition produces both a greater target identification rate and a lesser false identification rate, it is difficult to determine which condition was associated with superior performance. This limitation also affects logistic regression, which has been used in some lineup studies (Gronlund & Neuschatz, 2014) and linear mixed effects modelling (Weber & Varga, 2012). Competing lineup procedures are entered in to a model as predictors of binary correct/incorrect decision outcomes. This gives

the odds ratio for one binary outcome over another, e.g. the odds of a target identification on a target present lineup compared to an incorrect response (foil identification or miss), for a given procedure (e.g. Key et al., 2015). Recent studies have employed logistic regression as a complement to other measures of lineup performance (e.g. Horry et al., 2015; Wooten et al., 2020).

Another commonly used measure called the diagnosticity ratio (Wells & Lindsay, 1980) captures changes in both target and false identifications in a single index. It is calculated as the ratio of the target identification rate to the false identification rate and gives the probability that a suspect selected from a given lineup procedure is guilty rather than innocent. This aligns with the legal notion of "probative value", the extent to which evidence produced at trial supports the guilt or innocence of the defendant (Friedman, 1986). An identification from a lineup procedure with a greater diagnosticity ratio is, all else being equal, more likely to indicate that a suspect is guilty, although this ignores witness's post-decision confidence estimates (Wixted et al., 2014). Thus, a lineup procedure that produces a greater diagnosticity ratio is preferred. This measure was subsequently employed in studies that developed the procedural guidelines for lineups proposed by the U.S. National Department of Justice (1999), in addition to many studies investigating the sequential stopping rule lineup (Steblay et al., 2011b).

The major limitation of measures based on choosing rates is that they provide no insight into latent variables. Within a signal detection framework, observed decision outcomes on a lineup task reflect some combination of discriminability and response bias, in addition to other factors such as the format of the memory test. Recall the example in Section 1.4 of the difference in observed recognition performance between the 2AFC and yes/no test formats. This highlights the major shortcoming of comparing procedures using these measures; discriminability and response bias are confounded (Macmillan & Creelman, 2005;

Wixted & Mickes, 2012). This confound impedes understanding because it is unclear which latent variable, or other factor, is responsible for differences in decision outcomes between lineup procedures. The changing interpretation of the diagnosticity ratio from its introduction to the present day provides an instructive case study.

For much of its history, the diagnosticity ratio was thought to reflect the extent to which a procedure was diagnostic of suspect guilt (Lindsay & Wells, 1980). That is, procedures that produce a higher diagnosticity ratio are to be preferred. When viewed from a signal detection standpoint, the statistic largely captures changes in response bias – as opposed to discriminability – because it is very sensitive to changes in the false identification rate (Clark, 2012a; Wixted & Mickes, 2012). In most lineup studies, the target identification rate is greater than the false identification rate. As responding becomes more conservative, the false identification rate denominator approaches zero. This can result in a very large diagnosticity ratio. This property leads the diagnosticity ratio to favour conservative lineup procedures that reduce both target and false identifications, which was not widely appreciated for much of the period where it dominated eyewitness memory research. The use of this measure contributed to the development of the "no cost view" (Clark, 2012a, 2012b). Many lineup studies claimed that reforms that reduced false identifications either did not reduce target identifications or that the benefit of the larger reduction in false identifications outweighed a small reduction in target identifications (e.g. Lindsay & Pozzulo, 1999; Steblay et al., 2011b). Despite the findings of earlier studies, Clark (2012a) showed that the strong version of the "no cost view" does not hold for any of the major lineup reforms that have since been adopted as policy; in each case, a reduction in false identifications was accompanied by a reduction in target identifications. Additionally, the argument that the benefit of reducing false identifications outweighs the small loss of correct identifications is subjective, depending on the weighting of each major lineup error, failing to identify a target

and falsely identifying an innocent suspect. In effect, this weaker version of the "no cost view" presumes that a false identification is a far more serious error than a miss, which aligns with the values of many countries that employ an English legal system (Epps, 2015).

Research highlighting the tendency of the diagnosticity ratio to favour conservative procedures proposed that lineup procedures should instead be compared in terms of discriminability (Wixted & Mickes, 2012). This required the adoption of a signal detection framework that had seen limited use in eyewitness memory (Meissner et al., 2005), despite its long history in theoretical recognition memory research.

1.7.2.2 Receiver Operating Characteristic Analysis

In Sections 1.3 and 1.6, discriminability is formalised as the distance between the mean of two Gaussian distributions in units of their pooled standard deviation. This quantity is specific to a given signal detection model. For the equal variance model of the yes/no task, discriminability can be estimated using a particular closed-form solution. For the unequal variance lineup models, it must be estimated computationally using the likelihood functions that give the probability of each decision lineup decision outcome. Receiver Operating Characteristic (ROC) analysis provides an alternative method to signal detection modelling for measuring discriminability independent of response bias (Macmillan & Creelman, 2005). In general, a ROC curve can be constructed for a test procedure by plotting cumulative hit rates against cumulative false alarm rates at multiple levels of response bias. Discriminability can be measured by joining the hit rate/false alarm rate points and summing the resulting area under the ROC curve (AUC). All hit rate/false alarm rate pairs on a ROC curve are associated with the same level of discriminability, so greater AUC reflects greater discriminability.

ROC curves are constructed from lineup data by plotting the cumulative target identification rate against the cumulative false identification rate at multiple post-decision confidence categories, from conservative to lenient (Gronlund et al., 2014). The area under

the lineup ROC curve therefore provides a task-level measure of ability to discriminate targets from innocent suspects. If there is no designated innocent suspect in target absent lineups, then the false identification rate may be estimated by dividing the target absent foil identification rate by the lineup size. Some studies without a designated innocent suspect on target absent lineups plot the correct identification rate against the target absent foil identification rate (Mickes et al., 2017; Seale-Carlisle et al., 2019; Wilson et al., 2018), ostensibly to increase statistical power. In this case, AUC measures discriminability for guilty suspects compared to foils from a target absent lineup. Foil identifications on target present lineups and, in most cases, target absent lineups are not represented by a lineup ROC curve. As a result, these curves do not extend over the full range from zero to one in ROC space, as do ROC curves from tasks that share the 2 x 2 decision outcome structure of the yes/no task shown in Table 1.1. A lineup ROC curve terminates at the false identification/false alarm rate for the most lenient confidence category. As a result, lineup procedures are compared using partial AUC (pAUC), where the cut point for comparing pAUC for each curve is typically set at the most lenient false identification/false alarm rate for the more conservative procedure (Gronlund et al., 2014). This means that some portion of the ROC curve of the more lenient procedure will be discarded when making the comparison. This limitation has led some researchers to argue against the use of ROC analysis (Lampinen et al., 2019; Smith et al., 2019).

The aim of ROC analysis is to measure discriminability without relying on an underlying signal detection model. One benefit to this approach is that any two lineup procedures can be compared in terms of AUC, whereas comparing discriminability between lineup procedures requires careful selection and specification of models that can account for the structure of the tasks. However, “theory-free” quantification of discriminability requires compromises, even under ideal conditions. AUC is most often calculated is by connecting the

hit rate/false alarm rate pairs comprising the ROC curve using straight lines and summing the areas of the resulting trapezoids. As ROCs are generally curvilinear, this trapezoidal AUC calculation underestimates true AUC (Macmillan & Creelman, 2005; Wickens, 2002), a bias which worsens for ROC curves with fewer points. As participants in eyewitness memory experiments generally provide only a single lineup decision, a large sample is required in order to plot ROCs curves with many points. In the case of a single hit rate/false alarm rate pair, trapezoidal AUC cannot be calculated. This might be the case for a dataset that does not contain confidence ratings or when working with summary data from published studies. The next-most "theory-free" option is an average area statistic, A' (Pollack & Hsieh, 1969). A' behaves poorly at very high levels of performance and the formula for its calculation implies that the signal and noise distributions have equal variance, although it does not specify the form of these distributions (Wickens, 2002). It is therefore ill-suited to recognition memory tasks in general, where the respective variances of the signal and noise distributions are known to be unequal (Mickes et al., 2007; Wixted et al., 2018). The other alternatives for a single hit rate/false alarm rate point are to calculate discriminability from the yes/no task, using the formula described in Section 1.3 (Mickes et al., 2014), or the area under the theoretical isosensitivity curve A_z, both of which assume underlying Gaussian distributions for signal and noise. At this point, the measures are no longer atheoretical as some sort of underlying model is assumed. A_z is derived from a model for the two-alternative 2AFC decision task (Green & Swets, 1966). The 2AFC model is unlikely to adequately characterise data from the police lineup, as lineup tasks are not forced choice and generally contain three or more items.

1.7.2.3 Area under the Curve and Discriminability

AUC is perhaps best thought of as a measure of a detector’s ability to sort stimuli in to the correct response category, whereas discriminability is a measure of the separation of

underlying signal and noise distributions for a given decision task according to a particular signal detection model. The original justification for using AUC as a "model free" measure of decision performance is the finding that AUC for the 2AFC task is equal to discriminability for a model of the 2AFC task with underlying Gaussian distributions for signal and noise, the so called Area Theorem (Green & Swets, 1966). It is important to note that this one-to-one relationship between AUC and discriminability does not hold for other tasks. This means that AUC and discriminability can be dissociated in cases where some structural feature of the memory test affects the shape of the ROC curve (Rotello & Chen, 2016; Rotello et al., 2015; Stephens et al., 2019; Wixted & Mickes, 2018).

Once again, the comparison of AUC for the two-alternative forced choice (2AFC) task and yes/no task provides an example of such a dissociation. As discussed in Section 1.4, performance in terms of observed hit and false alarm rates is reliably better on a 2AFC test compared to a yes/no test, even when discriminability for each task, as estimated by appropriate models, is very similar (Jang et al., 2009; Macmillan & Creelman, 2005; Wixted & Mickes, 2018). AUC is calculated using these observed hit and false alarm rates, which means that AUC is reliably greater for a 2AFC memory test than for a yes/no memory test under the same conditions. This is a shortcoming of ROC analysis for comparing data generated from these tasks; it can neither separate nor quantify the respective contributions of discriminability and task structure to any differences in observed performance.

The potential for dissociation between AUC and discriminability is relevant to theoretical development in lineup research. When aspects of the lineup task are manipulated that do not alter the structure of the task, differences in AUC are likely to reflect differences in discriminability. For example, Carlson et al. (2019, experiment one) manipulated the number of features shared between lineup items using computer generated stimuli. They found that AUC was greater in conditions where there were fewer shared features. In all

experimental conditions, participants were tested using six item simultaneous lineups. As the lineup task was the same in each experimental condition, the greater AUC for the conditions in which fewer features were shared between the lineup items likely reflects greater discriminability. This result is theoretically informative; greater performance for one condition is explained by a corresponding increase in a latent variable. It also conforms to a prediction of the diagnostic feature detection theory of eyewitness memory, described in Section 1.8.3.

As foreshadowed in Section 1.6, some procedural manipulations alter the structure of lineup task, which affects the shape of the corresponding ROC curve. There are three major lines of research affected by this issue; the comparison of the single-suspect showup to a lineup, the comparison of lineups that differ in the number of items they contain and the comparison of the simultaneous lineup and the sequential stopping-rule lineup.

The showup is a special case of the lineup with only one item, so the first two issues can be discussed together. As the number of lineup items increases, the maximum possible false identification rate decreases. Recall that a lineup ROC curve terminates at the false identification rate of the lowest confidence category, i.e. the choose/no-choose threshold. Under the most lenient response bias, i.e. always choose an item, the false identification rate is equal to 1 divided by the lineup size. This means that lineups with fewer items necessarily have higher false alarm rates at the choose/no-choose threshold and extend further along the *x*-axis in ROC space. As a result, AUC (and pAUC at the same cut point) will differ between ROC curves from lineups containing different numbers of items, even if discriminability is the same. A simulation presented in Section 2.4.4 illustrates this effect. The shape of the sequential stopping rule lineup ROC is also affected by the structure of the task. Through simulation, Rotello and Chen (2016) showed that the stopping rule constraint of the sequential lineup leads to a non-monotonic ROC curve (also see Cohen et al., 2020), which

has been demonstrated empirically by Wilson et al. (2019). As responding becomes increasingly lenient, target and false identification rates decrease, because the first sequentially presented item will tend to be identified, resulting in the sequential lineup ROC curve falling back toward the line of chance performance. In contrast, the simultaneous ROC curve is monotonic; responding becomes more lenient, target and false identification rates increase. As in the situation of comparing lineups with different numbers of items, a difference in AUC between simultaneous and sequential presentation using a stopping rule does not necessarily indicate a difference in discriminability.

Many studies have used ROC analysis to compare showups and lineups (Colloff & Wixted, 2019; Gronlund et al., 2012; Key et al., 2015; Neuschatz et al., 2016; Wetmore et al., 2015; Wooten et al., 2020), reporting that lineups reliably lead to greater AUC. A number of studies have also compared simultaneous presentation to sequential presentation using ROC analysis. Some have reported greater AUC for simultaneous presentation (Experiment 1a Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Mickes et al., 2012; Neuschatz et al., 2016), while others have reported no difference between the procedures (Experiment 1b and 2 Andersen et al., 2014; Flowe et al., 2016; Gronlund et al., 2012; Mickes et al., 2012; Sučić et al., 2015). The results of these ROC studies are of great applied significance. Lower AUC for showups indicates that police should prefer lineups to showups, despite the relative ease of constructing administering showups (Wells et al., 2020). Greater AUC for the simultaneous lineup compared to the sequential lineup is contradicts the previous consensus in the literature. For many years, sequential presentation was thought to be superior to simultaneous presentation, based on the finding that it reliably produces a greater diagnosticity ratio (Steblay et al., 2011b). As previously discussed, the diagnosticity ratio is effectively a measure of response bias that favours conservative procedures. Consistent with this, subsequent research has found that the sequential lineup leads to more conservative

responding than the simultaneous lineup (Palmer & Brewer, 2012). If the simultaneous lineup is associated with greater AUC than the sequential lineup, then it may give policymakers cause to re-examine its use in real police lineups. From a theoretical perspective, the results of ROC analysis provide minimal explanation for why lineups outperform showups or why simultaneous presentation outperforms sequential presentation, because differences in AUC between these lineup tasks reflect some combination of task structure and discriminability.

1.7.3 Theoretical Limitations of Non-Model Based Measures

The introduction of ROC analysis is a positive development for addressing applied research questions in lineup research. AUC provides a measure of discrimination ability dissociated from response bias, in contrast to previously employed choosing rate measures. The advantages of ROC analysis have been noted by policymakers (National Research Council, 2014). However, the significant effort expended on debating which measure should be used to compare lineup procedures in terms of their applied utility (Lampinen, 2016; Wells et al., 2015b; Wixted et al., 2017) does not address the question of their theoretical utility.

Choosing-rate based measures and ROC analysis are of minimal utility for addressing theoretical questions because they are uninformative with respect to how changes in observed responses reflect changes in latent variables discriminability and response bias. That is, they are unable to explain why one procedure might lead to superior applied outcomes, however that might be defined. An ability to explain how differences in observed decision outcomes arise, rather than simply describe them with respect to applied aims, lies at the heart of developing and testing theories of eyewitness memory. These theories can be used to guide the development of future lineup procedures, rather than relying on intuition or instinct. The importance of theoretical development in this regard has been emphasised by numerous authors (e.g. Brewer et al., 2007; Clark, 2008; Clark et al., 2015a; Clark et al., 2014; Lane &

Meissner, 2008; Wells, 2008). However, much lineup research remains resolutely applied in focus, tending to address theoretical concerns only when there is overlap with some applied question (Carlson et al., 2019; Wooten et al., 2020).

In contrast to choosing rate based measures and ROC analysis, signal detection models measure latent variables that determine observed performance. Using these models, it is possible to quantify the effects of changes to lineup procedure on discriminability and response bias, in addition to the contribution of task structure. The ability to explain how changes to lineup procedure result in observed decision outcomes provides a starting point for developing and testing theories of eyewitness memory that choosing rate based measures and ROC analysis cannot.

1.8 Theories of Eyewitness Memory

Section 1.7.1 explored how the modern period of lineup research came to focus more on applied rather than theoretical research. This focus is reflected in the decision to compare procedures in terms of observed decision outcomes rather than latent variables. It has also influenced the theory advanced to explain patterns in observed decision outcomes (Bornstein & Penrod, 2008; Gronlund & Benjamin, 2018). This section reviews the two major theoretical positions in lineup research. The first of these is the distinction between absolute and relative judgment distinction (Wells, 1984), which was the dominant framework for much of the field's history and is closely linked to the diagnosticity ratio statistic. It proposes that differences in observed decision outcomes between lineup procedures result from the extent to which they encourage witnesses to adopt an absolute as opposed to a relative judgment strategy. The second is a signal detection-based diagnostic feature detection theory of eyewitness memory proposed by Wixted and Mickes (2014). This theory states that differences in observed decision outcomes between lineup procedures result from differences

in discriminability, which is facilitated by the extent to which a procedure allows witnesses to isolate distinctive features shared by the target and the lineup items.

1.8.1 Absolute vs Relative Judgment

Wells (1984) proposed a theoretical distinction between absolute and relative judgment strategies. Witnesses who adopt a relative judgement strategy select the lineup item most similar to the target *relative* to the other lineup items. This results in them identifying more readily when the target is absent from the lineup. In contrast, witnesses who adopt an absolute strategy directly compare each lineup item to their memory for the target. This results in them identifying less readily from target absent lineups, provided that the innocent suspect does not resemble the target to a greater degree than the other foils.

This distinction was proposed based on the results of early lineup studies (Lindsay & Wells, 1980) that manipulated the similarity between foils and the innocent suspect on target absent lineups but held constant the similarity between the innocent suspect and the target. Participants were more likely to select the innocent suspect when the foils were dissimilar rather than when they were similar (also see Wells et al., 1993). Wells (1984) interpreted this as evidence of a relative judgment strategy, because the similarity of the innocent suspect changed relative to the foils surrounding it, not to the target. Malpass and Devine (1981) found that informing witnesses that the perpetrator may or may not be present in the lineup reduced the tendency for witnesses to choose on a target absent lineups but not target present lineups. This was interpreted as an instruction that successfully discouraging participants from adopting relative judgement strategy.

In an experimental demonstration of the operation of relative judgements, Wells (1993) compared the choosing rates of participants shown a simultaneous six-item target present lineup to those shown the same lineup but with the target removed. In the target removed condition, a large proportion of participants identified the next-most frequently

chosen item from the target present condition rather than rejecting the lineup. This pattern of results was presumed to reflect the operation of a relative judgement strategy, because participants employing an absolute judgement strategy would have rejected the target removed lineup rather than identifying another lineup member (Quinlivan et al., 2017; Wells et al., 1998). Based on this empirical evidence, the absolute vs. relative judgement distinction became a guiding framework for motivating the reforms proposed by Wells et al. (1998), which influenced the guidelines developed by US National Department of Justice (1999). That is, some reforms proposed in this period of eyewitness research sought *a priori* to reduce participants' use of a relative judgement strategy, while other studies argued *post hoc* that a reduction in false identification rates was due to some procedural manipulation discouraging the use of a relative judgement strategy (Wells et al., 1998).

1.8.1.1 Limitations of the Absolute vs. Relative Judgment Distinction

A major limitation of the absolute vs. relative judgement distinction is that it is verbally specified. Verbal theories are difficult to falsify because their predictions are imprecise and therefore difficult to test. Many studies that were interpreted as supporting the absolute vs. relative distinction manipulated some factor that is presumed to influence judgement strategy and showed the expected effect on the rate of target and false identification, often summarised using the diagnosticity ratio (e.g. Lindsay & Wells, 1985; Lindsay et al., 1991b). This indirect method is necessary because the theory does not provide a measurement model that maps observed performance to latent variables that represent absolute and relative judgment. It also rests heavily on the validity of the experimental manipulation. Some studies sought to measure the contribution of absolute and relative judgement through self-report (Dunning & Stern, 1994; Lindsay et al., 1991a) although this rests on the assumption that participants have insight in to their cognitive processes, which is

known to be false under a variety of conditions (Haeffel & Howard, 2010; Nisbett & Wilson, 1977).

Realising the limitations of verbal specification, Clark (2003) formalised the absolute vs. relative judgment distinction in a global memory matching model called WITNESS. This model has parameters that correspond to the contributions of an absolute vs. relative judgment strategy. Fits of the model to data from previous studies (Clark et al., 2011) found some evidence for the superiority of an absolute strategy. However, the parameters in the WITNESS model for absolute and relative judgement were found be difficult to identify by Fife et al. (2014). This means that WITNESS can predict the same decision outcome rates using most combinations of parameter values for the respective contributions of absolute and relative judgment. Thus, the extent to which absolute and relative judgement determine witness decision performance remains unclear.

Finally, the theory alludes to constructs from formal models of recognition memory without adopting the formal aspects of these frameworks to understand performance (Clark, 2012b). Models of recognition memory often include some kind of threshold for degree of match to memory that separates one kind of response from another, e.g. a rejection from an identification. The absolute decision strategy is described as comparing each individual lineup item to a decision criterion. In contrast, the relative judgement strategy is described as involving no criterion and nothing preventing a witness from making an identification, other than that the items cannot all be equally familiar (Wells, 1984; Wells et al., 1998). Gronlund et al. (2015) reinterpreted the results of Wells (1993) through a signal detection theory lens, arguing that their results could reflect the situation where the target and the next most familiar foil exceed the decision criterion. Once the target is removed, the next most familiar foil will be selected under the same criterion. That is, the results of Wells (1993) are also consistent with a single criterion explanation. In fact, some researchers have proposed that

the absolute vs. relative judgement distinction is a theory of response bias (Wixted & Mickes, 2014). The difference between the two strategies is essentially that one strategy has a much lower criterion for identifying an item. Additionally, McAdoo and Gronlund (2016) formalised the absolute vs relative judgement distinction as a dual process signal detection model, where absolute judgements are analogous to recollection and relative judgements are analogous to recognition (Yonelinas, 1994). Using a ranking task, they showed that memory appeared to be based on recognition alone, contrary to the verbal description of relative judgement theory in Wells et al. (2012), which implied that some target identifications are based on recollection and some on recognition. Given the known limitations of dual-process theories of recognition memory in explaining patterns in observed data (Dunn, 2004; Wais et al., 2008), this development is unlikely to advance the absolute vs relative distinction to any great extent.

These theoretical issues do not just impede the building of scientific understanding, they also limit the utility of the theory for addressing applied problems. The limitations of the absolute vs. relative judgement distinction as a guiding framework are most evident when examining its role in the development of the sequential stopping rule lineup.

1.8.1.2 Development of the Sequential Stopping Rule Lineup

In their rationale for developing the sequential stopping rule lineup, Lindsay and Wells (1985) proposed that presenting lineup items sequentially would discourage witnesses from adopting a relative judgment strategy, thereby reducing the false alarm rate compared to simultaneous item presentation. Presumably, this is because sequential item presentation restricts the ability of witnesses to compare across lineup items. An experiment comparing the two procedures showed that the sequential stopping rule lineup led to a small reduction in the target identification rate and a larger reduction in the false identification rate compared to the simultaneous lineup. Lindsay and Wells (1985) compared the two procedures using the

diagnosticity ratio statistic discussed in Section 1.7.2.1, finding that sequential presentation generated a higher diagnosticity ratio. The finding that sequential presentation led to a higher diagnosticity ratio than simultaneous presentation, largely due to a reduction in the false identification rate, was replicated in many subsequent studies (Steblay et al., 2001; Steblay et al., 2011b).

This pattern of results was considered to support the absolute vs. relative judgement distinction and was also interpreted as evidence for the efficacy of sequential presentation in discouraging relative judgements. However, subsequent research employing signal detection theory has shown that this apparent sequential lineup advantage as indicated by the diagnosticity ratio is a result of more conservative responding (Mickes et al., 2012; Palmer & Brewer, 2012). The sequential lineup is likely to be either no different or possibly inferior to the simultaneous lineup in terms of discriminability (Mickes et al., 2012; Palmer & Brewer, 2012). The absolute vs. relative distinction is not informative in this regard because it does not make predictions about discriminability. Rather, this new understanding of the sequential lineup was reached by adopting the kinds of analysis used in theoretical recognition memory research. This potential misunderstanding of the effects of sequential item presentation on decision behaviour emphasises the limitations of a heavy focus on applied outcomes and a disconnection with relevant theoretical research (Gronlund & Benjamin, 2018). The combination of a limited theory, in the form of the absolute vs. relative judgement distinction, and a misleading dependent measure, in the form of the diagnosticity ratio, led researchers to develop the sequential lineup and advocate for its use. Its utility for real police jurisdictions is currently unclear, as it may not outperform the simultaneous lineup procedure.

1.8.1.3 Summary

The absolute vs. relative judgement distinction is ill-suited to guiding the development of future lineup procedures. In verbal form, it is difficult to test and, when

formalised, the predictions it makes about latent variables corresponding to the contributions of each decision strategy have found minimal support. Finally, the theory makes no clear predictions about the effect of changes to lineup procedure on discriminability, the latent variable of greatest theoretical interest to memory researchers.

1.8.2 Diagnostic Feature Detection Theory

Wixted and Mickes (2014) proposed a theory based on a signal detection framework to account for the results of ROC studies showing that simultaneous presentation was associated with greater AUC than sequential presentation (Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Mickes et al., 2012). Dubbed the Diagnostic Feature Detection Theory (DFDT), it proposes that discriminability on the lineup task results from the isolation of distinctive features shared uniquely by the perpetrator and the lineup items. Drawing on research in perceptual discrimination (Gibson, 1969), Wixted and Mickes (2014) suggested that affording witnesses greater opportunity to compare across lineup items may enhance discriminability by facilitating the isolation of distinctive features. This is in contrast to the absolute vs relative judgment distinction, which proposes that comparing across lineup members is detrimental to decision accuracy (Wixted & Mickes, 2014). DFDT predicts that simultaneous presentation should be associated with greater discriminability than sequential presentation because witnesses are better able to compare across items on the simultaneous lineup compared to the sequential stopping rule lineup. This explains the observed AUC advantage for simultaneous presentation.

Wixted and Mickes (2014) produced a simulation to show how the presence of additional items in a simultaneous lineup might allow the detection of features compared to a single suspect showup. They modelled four features; age, race, face shape and eye size. The features shared by a lineup item and the witness's memory of the perpetrator were modelled as a random draw from a Gaussian distribution with a mean of 1 and a standard deviation of

1.22. Features not shared by a lineup item and the witness's memory for the perpetrator were modelled as a random draw from a Gaussian distribution with mean of 0 and a standard deviation of 1. The higher mean for shared feature distributions is because features of the perpetrator encoded during the event should elicit greater memory strength than features not encoded during the event and the larger standard deviation reflects encoding variability (Mickes et al., 2007). Memory strength for each feature was combined by summing their means and variances to calculate an overall discriminability value for each test format. In the simulation, the innocent suspect shared age and race, but not face shape and eye size, with the target. It was assumed that the simultaneous lineup allows witnesses to discount the common features of age and race and focus only on the distinctive features of face shape and eye size, which leads to greater discriminability between the target and the innocent suspect compared to the showup, on which witnesses could not discount the common features of age and race.

This formalisation of DFDT makes a number of predictions about the effect of structural changes to lineup procedures on discriminability. It predicts that the simultaneous lineup will be associated with greater discriminability than the single suspect showup and the sequential lineup, as the simultaneous lineup offers more opportunity to compare across lineup members. It also predicts that discriminability will increase over the course of the sequential stopping rule lineup, as the presentation of each item provides an additional opportunity to isolate distinctive features. These predictions pertain to structural aspects of the lineup task. As discussed in Section 1.7.2.3, when two tasks differ in structure, greater AUC for one task compared to another does not necessarily indicate greater discriminability. In this case, comparing tasks in terms of discriminability requires signal detection models that can account for task structure. This is pertinent to studies employing ROC analysis that either aimed to test DFDT or have been retrospectively interpreted as supporting its predictions.

1.8.2.1 Simultaneous vs. Sequential Item Presentation Prediction

DFDT predicts that discriminability will be lower for both the sequential stopping rule lineup and the UK lineup compared to the simultaneous lineup. As discussed in Section 1.7.2.3, AUC and discriminability are dissociated for the sequential stopping rule lineup, so it is not possible to say whether greater AUC for simultaneous presentation (e.g. Mickes et al., 2012) indicates greater discriminability. The only study comparing the two procedures by fitting a signal detection model was conducted by Palmer and Brewer (2012). Employing the SDT-INT model, they found that the two procedures did not differ in discriminability, contrary to the prediction of DFDT. However, as discussed in Section 1.6.1.1, SDT-INT is not consistent with the stopping rule constraint, which may have affected Palmer and Brewer's (2012) results. Study one of this thesis is the first since the introduction of DFDT (Wixted & Mickes, 2014) to test this prediction of the theory by measuring discriminability using a suitable model of the sequential lineup, SDT-SEQ, rather than using AUC as proxy.

Also relevant to this prediction of DFDT are studies comparing the simultaneous lineup to the sequentially presented UK lineup (Seale-Carlisle & Mickes, 2016 ; Seale-Carlisle et al., 2019). On both tasks, only a single identification decision is made after the presentation of all items, as opposed to the yes/no decision for each item on the sequential stopping rule lineup. This means that simultaneous and UK lineups do not fundamentally differ in structure, despite the fact that lineup items are presented sequentially on the UK lineup and that the set of items must be viewed twice before making an identification decision. Holding all else equal, a difference in AUC between the procedures therefore likely reflects a difference in discriminability. Seale-Carlisle et al. (2019) and Seale-Carlisle and Mickes (2016) employed ROC analysis and signal detection modelling, finding that the simultaneous lineup is associated with both greater AUC and discriminability than the UK lineup, consistent with the predictions of DFDT. It is unclear whether the sequential

presentation of items is the primary cause of this decrement, as concluded by Seale-Carlisle et al. (2019), or whether it is due to some other aspect of the UK lineup procedure. Study two of this thesis seeks to clarify the effects of sequential item presentation by comparing simultaneous, sequential and UK lineup presentation, extending the findings of study one.

1.8.2.2 Sequential Lineup Position Effect Prediction

Wilson et al. (2019) found an increase in discriminability from serial position one to position two on the sequential stopping rule lineup task. This partially conforms to the predictions of DFDT, although the theory is somewhat unclear about whether discriminability should continue to increase with lineup size or whether this effect is mitigated by some other factor, such as newly presented test items interfering with memory for the target (Seale-Carlisle et al., 2019). As discussed in Section 1.6.3, Wilson et al. (2019) fit the SDT-SEQ model to estimate discriminability at each serial position, which does not account for the position at which an identification was made and is therefore ill-suited to investigating position effects. Study three of this thesis develops the Independent Sequential Lineup (ISL) model, which does account for identification position, and uses it to reanalyse the Wilson et al. (2019) data in addition to new experimental data, providing a further test of the DFDT prediction that discriminability increases with over the course of the sequential lineup.

1.8.2.3 Showup vs. Simultaneous Lineup Prediction

A number of studies have used ROC analysis to compare discriminability on the showup and the simultaneous lineup. All studies reported that AUC was greater for the simultaneous lineup than the showup. As for the sequential lineup, the two procedures differ in structure, which means that differences in AUC do not necessarily indicate differences in discriminability. Of these studies, Colloff and Wixted (2019) also fit a signal detection model to the data, which indicated that the simultaneous lineup was associated with greater discriminability than the showup. Colloff and Wixted (2019) also conducted an informative

experimental test of DFDT, comparing performance on the single suspect showup to a showup where the single suspect for identification was surrounded by foils that could not be chosen. Both AUC and discriminability were higher in the "simultaneous showup" condition compared to the single suspect showup condition. The presence of foil items, even if they could not be chosen by the witness, seems to have improved discriminability, which is consistent with DFDT.

1.8.2.4 Other Predictions

Predictions of the theory not related to task structure are also supported by empirical results. Carlson et al. (2019) manipulated the number of distinctive features shared the target and the lineup items using both computer generated and real stimuli. They found that AUC was greater when fewer features were shared between the lineup items, i.e. when there were more distinctive features, concluding that this result supported the predictions DFDT. As the same six-item simultaneous lineup was used for each experimental condition, differences in AUC between the conditions likely reflect differences in discriminability caused by the manipulation of the number of shared features. The prediction that increasing the similarity of lineup foils improves discriminability because it better allows witnesses to isolate diagnostic features has also found empirical support (Colloff et al., 2016; Colloff et al., 2018; Colloff et al., 2017; Wetmore et al., 2017).

1.8.2.5 Summary

DFDT improves on the absolute vs. relative judgements distinction because it makes testable predictions about the effect of procedural reforms on discriminability, one of the most important latent variables measured by models of recognition memory. Testing the DFDT predictions that relate to structural aspects of lineup procedure therefore requires signal detection models that can capture the structure of the tasks under investigation.

1.8.3 Limitations of Existing Modelling Research for Theoretical Development

Wixted and Mickes (2018) proposed that AUC as measured by ROC analysis, which they term empirical discriminability, is relevant to policy concerns, while discriminability as measured by signal detection models, which they term underlying discriminability, is relevant to theoretical questions. Aside from legal considerations of procedural fairness, this implies that policy makers should prefer the lineup procedure that produces the greatest AUC regardless of whether this results from greater discriminability for that procedure or some structural aspect of the task. As argued in Section 1.7.3, it is important to understand why one procedure leads to better performance in terms of observed decision outcomes than another. This knowledge can be built in to coherent theories, which can be used to develop lineup procedures that better achieve applied aims. Wixted and Mickes (2018) also state that the measures most often agree about which procedure is superior, a claim used by others (e.g. Colloff & Wixted, 2019; Wooten et al., 2020) as a justification for interpreting AUC as a proxy for discriminability. As discussed in Section 1.7.2.3, AUC is only a proxy for discriminability when the procedural manipulation in question does not alter the structure of the lineup task.

While some studies have used signal detection models to specifically address theoretical questions (Wilson et al., 2019; Wixted et al., 2018), many studies have employed model based analyses as a complement to ROC analysis (Colloff et al., 2016; Colloff et al., 2017; Colloff & Wixted, 2019; Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019) or choosing rate based analyses (Smalarz et al., 2019). For example, Seale-Carlisle and Mickes (2016) and Seale-Carlisle et al. (2019) reported both AUC and discriminability as estimated by a signal detection model for simultaneous lineup presentation compared to UK lineup presentation. They largely focused on the applied utility of the procedures as measured by AUC when discussing their results, commenting that discriminability and AUC agreed in all

cases about which procedure should be preferred. While researchers take care to highlight the conceptual difference and potential for dissociation between AUC and discriminability (Colloff & Wixted, 2019), this approach has the potential to obscure the different purpose of the two measures as outlined by (Wixted & Mickes, 2018). This body of research may give the impression that modelling is most useful as a kind of "sanity check" for the results of ROC analysis with respect to the applied utility of a given procedure. In fact, model based analyses aim to answer theoretical questions and are a necessity when procedural manipulations alter the structure of the lineup task, leading to a dissociation between AUC and discriminability. Theoretical development in lineup research would be hampered if signal detection models were primarily used to support the result of ROC analysis with respect to applied outcomes.

1.9 Aims and Study Summaries

This thesis aims to improve understanding of how the sequential presentation of lineup items affects memory strength, i.e. discriminability, and response bias. Much prior research examining the sequential stopping rule lineup (Lindsay & Wells, 1985) has employed ROC analysis as a proxy for discriminability (e.g. Mickes et al., 2012), however AUC does not measure discriminability on this task because the stopping rule constraint affects the shape of the ROC curve (Rotello & Chen, 2016; Wilson et al., 2019). Other research has measured discriminability on the sequential stopping rule lineup task using signal detection models that do not account for the stopping rule constraint (Carlson et al., 2016; Palmer & Brewer, 2012), potentially compromising measurement accuracy. Measuring discriminability on the stopping-rule sequential lineup requires the development of models that can account for the unique constraint that an identification terminates the memory test.

The first study in this thesis develops a signal detection model for characterising data from the sequential lineup task, SDT-SEQ. This model, along with two extant candidate

models of the simultaneous lineup, SDT-MAX and SDT-INT, is fit to lineup data to compare the simultaneous and sequential stopping rule lineups in terms of discriminability and response bias. This tests the prediction of the diagnostic feature detection hypothesis that the sequential presentation of items impairs discriminability relative to the simultaneous presentation of items. Two corpora of previously published studies comparing simultaneous and sequential stopping-rule presentation are analysed, in addition to data from a new experiment comparing simultaneous and sequential presentation. This experiment addresses the limitations of earlier studies by employing a larger sample size. In an attempt to minimise the potential for stimulus effects, a pool of stimuli that acted as both targets and foils was used, rather a single target with a set of accompanying foils. This study also allows some comparison of the performance of SDT-MAX and SDT-INT in accounting for simultaneous lineup data.

The second study extends the first by comparing simultaneous, sequential stopping rule and UK lineup presentation in a more powerful experiment. This is the first study to compare the sequential stopping rule lineup and the UK lineup directly. The aim is to unify the results of previous studies that have separately compared these tasks to the simultaneous lineup in order to clarify whether the sequential presentation of items itself negatively affects discriminability, as proposed by the diagnostic feature detection hypothesis, or whether some other aspect of the UK lineup task might be responsible for its apparent decrement in discriminability compared to the simultaneous lineup (Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019; Wixted et al., 2018).

The third study develops a model for examining changes in discriminability and response bias over serial position in the sequential stopping rule lineup, the ISL model. The critical difference between the ISL model and the SDT-SEQ model developed in the first study is that the ISL model accounts for the serial position at which an identification is made,

whereas SDT-SEQ does not. This model is used to reanalyse the results of a previous study that reported an increase in discriminability with serial position (Wilson et al., 2019), consistent with diagnostic feature detection theory, in addition to newly collected experimental data.

Together, these three studies employ a measurement model approach informed by research in theoretical research on recognition memory. They develop and apply new models to address both theoretical and applied questions about how item presentation format affects decision behaviour. The broad aim is to increase understanding of eyewitness memory and the procedures used to collect identification evidence.

Chapter 2

2.1 Preface to Study One

Previous studies evaluating the performance of the simultaneous and sequential stopping rule lineup procedures have compared the procedures in terms of observed decision outcomes, rather than the latent variables that determine them. This body of research is of some applied utility in determining which procedure should be preferred by policymakers. However, it is limited in its ability to inform theory because it does not explain why changing the presentation format of the lineup leads to changes in the observed decision outcomes of interest.

Study one introduces a measurement model approach based in signal detection theory that can be used to quantify two important latent variables that govern observed decision outcomes, discriminability and response bias. In order to accurately estimate these parameters, it is necessary for the models to capture the structure of the task under investigation. This is relevant to the stopping rule constraint used in the most commonly studied version of the sequential lineup tasks, which means that not all items will be seen by every participant on every trial. This study develops and describes a model that accounts for this constraint of the sequential stopping rule lineup, SDT-SEQ.

This model, in conjunction with models of the simultaneous lineup, SDT-MAX, SDT-INT and the Ensemble model, are used to compare discriminability and response bias on the simultaneous and sequential stopping rule lineups. The models are fit to data from a large number of previously published studies comparing the simultaneous and sequential stopping rule lineup tasks. A new experiment was also conducted to address the statistical power and methodological limitations of previously collected data. In this experiment, participants were randomised to see either a target present or target absent lineup, presented either simultaneously or sequentially with a stopping rule. Each participant provided only a single

lineup decision. A target was randomly selected on each trial from a pool of sixteen faces, with the necessary foils also randomly selected from the pool. The position of the target in the lineup was also randomised. This approach to lineup composition was motivated by a desire to avoid stimulus effects that may occur if only a single target and accompanying set of foils is used for an experiment, as is the case in some older lineup studies.

2.2 Statement of Authorship

Title of Paper	Do Sequential Lineups Impair Underlying Discriminabiltiy?
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Kaesler, M., Dunn, J. C., Ransom, K., & Semmler, C. (2020). Do sequential lineups impair underlying discriminability? *Cognitive Research: Principles and Implications*, *5*(1), 35.

Principal Author

Name of Principal Author (Candidate)	Matthew Kaesler		
Contribution to the Paper	Designed the study, selected stimuli for experiment, performed hypothesis tests for Palmer and Brewer (2012) corpus data, processed and fit mathematical models to the experimental data, extracted and analysed data for the corpus of post-2012 studies, wrote the manuscript, submitted the manuscript, and completed necessary revisions and proofs.		
Overall percentage (%)	70		
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	13/07/2021

Co-Author Contributions

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	Prof John C Dunn

Contribution to the Paper	Assisted in study design, developed the mathematical models used to analyse the data and wrote code for fitting them, extracted and analysed data from the Palmer and Brewer (2012) corpus of studies, provided feedback on manuscript drafts.		
Signature		Date	13/7/2021

Name of Co-Author	Dr Keith Ransom		
Contribution to the Paper	Programmed the web-based experiment for collecting data, assisted in stimuli selection process, collected the data, provided a base script for processing the raw data, which was adapted by MK. Provided feedback on (final) manuscript draft.		
Signature		Date	24/2/2021
Name of Co-Author	A/Prof Carolyn Semmler		
Contribution to the Paper	Supervised the development of the study, the design of the experiment and its execution, and the analysis of the data. Provided feedback on manuscript drafts.		
Signature		Date	23/02/2020

Do sequential lineups impair underlying discriminability?

by

Matthew Kaesler[1], John C. Dunn[2,3], Carolyn Semmler[1]

[1] University of Adelaide

[2] University of Western Australia

[3] Edith Cowan University

Address for correspondence:

Mr Matthew Kaesler
University of Adelaide
North Terrace
Adelaide, SA, 5005
Email:

Publication Details

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

2.3 Abstract

Debate regarding the best way to test and measure eyewitness memory has dominated the eyewitness literature for more than thirty years. We argue that resolution of this debate requires the development and application of appropriate measurement models. In this study we develop models of simultaneous and sequential lineup presentations and use these to compare these procedures in terms of underlying discriminability[2] and response bias, thereby testing a key prediction of diagnostic feature detection theory, that underlying discriminability should be greater for simultaneous than for stopping-rule sequential lineups. We fit the models to the corpus of studies originally described by Palmer and Brewer (2012, *Law and Human Behavior, 36*(3), 247-255), to data from a new experiment and to a corpus of eight recent studies comparing simultaneous and sequential presentation. We found that although responses tended to be more conservative for sequential lineups there was little or no difference in underlying discriminability between the two procedures. We discuss the implications of these results for the diagnostic feature detection theory and other kinds of sequential lineups used in current jurisdictions.

[2]The term "underlying discriminability", used throughout the manuscripts presented in chapters two, three and four of this thesis, refers to model-based discriminability. The term "empirical discriminability" refers to Area under the ROC Curve (AUC). See Wixted and Mickes (2018) for a full discussion. This terminology was adopted based on reviewer comments for this published manuscript, which led to its use in the prepared manuscripts that comprise chapters three and four. My personal view is that referring to model-based measures as "discriminability" and ROC analysis measures as "AUC" is both clearer and more accurate. I use my preferred terminology in the general introduction and discussion sections of this thesis, chapters one and five, respectively.

2.4 Introduction

A major goal of eyewitness research is to develop procedures that maximize correct identifications and minimize incorrect identifications by eyewitnesses. The sequential[3] lineup has been proposed as one such procedure (Lindsay & Wells, 1985). In contrast to the more traditional simultaneous lineup, in which all items are presented to the eyewitness at the same time, items in the sequential lineup are presented one at a time. Past research had suggested that the sequential lineup is superior to the simultaneous lineup because it leads to a reduced number of incorrect identifications without affecting the number of correct identifications (e.g. Wells et al., 2006), suggesting that memory for the perpetrator is expressed more efficiently in the sequential lineup. However, recent studies have drawn the opposite conclusion, finding that simultaneous presentation is superior (e.g. Clark, 2012a; Mickes et al., 2012). This raises the question of whether memory for the perpetrator is greater in the sequential lineup compared to the simultaneous lineup or vice versa. In order to answer this question, we argue that it is necessary to apply formal models specific to each procedure in order to measure underlying memory strength and response bias. Our aim in this paper is to develop such models and to apply them to both existing and new data to answer the question whether memory is the same or different between simultaneous and sequential lineups.

2.4.1 The Sequential Lineup

Lineups are typically presented simultaneously, with all lineup items shown at the same time in a single array. A witness may either identify an item as the target (i.e., corresponding to their memory of the perpetrator) or reject the lineup, indicating that no item is a suitable match. In a sequential lineup, as originally proposed by Lindsay and Wells (1985), each lineup item is presented one at a time and, for each item, the witness is asked to

[3]By convention, the term "sequential lineup" in lineup research refers to the sequential lineup conducted with a stopping rule proposed by Lindsay and Wells (1985). This is the terminology used in the published version of this manuscript.

judge if it matches their memory of the target by making a "yes/no" judgment. If the witness responds "yes", the procedure terminates and the remaining lineup items (if any) are not shown. If they respond “no”, they are shown the next lineup item if there is one. The lineup is implicitly rejected if the witness responds “no” to all available lineup members. Variations of this procedure have also been proposed which do not enforce the immediate stopping rule. These alternatives may permit witnesses to see remaining lineup members after an identification is made (Wilson et al., 2019), require witnesses to view all lineup members before making an identification, or allow (or require) witnesses to lap through the procedure a second time (Horry et al., 2015; Seale-Carlisle et al., 2019).

Lindsay and Wells (1985) originally proposed the sequential lineup based on a theoretical distinction between absolute and relative judgment strategies (Wells, 1984). A relative judgment is said to occur when a witness selects the lineup item most similar to their memory of the target *relative* to the other items. Such a strategy would tend to lead to a high false positive rate because there is a basis for identification even when memory for the perpetrator is poor or the target is not a member of the lineup. An absolute judgment is said to occur when an identification judgment does not depend on the similarity of other lineup items to the witness’ memory of the target. Such a strategy would tend to lead to lower false positive rates because witnesses have a basis to reject the lineup when memory for the target is poor or if the target is not present. Lindsay and Wells (1985) suggested that the sequential lineup would encourage an absolute decision strategy by removing the opportunity to compare lineup items. Consistent with this, Lindsay and Wells (1985) found that sequential presentation led to significantly fewer innocent suspect identifications than simultaneous presentation, accompanied by a relatively small reduction in target identifications. This pattern of results, termed the *sequential superiority effect*, has been found by many subsequent studies as well as by two meta-analyses (Steblay et al., 2001; Steblay et al.,

2011b). Based on this evidence, researchers have successfully advocated a policy shift toward sequential presentation, which has led to its adoption in various forms in 30% of US jurisdictions as well as in Canada and the United Kingdom (Police Executive Research Police Executive Research Forum, 2013; Seale-Carlisle & Mickes, 2016).

2.4.2 Diagnostic Feature Detection Theory

The interpretation of the sequential superiority effect has recently been challenged by Wixted and Mickes (2014). They have proposed the diagnostic feature detection theory (DFDT) of lineup identification, which predicts a memory advantage for simultaneous lineups compared to sequential lineups. According to this theory, correct identification (and rejection) of a lineup is based on identifying *diagnostic features* of the different lineup items. A diagnostic feature is one that is uniquely shared by a lineup item and the witness' memory of the target which, if identified, would support a correct identification. A non-diagnostic feature is one that is shared by all lineup items (e.g., hair colour) which, even if it matches the witness' memory of the target, cannot support a correct identification. Wixted and Mickes (2014) argued that because a witness is better able to compare the features of different lineup items in a simultaneous lineup, they are better able to identify features that are diagnostic and to discount those that are not.

The distinction between absolute and relative identification strategies proposed by Lindsay and Wells (1985) and DFDT make opposite predictions concerning the relative merits of simultaneous and sequential lineups – both cannot be correct. This has led to a re-evaluation of the sequential superiority effect and a re-examination of how eyewitness performance is measured. Specifically, researchers have argued that much of the early sequential lineup research has obscured potential shortcomings of the sequential procedure by treating the accompanying small reduction in perpetrator identifications as inconsequential (Clark, 2012a; Moreland & Clark, 2016). In addition, recent research, employing Receiver

Operating Characteristic (ROC) analysis derived from signal detection theory, has found evidence that simultaneous presentation may, in fact, outperform sequential presentation (e.g. Carlson & Carlson, 2014; Dobolyi & Dodson, 2013). We discuss each of these issues in turn.

2.4.3 Measuring Identification Performance

In many earlier studies of the sequential superiority effect, eyewitness performance was measured using the diagnosticity ratio statistic, defined as the ratio of the proportion of correct target identifications (TID rate) to the proportion of incorrect suspect identifications (SID or false positive rate). A TID is made when the witness correctly identifies the target in the lineup. An SID is made when the target is not a member of the lineup and the witness incorrectly identifies the innocent suspect. On this measure of performance, an identification made from a lineup procedure that reliably generates a higher diagnosticity ratio is to be preferred to one that does not.

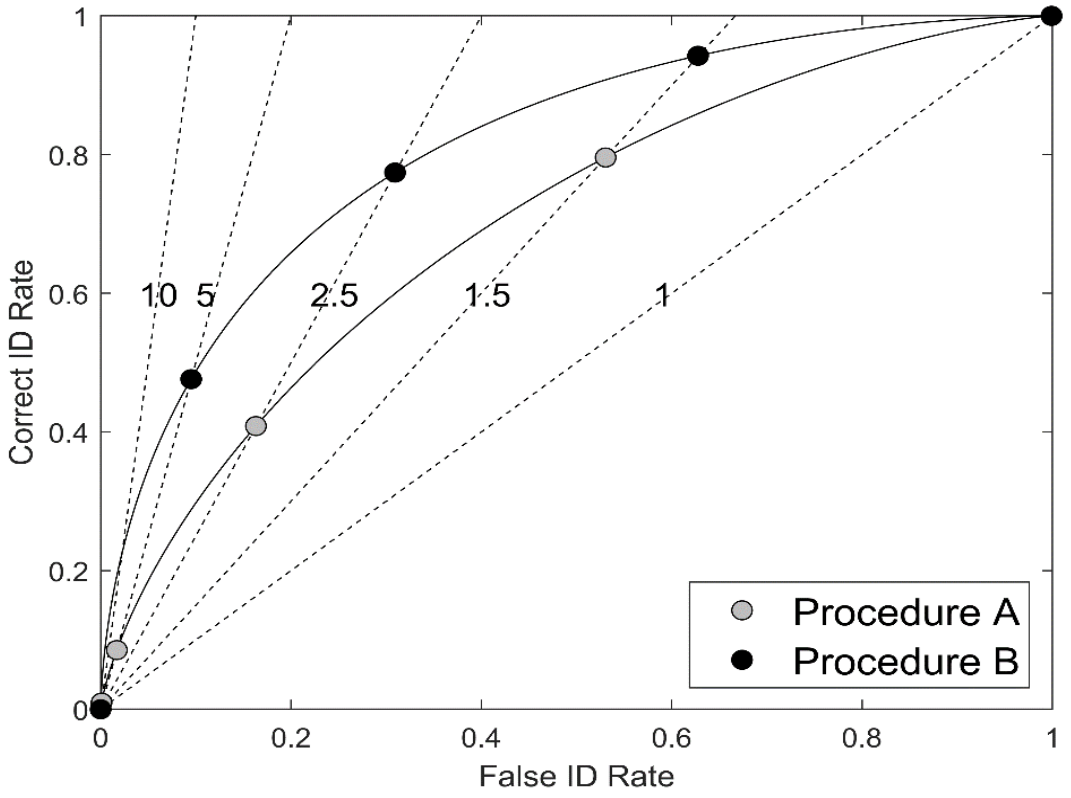
An alternative performance measure is based on signal detection theory (Wixted & Mickes, 2012; Wixted & Mickes, 2015a, 2015b) and proposes that performance should be judged in terms of the level of correct identifications that can be obtained for a given level of incorrect suspect identifications. This is termed *empirical discriminability* and it minimizes the two kinds of identification error discussed previously (Wixted & Mickes, 2018). Empirical discriminability can be measured by constructing a Receiver Operating Characteristic (ROC) curve. In the context of lineup tasks, this is a plot of TID rates against SID rates at different levels of *response bias* – the general willingness of a decision maker to make an identification. In perceptual research, different levels of response bias are achieved by varying payoffs that differentially weight correct and false positive responses, leading decision makers to be biased towards one kind of response over another. In many recognition memory experiments, post-decision confidence estimates are used as a proxy for different

levels of response bias. These may be recorded on a Likert scale or a 0-100% scale with the number of bins set by the researcher.

Figure 2.1 displays ROCs for two hypothetical show up procedures. A show up is a lineup consisting of only one item.

Figure 2.1

The Diagnosticity Ratio at different points in ROC Space for two Hypothetical Lineup Procedures



These ROC curves have the same form as found in laboratory-based yes-no recognition memory tasks, extending from the extreme lower left to extreme the upper right. The two curves in Figure 2.1 differ in empirical discriminability, which is greater for the curve that is closer to the top-left corner. This curve, corresponding to Procedure B in this example, always has a higher correct identification rate for any given incorrect identification rate. If empirical discriminability is zero, the ROC curve falls on the main diagonal indicating

chance performance. Following this logic, empirical discriminability can be measured by calculating the area under the ROC curve (AUC). The greater the AUC, the greater the empirical discriminability. The AUC measure is independent of response bias because any combination of correct and incorrect identification rates on the same ROC curve is associated with the same AUC. Accordingly, because Procedure B has greater AUC than Procedure A, it has greater empirical discriminability.

Each point on an ROC curve corresponds to a different response bias and is associated with a given diagnosticity ratio. It is here that the contrast between empirical discriminability and the diagnosticity ratio becomes apparent – the same ratio can be found on different ROC curves corresponding to different levels of discriminability (Gronlund et al., 2014; Rotello et al., 2015). This feature is shown in Figure 2.1 by the set of dashed lines each of which corresponds to a different diagnosticity ratio (either 1, 1.5, 2.5, 5, or 10). As can be seen, these lines intersect each of the two ROC curves at different points showing that, all else being equal, the more conservative is the response bias (associated with lower false positive rates), the larger is the diagnosticity ratio. It is clear from this that the diagnosticity ratio is simply a measure of response bias, independent of empirical discriminability.

2.4.4 Task Dependence of ROC Curves

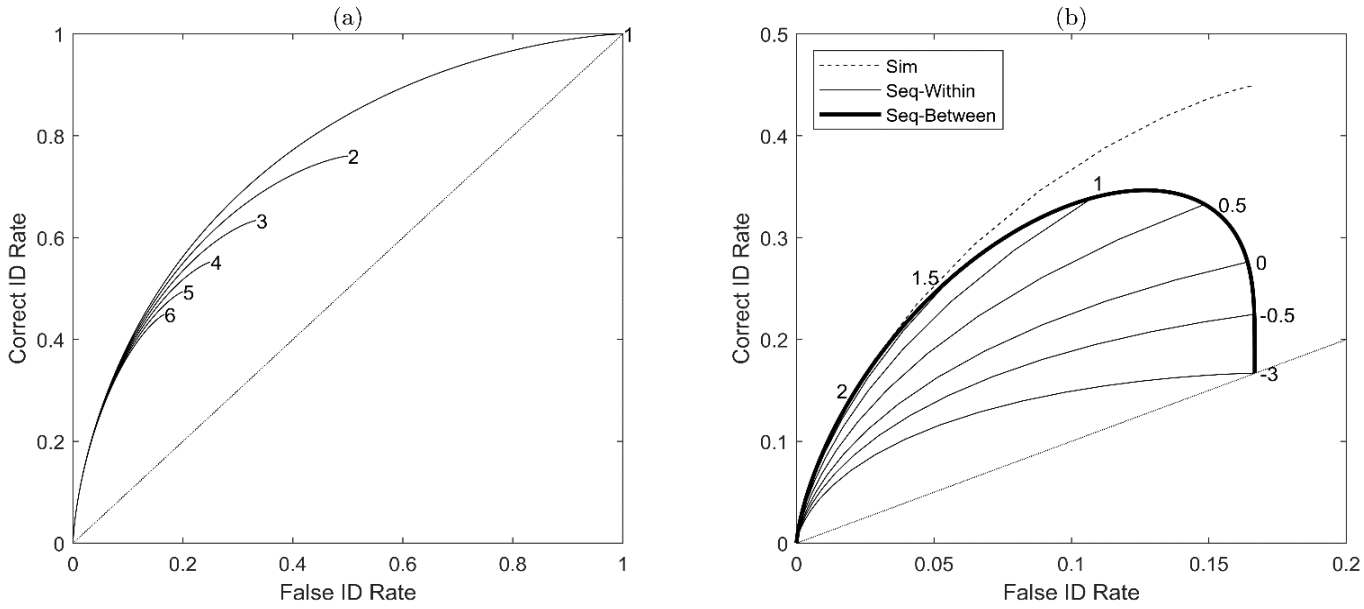
Empirical discriminability provides an objective criterion against which different lineup procedures may be compared. On this view, any procedure that leads a higher correct identification rate for any given false positive rate is to be preferred (Wixted & Mickes, 2012). However, DFDT is concerned with *underlying* discriminability, i.e. memory strength (Wixted & Mickes, 2018). It proposes that the feature detection mechanism facilitated by simultaneous presentation leads to greater underlying discriminability compared to sequential presentation, and that this explains the superior empirical discriminability for simultaneous presentation observed in some ROC studies (e.g. Carlson & Carlson, 2014). ROC analysis

may be uninformative with respect to underlying discriminability when the procedures being compared have different structural characteristics. In this case, the shapes of the ROC curves and the resulting empirical discriminability associated with each procedure may differ substantially even when underlying discriminability is the same (Rotello & Chen, 2016; Stephens et al., 2019).

A dissociation between empirical and underlying discriminability due to structural features of a task is illustrated in Figure 2.2 Panel A. This shows a family of hypothetical ROC curves derived from lineups of different sizes. These curves were generated using the simultaneous lineup model SDT-MAX which we define later (the relevant formulas are given in Appendix A).

Figure 2.2

The effect on ROC shape of changes to a) Lineup Size on the Simultaneous Lineup and b) Response Criterion on the Sequential Lineup



This model is based on a signal detection framework in which there is a normal distribution of familiarity values for the target item and another normal distribution for foil items, including the innocent suspect. For each lineup size, although underlying discriminability (i.e., the difference between the familiarity distributions of the target and foils) is the same, the shape and termination point of each ROC curve is different. Each curve terminates at a different point because, under the most lenient response bias (i.e., always select a lineup member) there is a 1/*n* chance of choosing the innocent suspect, where *n* is the lineup size. Thus, because *n* differs between the curves, each must terminate at a different point corresponding to a false positive rate of 1/*n*.

Because the ROC curves in Figure 2.2 Panel A were all generated from the same underlying signal detection model, the differences are due to a structural characteristic of the lineup task – specifically the lineup size. This means that differences in empirical discriminability between these tasks do not indicate differences in underlying discriminability (which is the same for each curve).

From the foregoing, it should come as no surprise that structural characteristics of the sequential lineup also change the shape of the ROC curve. In this case, it is not the size of the lineup that is critical, but the minimum level of evidence required to make an identification. Figure 2.2 Panel B shows a set of ROC curves for a sequential lineup of size six, each constructed with a different minimum level of evidence. The ROC curves shown by thin solid lines in Figure 2.2 Panel B illustrate different choices for the minimum level of evidence expressed in terms of a decision criterion on the familiarity axis. The value of this criterion is indicated at the end of each corresponding ROC curve. A large value indicates a conservative response bias for which a relatively high level of familiarity is required for a lineup item to trigger identification. A small value indicates a lenient response bias for which a relatively low level of familiarity is sufficient to trigger identification. Each of these ROC curves

terminate at different points. In the limit, when the minimum evidence is very low, the ROC curve terminates on the main diagonal (indicated by the dotted line in Figure 2.2 Panel B). The ROC curve shown by the thick solid line corresponds to the situation in which each witness has a different level of minimum evidence. It encloses the set of confidence-based ROC curves and is clearly non-monotonic. Rotello and Chen (2016) found a similar shaped curve in their simulations of the sequential lineup, as did Wilson et al. (2019) in empirical sequential lineup data.

Figure 2.2 Panel B also shows the ROC curve generated from a simultaneous lineup of size six as shown in Figure 2.2 Panel A (by the curve labelled 6). Altogether, these curves show that even when underlying discriminability is held constant, the shapes of ROCs and the corresponding empirical discriminability values differ to a considerable degree. It is therefore important to distinguish two research questions. One question is about empirical discriminability– for any given false identification rate, which procedure leads to higher correct identification rates? The ROC curves shown in Figure 2.2, suggest that simultaneous lineups are preferred to sequential lineups and, within the class of simultaneous lineups, smaller lineup sizes are preferred to larger lineup sizes. Empirical research also supports this conclusion, at least with respect to simultaneous, as compared to sequential lineups (Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Experiment 1a Mickes et al., 2012; Neuschatz et al., 2016), although this has not always been found (Flowe et al., 2016; Gronlund et al., 2012; Experiment 1b and 2 Mickes et al., 2012; Sučić et al., 2015).

The second question bears on DFDT and concerns underlying discriminability – which eyewitness test procedure reveals higher levels of memory strength? ROC curves and AUC cannot be used to answer this question. As shown above, they may not reflect underlying discriminability across different lineup procedures. In order to measure underlying discriminability, it is necessary to use a formal model to measure the parameter of

interest. In this section we outline two models of the simultaneous lineup task based on signal detection theory (SDT-MAX and SDT-INT) and develop a comparable model of the simple stopping rule version of the sequential lineup task (SDT-SEQ). We then apply these models to extant and new data to estimate memory strength across the two procedures.

2.4.5 Unequal Variance Signal Detection Model

The starting point for all the lineup models we consider is the unequal variance signal detection (UVSD) model. This common underlying framework ensures that parameters values estimated by each model are directly comparable. The UVSD model accounts well for data in laboratory-based recognition memory tests (Jang et al., 2009; Mickes et al., 2007) and can be extended to account for lineup tasks. In a typical eyewitness experiment, a participant views a simulated crime conducted by a perpetrator and is subsequently shown an *n*-item lineup. In a *target present* (TP) lineup, one item is the *target* (a picture of the perpetrator) and the remaining items are foils or fillers (pictures of other people). In a *target absent* (TA) lineup, one item may designated as the innocent *suspect* with the remaining items being foils. The participant is required to judge whether the lineup contains the target and, if they believe it does, to identify the corresponding item. We assume that each lineup item is associated with a familiarity value that reflects its similarity to the participant's memory of the perpetrator. Each familiarity value is considered a random draw from one of several distributions – a target distribution if the item is a target, an innocent suspect distribution if it is an innocent suspect[1], and a foil distribution if it is a foil. In order for the models to be testable we assume that each distribution is Gaussian. Consistent with most signal detection models, the foil distribution is assigned a mean of zero and a standard deviation of one. The target distribution has d_t and standard deviation s_t, both of which can be estimated from the data. Because s_t may not equal one the model is called the *unequal variance* signal detection

model. In addition, because the innocent suspect may be distinct from the remaining foils, the suspect distribution has mean d_s and standard deviation s_s.

A lineup can be considered as a combination of a detection question, "Is the target present?", and an identification, "If so, which item is the target?" (Duncan, 2006). While the answer to the identification question is relatively straightforward – always choose the lineup member associated with the greatest familiarity – the answer to the detection question is less clear-cut. This leads to different models based on different decision rules. Although there is a wide range of possible decision rules, we consider two in particular which we call SDT-MAX and SDT-INT. In the SDT-MAX model, the decision rule is to compare the familiarity value of the most familiar lineup item (the maximum) to a response criterion. In the SDT-INT model, the decision rule is to compare the *sum* of the familiarity values of the lineup items to a response criterion. For both of these models, if the relevant value exceeds the criterion, the most familiar item is identified as the target. We also develop a model of the sequential lineup. In this case, because the witness does not see all the lineup items until the end, and may not see all items if they choose before reaching the end, it not possible before that point to identify either the maximum or the sum, or any other function of the familiarity values of the entire lineup. For this reason, we develop a model of the sequential lineup, here called SDT-SEQ.

2.4.5.1 SDT-MAX

SDT-MAX, also known as the Independent Observations model (Duncan, 2006; Wixted et al., 2018), is perhaps the simplest model of the simultaneous lineup. In this model, identification decisions are made with respect to a set of *k* decision criteria, $C = \{c_1, \ldots, c_k\}$ such that $c_1 < c_2 < \ldots < c_k$, that define a set of $k+1$ confidence levels. Let $X = \{x_1, \ldots, x_n\}$ be the set of familiarity values associated with each of *n* lineup items and let *m* be the item

number such that $x_m = \max(X)$. The decision rule is this: If $x_m < c_1$ then reject the lineup, otherwise choose lineup item *m* with confidence level *l* where c_l is the largest element of the set, $\{c_i \in C : x_m \geq c_i\}$.

As detailed in Appendix A, we derive general formulas for the probability of a correct identification and the probability of a false identification under the SDT-MAX model. We summarize these below under the assumption that all the underlying target and foil distributions are Gaussian. Let $\phi(x;\mu,\sigma)$ be the normal probability density function and let $\Phi(x;\mu,\sigma)$ be the normal cumulative distribution function evaluated at $x \in \mathbb{R}$. Recall that the foil distribution takes the form of the standard normal distribution with $\mu = 0$ and $\sigma = 1$. In this case, we write $\phi(x;0,1) = \phi(x)$ and $\Phi(x;0,1) = \Phi(x)$. Let $c \in C$ be a decision criterion and let $P_{TID}(c)$ be the probability of a correct target identification with confidence greater than or equal to *c*. Then

$$P_{TID}(c) = \int_c^{\infty} \phi(x;d_t,s_t)\Phi(x)^{n-1}\,dx\,.$$

Similarly, let $P_{SID}(c)$ be the probability of an incorrect suspect identification with confidence greater than or equal to *c*. Then, if there is a designated innocent suspect,

$$P_{SID}(c) = \int_c^{\infty} \phi(x;d_s,s_s)\Phi(x)^{n-1}\,dx\,,$$

otherwise,

$$P_{SID}(c) = \frac{1}{n}\left(1-\Phi(c)^n\right).$$

2.4.5.2 SDT-INT

Let $\text{sum}(X)$ be the sum of familiarity values of all the lineup items. The decision rule is this: If $\text{sum}(X) < c_1$ then reject the lineup, otherwise choose lineup member *m* with confidence level *l* where c_l is the largest element of the set, $\{c \in C : \text{sum}(X) \geq c\}$.

The equations for the probability of a correct identification and probability of a false identification under the SDT-INT model are summarized below (see Appendix A for details).

$$P_{TID}(c) = \Pr(\text{sum}(X) \geq c \mid m = t) \cdot \Pr(m = t)$$
$$\approx \int_{-\infty}^{\infty} \left(1 - \Phi\left(c - x; (n-1)\mu_x, \sqrt{(n-1}\sigma_x\right)\right) \phi(x; d_t, s_t) \Phi(x)^{n-1} dx$$

where *t* is the position of the target item and μ_x and σ_x are the mean and standard deviation, respectively, of the standard normal distribution truncated at the upper limit of *x*. The equation is not exact because it assumes that the sum of $n-1$ truncated distributions is approximately normal (by the Central Limit Theorem). Similarly, if there is a designated innocent suspect, then

$$P_{SID}(c) \approx \int_{-\infty}^{\infty} \left(1 - \Phi\left(c - x; (n-1)\mu_x, \sqrt{(n-1)}\sigma_x\right)\right) \phi(x; d_s, s_s) \Phi(x)^{n-1} dx,$$

otherwise,

$$P_{SID}(c) = \frac{1}{n}\left(1 - \Phi\left(c; 0, \sqrt{n}\right)\right).$$

2.4.5.3 SDT-SEQ

Our model for sequential presentation is also based on the UVSD framework and incorporates the "first-above-criterion" decision rule where presentation of the lineup items is terminated as soon as an identification is made. As detailed in Appendix A, we derive the following equations for the probability of a correct identification and probability of a false identification under the SDT-SEQ model. Let p_i be the probability that the lineup item at position *i* is a target. Then

$$P_{TID}(c) = \left(1 - \Phi(c; d_t, s_t)\right) \sum_{i=1}^{n} p_i \Phi(c_1)^{i-1} \quad .$$

If there is a designated innocent suspect, let q_i be the probability that the lineup item at position *i* is the suspect. Then,

$$P_{SID}(c) = \left(1 - \Phi(c; d_s, s_s)\right) \sum_{i=1}^{n} q_i \Phi(c_1)^{i-1} ,$$

otherwise,

$$P_{SID}(c) = \frac{1}{n}\left(1 - \Phi(c)\right) \sum_{i=1}^{n} \Phi(c_1)^{i-1} .$$

2.4.6 Palmer and Brewer (2012) Database

Palmer and Brewer (2012) conducted an extensive analysis of previously published studies that compared simultaneous and stopping-rule sequential lineups under the same conditions. They fit a signal detection model equivalent to the SDT-INT model described previously, to data from 22 previous studies. Their aim was to determine if either underlying discriminability and/or response bias differs between sequential and simultaneous lineups. Their analysis revealed that, across the datasets, the two presentation methods did not differ in terms of underlying discriminability but that the sequential procedure was associated with more conservative responding.

While the finding of equal underlying discriminability is not consistent with DFDT, the difference in response criteria was consistent with the view that a sequential lineup produces a higher diagnosticity ratio. It is now widely accepted that sequential presentation leads to more conservative responding than simultaneous presentation (Clark, 2012a; Clark et al., 2014; Wells, 2014; Wixted & Mickes, 2014). The apparent success of the modelling approach employed by Palmer and Brewer (2012) has also led researchers to use SDT-INT to examine other aspects of the sequential lineup (Carlson et al., 2016; Horry et al., 2015; Horry et al., 2012b).

However, there are aspects of the Palmer and Brewer's (2012) approach that challenge the validity of their conclusions. First, and most critically, the SDT-INT model was fit to data from both simultaneous and sequential lineups. No attempt was made to model the unique task demands of sequential presentation. It is therefore unknown whether the same results would be found if a more appropriate model were used, such as SDT-SEQ described previously. Second, the SDT-INT model does not exhaust the set of decision rules for simultaneous lineups (Wixted et al., 2018). A different decision rule, such as SDT-MAX, may lead to different results. Third, Palmer and Brewer (2012) fit the SDT-INT model using an inefficient and potentially inaccurate manual grid search of parameter space. Finally, because confidence judgments were not available, it was only possible to fit an equal variance signal detection model in which $s_t = s_s = 1$. If this is not an appropriate model of their data, the results may be distorted.

2.4.7 Summary and Aims

The aim of the present paper is to compare simultaneous and sequential lineups in order to test the central prediction of DFDT that simultaneous presentation is associated with greater underlying discriminability than sequential presentation. To do this, we first reanalysed the corpus of simultaneous and sequential data from Palmer and Brewer (2012), addressing the previously described problems in their analysis. Principally, we fit a model of the sequential lineup, SDT-SEQ, specifically developed for this task, as well as two models of the simultaneous lineup - the SDT-INT model as used by Palmer and Brewer (2012) and the alternative SDT-MAX model. Third, we fit each model using an efficient optimisation procedure that leads to more accurate solutions. Second, we conducted a new experiment from which we obtained confidence judgments enabling us to fit models based on the assumption of unequal variances.

2.4.8 Predictions

Predictions were preregistered on the Open Science Framework, available at https://osf.io/xwp9d/. DFDT predicts that simultaneous presentation should lead to greater underlying discriminability than sequential presentation. Specifically, this means that the estimate of d_t (or the difference $d_t - d_s$ if there is a designated suspect) should be greater for simultaneous lineups. Based on the conclusions reached by Palmer and Brewer (2012), sequential presentation is predicted to lead to more conservative responding than simultaneous presentation. This means that the estimate of c_1 (and possibly other criteria) should be greater for sequential lineups.

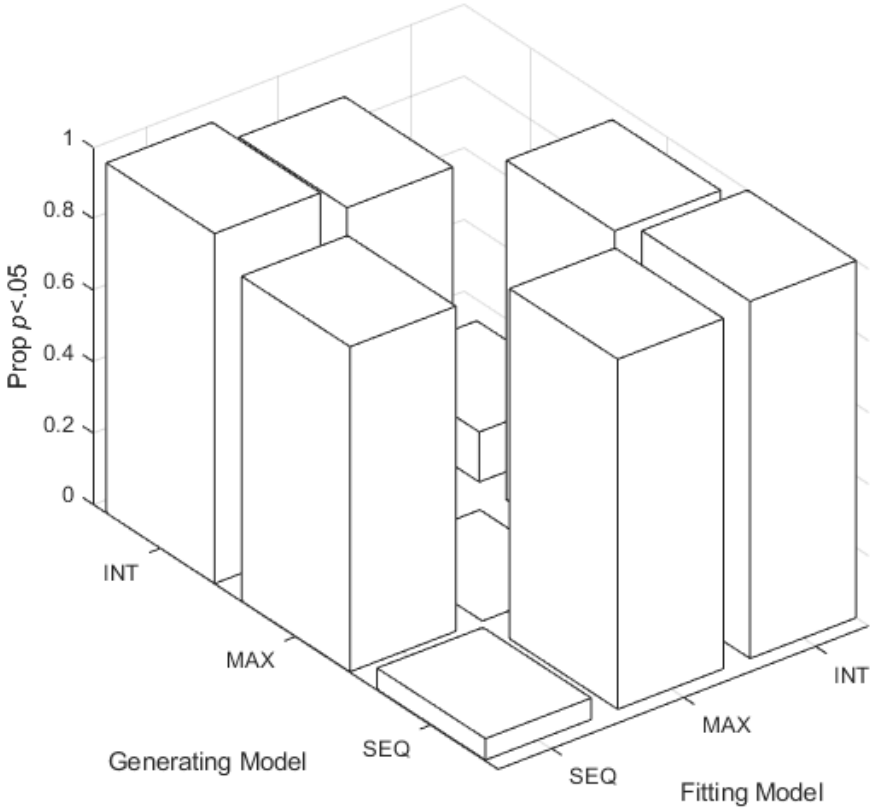
2.5 Model Cross Fit

We have described three models that we propose to fit to data. This is motivated in part by the idea that there are differences between the models that determine how well they fit different kinds of data. This means that if data is simulated from a model, while this model should fit the data well, other models should fit relatively poorly. In order to investigate this question, we conducted a cross fitting and parameter recovery analysis. First, we randomly generated 100 sets of parameter values for a 6-item lineup and then used each of these to generate 100 simulated datasets from each model. To avoid issues with low cell counts, we set the number of TP and TA lineups to 10,000, giving 20,000 simulated observations for each dataset. We then fit each model to its own sets of data and to those generated by the other models, recording the χ^2 value, p value and parameter estimates from each fit. Further detail regarding the simulation process and expanded results are available in Appendix B.

Figure 2.3 shows the proportion of datasets where the model could be rejected at $p < .05$. It shows that when a model is fit to data generated by any other model, it is highly likely to be rejected. In other words, the models are in principle distinct – given sufficient statistical power, if the data are consistent with one model then they should be poorly fit by any of the remaining models.

Figure 2.3

Results of Cross Fitting Models to Simulated Data



2.6 Parameter Recovery

We measured parameter recovery by examining the correlation between generating and recovered parameter values for each model fit. Scatterplots and tables of correlations are available in Appendix B. We were interested in two aspects of this analysis. First, it is desirable for the correlation to be close to one when the models are fit to their own data. Second, it is also important to understand how well the models recover the correct parameter

values when fit to data they did not generate as, in some cases, they may fit well but recover incorrect parameter estimates.

When fit to their own data, the models generally recover their own parameters well, with $r >= .90$ for generating vs. recovered parameter values. SDT-MAX recovers the generating parameters perfectly when fit to its own data, but both SDT-SEQ and, to a lesser extent, SDT-INT, recover a small number of outliers, affecting the correlation coefficients. These are most likely due to the presence of local minima which can be avoided by starting parameter search from different initial values. It is evident from the scatterplots in Appendix B that recovery is close to perfect once these outliers are excluded.

When SDT-MAX and SDT-INT are fit to data generated by SDT-SEQ, recovery of d_t is poor. This suggests that if SDT-SEQ is a good representation of the sequential lineup task, then fitting SDT-MAX or SDT-INT to sequential lineup data may lead to inaccurate estimates of d_t. Recovery of s_t was poor for all models when fit to data they did not generate, while recovery of the decision criteria ($c_1, \cdots, c_5$) was generally good for all fits, with $r >= .80$.

2.7 Reanalysis of the Palmer and Brewer (2012) Dataset

Palmer and Brewer (2012) selected a corpus of 22 studies (total $N = 3871$, simultaneous $n = 1952$, sequential $n = 1919$) that compared simultaneous and stopping-rule sequential presentation procedures using the "full diagnostic design" inclusion criteria described in Steblay et al. (2011b). That is, each study manipulated both presentation format (simultaneous vs. sequential) and target presence (present vs. absent), reported above-chance identification performance, defined as $P_{TID} - P_{SID} > 0.1$, in at least one of the four experimental conditions, and used only adult participants.

The simultaneous lineup data from each study were fit by both SDT-INT (as undertaken by Palmer and Brewer) and SDT-MAX. The corresponding sequential lineup data

were fit by SDT-INT (as undertaken by Palmer and Brewer) and SDT-SEQ. Each model was fit using the Matlab® fmincon function. Because each study required participants to make a single choose-no choose decision, there are not enough degrees of freedom to fit all of the model parameters, specifically c, d_t, d_s, s_t, and s_s, without the model becoming saturated (i.e., having no remaining degrees of freedom). Accordingly, we specified that $s_t = s_s = 1$, as was also assumed by Palmer and Brewer.

Some studies designated an innocent suspect while others did not. When a suspect had been designated, we estimated d_s, the mean of the suspect distribution, otherwise we stipulated that $d_s = 0$, the same as the mean of the foil distribution. In addition, studies differed in the probability of a target (and suspect if relevant) appearing at different sequential lineup positions. When specified, this information was used in fitting the SDT-SEQ model (see Appendix A for details), otherwise it was assumed that the target/suspect had the same probability of appearing at each lineup position.

2.8 Results and Discussion

2.8.1 Model Fit Performance

Table 2.1 presents the χ^2 goodness-of-fit values for each dataset and each fitted model. Each χ^2 test has one degree of freedom and we set $\alpha = .01$ to control the Type I error rate across the large number of tests conducted. We fit the SDT-MAX and SDT-INT models to the simultaneous lineup data and SDT-SEQ and SDT-INT to sequential lineup data. SDT-MAX fit 20 of 22 simultaneous data sets, as indicated by non-significant χ^2 values. The model did not fit data from two studies - Carlson et al. (2008) experiment two and Greathouse and Kovera (2009). SDT-INT performed similarly, also failing to fit the two studies above, in addition to Lindsay and Wells (1985). For the sequential lineups, SDT-SEQ fit 19 of 22 data sets, failing to fit data from Kneller et al. (2001), Lindsay and Wells (1985) and Pozzulo and Marciniak (2006). SDT-INT failed to fit the three datasets above, in addition

to experiment one and two from Carlson et al. (2008). In all, SDT-MAX and SDT-SEQ performed better than SDT-INT when fit to data from simultaneous and sequential lineups respectively. Similar results with respect to simultaneous lineup data were found by Wixted et al. (2018), who examined the performance of SDT-MAX and SDT-INT by fitting these models to a number of previous lineup datasets.

Table 2.1

χ^2 fit values for each dataset, presentation format, and model

	Simultaneous Lineup		Sequential Lineup	
Dataset	SDT-MAX	SDT-INT	SDT-SEQ	SDT-INT
Carlson et.al (2008, Exp 1)	0.01	2.29	2.01	11.48*
Carlson et.al (2008, Exp 2)	20.81*	36.53*	0.23	30.04*
Clark & Davey (2005, Exp 1)	0.39	0.08	0.06	0.05
Clark & Davey (2005, Exp 2)	0.30	0.03	1.14	0.51
Greathouse & Kovera (2009)	9.23*	10.28*	2.91	0.01
Kneller et al (2001)	2.68	3.20	10.91*	13.17*
Levi (2006)	0.08	0.72	0.17	0.10
Lindsay, Lea, & Fulford (1991)	1.24	1.39	0.17	4.99
Lindsay & Wells (1985)	5.99	11.86*	6.74*	22.25*
MacLin & Phelan (2007)	0.37	0.21	0.02	0.00
MacLin et al (2005, Exp 1)	0.25	0.22	1.41	1.39
MacLin et al (2005, Exp 2)	0.61	0.46	0.00	0.03
Melara et al (1989)	1.14	1.18	0.07	0.01
Memon & Gabbert (2003)	0.31	0.48	0.05	0.34
Parker & Ryan (1993)	1.38	4.33	0.00	0.27
Pozzulo et al (2008)	0.03	0.00	0.00	0.06
Pozzulo & Marciniak (2006)	0.09	0.03	12.18*	13.75*
Rose et al (2005)	0.49	1.62	0.01	0.10
Sporer (1993)	0.66	0.63	0.63	0.44
Steblay et al (2011)	0.72	1.24	0.00	0.07
Wells & Pozzulo (2006)	0.47	0.24	0.59	0.73
Wilcock et al (2005)	5.34	5.56	0.02	0.17

Note: Asterisks indicate a significant difference from zero, $\alpha = 0.01$ (critical value = 6.63).

Table 2.2

χ^2 fit values for Previously Non-fitting Datasets, disaggregated in to Original Experimental Conditions

	Simultaneous Lineup		Sequential Lineup	
Dataset	SDT-MAX	SDT-INT	SDT-SEQ	SDT-INT
Carlson et.al (2008, Exp 2) – biased	19.68*	19.66*	1.94	22.76*
Carlson et.al (2008, Exp 2) – intermediate	.81	3.02	.42	10.23*
Carlson et.al (2008, Exp 2) – fair	10.00*	16.88*	.85	2.61
Greathouse & Kovera (2009) – biased, single-blind	.15	0.38	.78	0.15
Greathouse & Kovera (2009) – biased, double-blind	.57	0.38	4.08	2.17
Greathouse & Kovera (2009) – fair, single-blind	5.44	6.29	4.06	0.44
Greathouse & Kovera (2009) – fair, double-blind	3.83	4.09	.25	0.15
Pozzulo & Marciniak (2006) – no appearance change	.89	0.09	1.82	3.81
Pozzulo & Marciniak (2006) – appearance changed	.21	0.06	12.60*	11.58*

Note: Asterisks indicate a significant difference from zero, $\alpha = 0.01$ (critical value = 6.63).

We examined the datasets that were not fit by one or more models. Our first observation was that each of these contained a limited number of observations, although this was also true for other datasets that were fit well. Second, in the case of Carlson et al. (2008, Exp 2), Greathouse and Kovera (2009) and Pozzulo and Marciniak (2006), Palmer and Brewer (2012) had collapsed the relevant data across different experimental conditions. In addition to presentation format, Carlson et al. (2008, Experiment 2) manipulated lineup fairness, Greathouse and Kovera (2009) manipulated administrator bias and lineup fairness, and Pozzulo and Marciniak (2006) manipulated appearance change from encoding to test.

Given that these manipulations may have affected the underlying signal detection parameters and that collapsing across these conditions may have caused the models to perform poorly, we disaggregated each dataset in to its original experimental conditions and re-fit the models to these datasets. The resulting χ^2 values are shown in Table 2.2, revealing improved model fits in 10 of 18 experimental conditions.

2.8.2 Parameter Estimates

In order to compare our results with Palmer and Brewer's (2012), we report parameter values recovered from fitting the models to the same 22-dataset corpus, rather than disaggregating each study in to its original experimental conditions. A full table of parameter estimates is available in Table 1, Appendix C. Table 2.3 shows the mean estimates of the model parameters and their standard deviations for each presentation format, weighted by sample size. The parameters are underlying discriminability, decision criterion, *c*, and a derived decision parameter *C*, which Palmer and Brewer (2012) used in their original analysis. *C* is defined as, $C = c - d_t / 2$, with zero indicating an "unbiased" criterion set at the midpoint between the target and foil distributions. Negative values indicate a lenient response criterion while positive values indicate a conservative criterion. This metric is only relevant in the equal variance case, as a change in target distribution variance will shift the point at which choosing would be truly unbiased. Our hypothesis tests are based on the estimated parameters from fitting SDT-MAX fit to the simultaneous data and SDT-SEQ fit to the sequential data. Mean weighted parameter values from fitting SDT-INT to both data types and as calculated from the original Palmer and Brewer (2012) fits are presented for comparison.

2.8.2.1 Underlying Discriminability

Figure 2.4 shows underlying discriminability plotted against criterion *c* estimated by SDT-MAX and SDT-SEQ fit to simultaneous and sequential lineups respectively. For studies

that specified a designated innocent suspect, underlying discriminability was calculated as $d_t - d_s$. Visual examination of Figure 2.4 reveals no particular relationship between underlying discriminability and presentation format. Mean weighted underlying discriminability shown in Table 2.3 does not differ between simultaneous and sequential presentation, as indicated by a Welch two-sample weighted *t*-test, $t(40.33) = -.40$, $p = .69$. We re-ran the analysis, excluding datasets that the models failed to fit, but this did not change the result. This result is consistent with the conclusion reached by Palmer and Brewer (2012) and fails to support our hypothesis that underlying discriminability is greater for simultaneous presentation.

Figure 2.4

Discriminability Plotted Against Criterion for All Studies in the Palmer and Brewer Corpus

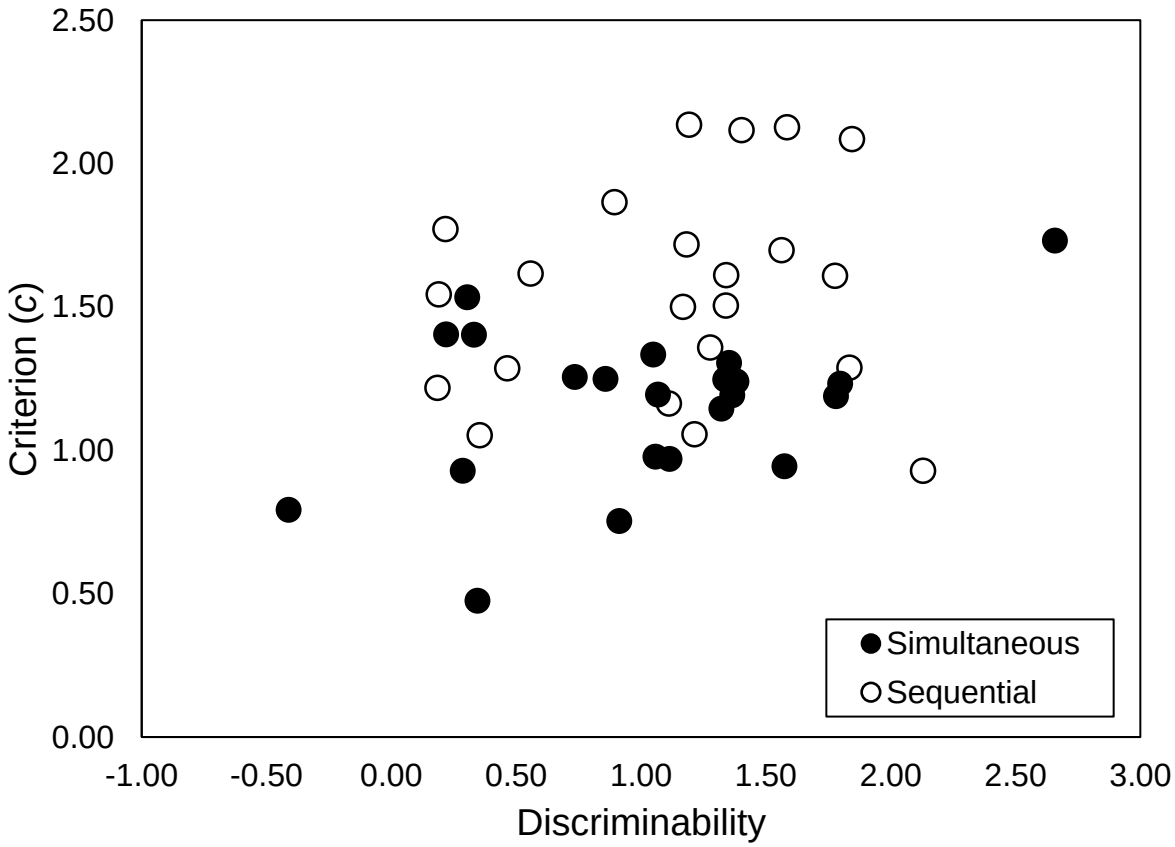


Table 2.3 shows that the mean-weighted estimates of underlying discriminability recovered by SDT-INT for each presentation format are similar to those recovered by SDT-

MAX and SDT-SEQ when fit to their respective data types. Welch two-sample weighted *t*-tests indicated that there is no significant difference for simultaneous, $t(37.53) = .08$, $p = .94$, or sequential presentation, $t(35.22) = -.26$, $p = .79$.

Table 2.3

Mean Weighted Parameter Estimates for the Palmer and Brewer Corpus

		Parameter					
		discriminability		*c*		*C*	
Format	Source	μ_w	σ_w	μ_w	σ_w	μ_w	σ_w
Simultaneous	Palmer & Brewer (2012)	1.64	.50	-0.07	.37	-0.89	.33
	SDT-MAX	.91	.72	1.24	.24	0.58	.25
	SDT-INT	.94	1.02	-0.17	.82	-1.01	.72
Sequential	Palmer & Brewer (2012)	1.75	.62	0.48	.59	-0.38	.49
	SDT-SEQ	.99	.58	1.61	.37	0.92	.39
	SDT-INT	.93	.93	1.07	1.37	0.18	1.25

Our estimates of mean weighted underlying discriminability shown in Table 2.3 are less than those calculated from Palmer and Brewer's (2012) original analyses and those reported in our preliminary analysis of this corpus (Kaesler et al., 2017). This is because we estimated d_s for studies that employed a designated innocent suspect selected to resemble the perpetrator more closely than foils, where previous analyses assumed that the innocent suspect and the foils were drawn from the same distribution with a mean of zero and standard deviation of one. In the case where d_s is greater than zero, DFDT does not predict a strong simultaneous advantage, because the features uniquely shared by the innocent suspect and the perpetrator will cause the innocent suspect to be identified at a higher rate in the simultaneous procedure compared to the sequential procedure. For this reason, we examined whether there was a simultaneous advantage in the subset of studies that did not use an innocent suspect. We found that the mean weighted difference in underlying discriminability between

simultaneous and sequential presentation as estimated by SDT-MAX and SDT-SEQ respectively was less for the eight studies that used an innocent suspect (M = -.23) compared to the fourteen that did not (M = .09). However, a Welch two-sample weighted t-test indicated that the difference between these means is not significant, $t(9.78) = -1.61$, $p = .14$.

2.8.2.2 Response Bias

Visual examination of Figure 2.4 shows an apparent difference between sequential and simultaneous datasets for values of the decision criterion, c. Analysis of mean weighted c values show that these are greater (indicating more conservative responding) for sequential than for simultaneous lineups, Welch two-sample weighted t-test, $t(35.83) = -3.88$, $p < .01$. Once again, excluding the datasets that the models failed to fit did not change the result.

2.8.3 Summary

The reanalysis of Palmer and Brewer's (2012) corpus of data reaffirmed their original finding of no significant difference in underlying discriminability between sequential and simultaneous presentation. SDT-MAX and SDT-SEQ performed similarly and recovered similar parameter estimates to SDT-INT when fit to their respective data types. This is in contrast to simulations we conducted that showed that the models behave differently over the entire parameter space. Both of these results may be attributable to low statistical power since each study on average had less than 100 participants. It is possible that, because of the relatively small number of participants in each study, each individual analysis lacked the statistical power to detect both differences in the fits of models and differences in underlying discriminability between simultaneous and sequential lineups.

In addition to a lack of statistical power, two other methodological issues limit the utility of the corpus for investigating differences in underlying discriminability. First, a designated innocent suspect was selected to resemble the perpetrator in some studies which may attenuate any simultaneous advantage in underlying discriminability. Second, the target

was fixed to appear in certain positions in many of the sequential lineup studies. While our modelling approach accounted for fixed target positions, there is some evidence to suggest that underlying discriminability may increases with target position (Wilson et al., 2019). As a result, in those studies in which the target was fixed to appear late in the lineup may have overestimated underlying discriminability compared to studies in which target position was randomised. In addition, because each study either recorded or reported only a binary (yes/no) decision, it was necessary to assume an underlying equal variance signal detection model. Although the resulting model fits were good, it is possible that the parameter estimates may have been systematically affected. For these reasons, we conducted a new experiment that sought to address each of these limitations.

2.9 Experiment 1

The aim of Experiment 1 is to compare a simultaneous lineup and a stopping-rule sequential lineup, extending the studies examined by Palmer and Brewer (2012) by increasing statistical power by using a large sample size, collecting confidence judgments, and avoiding using a designated innocent suspect.

2.9.1 Design

We employed a 2 x 2 between-participants factorial design, manipulating presentation format (simultaneous vs. sequential) and target presence (TP vs. TA).

2.9.2 Participants

Participants were n = 600 Amazon Mechanical Turk workers who were compensated 1 USD for the five- to ten-minute experiment. Eleven participants were excluded for failing attention check questions relating to the content of the stimulus video, leaving n = 589 participants (simultaneous TP = 139, simultaneous TA = 141, sequential TP = 161, sequential TA = 148) for the eventual analysis.

2.9.3 Materials

This study employed a pool of sixteen female lineup members, drawn from the Adelaide Lineup Database. This consists of a video and accompanying head-and-shoulders photographs taken front-on, at 90 degrees side-on, and approximately at 45 degrees for each of 194 persons. Only front-on photos were used in this study. In each video, the actor wears a black shirt t-shirt with a white University of Adelaide logo to remove the identifying potential of coloured clothing in the lineup phase. The scene opens with an actor (each of the 194 persons in turn) seated at computer with their back to the camera. After a few seconds during which they type on the computer keyboard, the actor picks up a mobile phone placed on the table to their left and turns to face the camera while looking at the phone. The actor then stands and walks towards the camera while looking at the phone, glancing up briefly to the camera as they pass by. Each video is approximately 10 to 15 seconds in duration. An example video can be found at https://osf.io/p2hck/.

2.9.4 Stimulus Pool Selection Process

In order to minimise the potential for stimulus effects, rather than a single target and set of foils, we used a pool of lineup members that could all act as both targets and foils. The starting point for selecting the pool members was similarity ratings previously collected for front-on photographs of ninety female faces in the Adelaide Lineup Database. Amazon Mechanical Turk workers (n = 76) were compensated USD 1.30 to rate 45 pairs of faces on a Likert scale from zero, most similar, to ten, least similar. Each participant rated a different subset of the possible face pairs to reduce participant burden and ensure timely collection of the data. The average number of ratings per similarity pair was 5.92, minimum 1, maximum 10. This resulted in a similarity matrix with each cell containing the mean similarity rating between each pair of faces.

We first summed across each row of the similarity matrix, giving the mean similarity of a face relative to all other faces. Faces were then sorted from most similar to all others to least similar to all others. While this ordering served as a guide, we also identified a set of feature-based exclusion criteria, some of which related to distinctive non-biological features that appear in the photographs and others that related to constraints in terms of isolating a suitably large feature-matched subset from within the corpus. We excluded participants with nose rings or other obvious piercings, those wearing glasses, those who were not Caucasian in appearance, those with "unnaturally" dyed hair, e.g. blue, those with hair shorter than shoulder length and those with their hair pulled back. This resulted in a pool of sixteen lineup female members of a similar ethnicity, skin tone, hair colour and hairstyle. One of the stimulus photographs required some editing to remove distinctive clothing features that were not obscured by the black t-shirt worn by all actors.

2.9.5 Procedure

The entire procedure took place within Amazon Mechanical Turk (AMT), with the experiment rendered on the participants' web browsers. Participants were allocated to one of the four conditions on a round robin basis. They were first questioned on their understanding of the task, being directed back to the instruction page if incorrect responses were recorded. They were then shown a video of a target randomly selected from the sixteen-member pool, before completing a visual search distractor task, similar in nature to a "Where's Waldo/Wally". Participants were then shown pre-lineup instructions corresponding to those in the U.S. National Department of Justice (1999) guidelines before viewing either a target present or target absent lineup presented simultaneously or sequentially, with the appropriate number of foils (5 for target present, 6 for target absent) randomly selected from the remaining fifteen members of the stimulus pool. The position of the target on target present

lineups and the order of the foils on both target present (TP) and target absent (TA) lineups was randomised.

In the simultaneous condition, participants could either identify a lineup item or choose a black silhouette to indicate that the target was not present in the lineup, after which they provided a confidence rating for their choice by typing a number from zero to one hundred, where zero was lowest confidence and one hundred was highest confidence. In the sequential condition, participants were shown each lineup item individually with an option either to identify or to reject it. If the item was rejected, the next item in the sequence was shown. If a lineup item was identified, the procedure terminated and the participant was asked to provide a typed confidence estimate for their identification. If all lineup items were rejected, participants were informed that the lineup had been exhausted, indicating a rejection decision, and were asked for a typed confidence rating. Participants then answered follow-up questions about instruction clarity and task difficulty, and were given the opportunity to provide feedback.

2.9.6 Analyses

We fit SDT-MAX to the simultaneous data and fit SDT-SEQ to the sequential data, estimating seven parameters, d_t, s_t, and $c_1, \ldots, c_5$, for each dataset. In Supplement 1 we provide annotated R code for fitting a multi-criteria, unequal variance version of SDT-MAX to simultaneous lineup data.

We tested our hypotheses using likelihood-ratio tests, comparing an unconstrained model to seven nested models where an equality constraint across the simultaneous and sequential data was imposed for one or the other parameter. The unconstrained models for the likelihood-ratio tests were an omnibus fit of SDT-MAX to the simultaneous data and SDT-SEQ to the sequential data, minimising $\chi 2$ for the overall fit to both datasets rather than

fitting each dataset separately. This allowed us to specify equality constraints across both conditions.

2.9.7 Results and Discussion

Table 2.4 shows the decision outcome frequencies for simultaneous and sequential lineups. The bin widths were set by collapsing over all conditions and partitioning the confidence judgments in to even-as-possible frequency quintiles. We used an alpha level of .05 for the model fits and hypothesis tests.

Table 2.4

Decision Outcome Frequencies for Simultaneous and Sequential Presentation

Simultaneous						
Confidence	100-91	90-81	80-66	65-51	50-0	Reject
TP – Target ID	24	25	30	9	11	19
TP – Foil ID	0	1	5	4	11	
TA – Foil ID	4	11	25	16	24	61

Sequential						
Confidence	100-91	89-81	80-66	65-51	50-0	Reject
TP – Target ID	32	22	21	13	6	41
TP – Foil ID	0	3	7	9	7	
TA – Foil ID	3	5	31	11	14	84

2.9.7.1 Model Fit Performance and Parameter Estimates

Table 2.5 shows the recovered parameter values and fit statistics for SDT-MAX and SDT-INT fit to the simultaneous data and SDT-SEQ fit to the sequential data. For the simultaneous condition, both SDT-MAX and SDT-INT fit the data well. For the sequential condition, SDT-SEQ provided an adequate fit to the data. Table 2.5 shows that simultaneous and sequential s_t are similar when SDT-MAX is the simultaneous lineup model. This means that the d_t values for each presentation format are comparable estimates of underlying discriminability. In contrast, s_t is twice as large for simultaneous presentation compared to sequential presentation when SDT-INT is the simultaneous lineup model. In this case, the d_t

values for each presentation format cannot be interpreted as directly comparable estimates of underlying discriminability. This is because, holding all else equal, increasing s_t increases the area of overlap between the target and foil distributions, reducing underlying discriminability.

Table 2.5

Estimates and from Fitting SDT-MAX and SDT-INT to the Simultaneous Data and SDT-SEQ to the Sequential Data from Experiment 1

	Simultaneous		Sequential
	SDT-MAX	SDT-INT	SDT-SEQ
d_t	1.83	2.56	1.89
s_t	.94	2.02	1.12
c_5	2.72	5.17	2.74
c_4	2.20	3.41	2.27
c_3	1.69	1.56	1.74
c_2	1.49	.79	1.54
c_1	1.16	-.54	1.41
χ^2	13.44	12.19	15.39
df	8	8	8
p	.10	.14	.05

It is also evident from Table 2.5 that the decision criteria (*c*) estimated by SDT-INT are spread wider than those estimated by both SDT-MAX and SDT-SEQ. This is because they are scaled according to the detection decision variable for SDT-INT, the sum of the familiarity of all lineup items. Consequently, the decision criteria estimated by SDT-INT are not directly comparable to those estimated by SDT-SEQ (or SDT-MAX). In contrast, the

decision variables of SDT-SEQ and SDT-MAX are both based on "untransformed" signal strengths and are therefore directly comparable.

These difficulties in comparing the parameter estimates of SDT-INT to SDT-SEQ mean that SDT-INT is not well suited to testing our hypothesis. As a result, we employ SDT-MAX as the simultaneous lineup model and SDT-SEQ as the sequential lineup model in all subsequent analyses.

2.9.7.2 Underlying Discriminability

Table 2.6 shows the results of the likelihood-ratio tests of the equality of each parameter between the simultaneous and sequential conditions as estimated by the SDT-MAX and SDT-SEQ models respectively. The estimates of d_t and s_t did not differ significantly between the simultaneous and sequential conditions.

Table 2.6

Likelihood Ratio Tests for Parameter Equality

	$\chi^2(1)$	p
d_t	.15	.70
s_t	.87	.35
c_5	.01	.91
c_4	.28	.60
c_3	.28	.60
c_2	.48	.48
c_1	10.54	< .01

Note. Significant *p*-values at $\alpha = .05$ indicate that model fit significantly worsened when a parameter was constrained to be equal across the simultaneous and sequential conditions. For each unconstrained model, we fit SDT-SEQ to the sequential data and SDT-MAX to the simultaneous data. The unconstrained models had 16 degrees of freedom, fixing one parameter increases the degrees of freedom to 17, $\chi^2(17) - \chi^2(16) = \chi^2(1)$, thus the χ^2 tests above have one degree of freedom.

The lack of a significant difference in underlying discriminability between simultaneous and sequential lineups is consistent with our previous re-analysis of the Palmer-Brewer database. It suggests that this result is not easily attributable non-random target position in sequential lineups or the use of a designated innocent suspect selected to resemble the target to a greater extent than the foils. We also attempted to address the lack of statistical power in many of the studies in the Palmer-Brewer database. Despite increasing the number of participants compared, we were unable to observe a statistically significant difference in underlying discriminability. This suggests that if there is a simultaneous advantage, it is small enough to be difficult to detect. The effect size as measured by Hedge's g for the difference between simultaneous and sequential underlying d_t is small, $g = .06$.

Additionally, our conclusion rests on the assumption that the SDT-MAX model is an appropriate model of the simultaneous lineup data. Recently, Wixted et al. (2018) proposed the *Ensemble model* based on the idea of comparing diagnostic features[2]. In this model, the item with the maximum familiarity (and potential target) is compared to the average familiarity of the remaining items. If this difference exceeds an evidential criterion, the potential target is identified, otherwise the lineup is rejected. We also fit this model to data from the simultaneous condition of experiment one and found that it provided an excellent fit, $\chi^2(8) = 6.96$, $p = 0.54$. However, we again found no statistically significant difference between its estimate of d_t and the estimate from the SDT-SEQ model, $\chi^2(1) = 0.29$, $p = 0.59$.

2.9.7.3 Response Bias

Table 2.5 shows that estimates of decision criteria (c_2, …, c_5) are comparable between simultaneous and sequential lineups for each criterion except c_1, which separates lineup identification and rejection decisions (the choose/no choose threshold). Table 2.6 shows that c_1 was significantly larger in the sequential condition, supporting our hypothesis and conforming to previous literature (Carlson et al., 2016; Clark, 2012a; Dobolyi & Dodson,

2013; Gronlund et al., 2009; Meissner et al., 2005). Interestingly, having made this decision, the assignment of additional confidence levels did not differ between the two procedures.

2.9.7.4 Target Distribution Variance

Table 2.6 shows that estimates of target distribution variance (s_t) did not differ between simultaneous and sequential presentation. The s_t values displayed in Table 2.5 are also close to one for both presentation formats, implying that equal-variance models may account for this data. Constraining the models so that $s_t = s_s = 1$ did not significantly worsen the fit for SDT-MAX, $\chi^2(1) = .28$, $p = .60$, or SDT-SEQ, $\chi^2(1) = .61$, $p = .43$. This indicates that equal-variance models adequately capture these data, in contrast to long-standing findings of unequal target and lure distribution variance in basic recognition memory literature (Egan, 1958; Mickes et al., 2007) and in recent lineup research (Wilson et al., 2019; Wixted et al., 2018).

2.9.8 Sequential Position One compared to the Simultaneous Lineup

In addition to greater underlying discriminability in the simultaneous lineup, DFDT also predicts that underlying discriminability should increase over the course of the sequential lineup (Wixted & Mickes, 2014). The presentation of each new sequential lineup item provides an additional opportunity to isolate distinctive features uniquely shared by the target and the lineup items. Consistent with this, Wilson et al. (2019) found greater underlying discriminability at sequential target positions two to six compared to position one. This suggest that the difference in discriminability between sequential and simultaneous presentation should be greatest at sequential position one and should reduce over the course of the lineup. Because position one in a sequential lineup is equivalent to a single item showup, this result is also consistent with the robust finding that the simultaneous lineup outperforms the single-suspect showup (Gronlund et al., 2012; Neuschatz et al., 2016; Wooten et al., 2020).

When comparing underlying discriminability between simultaneous and sequential presentation, differences between the simultaneous lineup and each sequential position are aggregated. Fully randomising the position of the target, as in our experiment, may have reduced the average simultaneous advantage, which may explain why we failed to find one. To investigate this possibility, we compared underlying discriminability between sequential position one and the simultaneous lineup.

2.9.8.1 Data

Table 2.7 shows the frequency counts for sequential serial position one (i.e. showup) data and the simultaneous lineup. Because of the comparatively small number of target present trials in sequential position one, it was not possible to classify the data further by confidence level. In order to treat sequential position one responses as a showup, we reclassified participants' responses as follows. A TP_1 showup trial occurred when the first sequential lineup item was the target. A TA_1 showup trial occurred when the first sequential lineup item was a foil. Note that this includes those participants who encountered the target at a later serial position in the lineup as well as those who never saw a target.

Decision outcomes frequencies for sequential serial position one, treated as a showup, and the simultaneous lineup

Table 2.7

Data from Sequential Position One and the Simultaneous Lineup

Showup (Sequential Serial Position One)		
	Identify	Reject
TP_1 – Target ID	15	13
TA_1 – Foil ID	19	262

Simultaneous Lineup		
	Identify	Reject
TP – Target ID	99	19
TP – Foil ID	21	
TA – Foil ID	80	61

2.9.8.2 Model Fits and Results

We used an equal variance (EVSD) model of the yes/no task to estimate showup d_t and c and SDT-MAX to estimate simultaneous d_t and c. As previously, we conducted likelihood ratio tests comparing the overall fit of an unconstrained model fit to each dataset simultaneously, to various constrained models where one parameter was set to be equal across the two sets of data.

We fit the EVSD model to the showup data. In this case, it has an analytic solution given by, $d_t = \Phi^{-1}(H) - \Phi^{-1}(F)$ and $c = \Phi^{-1}(1 - F)$, where H is the Target ID rate, F is the TA Foil ID rate and Φ^{-1} is the inverse normal cumulative distribution function. Because there are no degrees of freedom, this model necessarily fits perfectly. The estimated parameter values were, $d_t = 1.58$ and $c = 1.49$. We fit the SDT-MAX to the simultaneous data with the constraint that $s_t = 1$. It fit these data well, $\chi^2(1) = 2.54$, $p = .11$, with estimated parameter values, $d_t = 1.98$ and $c = 1.18$. Although underlying discriminability appeared to be greater for the simultaneous lineup, this difference was not significant, $\chi^2(1) = 1.87$, $p = .17$. Responding was significantly more conservative for sequential position one, $\chi^2(1) = 5.79$, $p < .05$, consistent with previous findings at the aggregate level.

Despite previous studies that have reported a simultaneous advantage in underlying discriminability over showups (e.g. Neuschatz et al., 2016) we failed to observe a similar effect in our data. Because the experiment was not designed with this analysis in mind, the number of participants in the TP_1 was relatively small (N = 28) which means that the analysis may not have sufficient statistical power. Nevertheless, it is possible to conclude that if there is an advantage for simultaneous presentation it is likely to be a relatively small effect.

2.10 Reanalysis of Simultaneous vs. Sequential Studies conducted since Palmer and Brewer (2012)

We failed to find an underlying discriminability advantage for the simultaneous lineup compared to the sequential lineup in a corpus of studies published prior to Steblay and Phillips (2011) and in our own experimental data. However, it is possible that such an effect occurs in studies published after Steblay and Phillips (2011), particularly those that report an empirical discriminability advantage for simultaneous presentation (e.g. Mickes et al., 2012). We conducted a literature search for studies published since 2011 that compared photographic simultaneous and stopping-rule sequential lineups. We isolated studies that reported results in such a way that we could extract the cell frequencies required to fit the SDT-MAX and SDT-SEQ models. Seven simultaneous vs. stopping sequential studies published since 2011 reported met our criteria; Pica and Pozzulo (2017), Flowe et al. (2016), Carlson et al. (2016), Pozzulo et al. (2016), Sučić et al. (2015), Carlson and Carlson (2014) and Pozzulo et al. (2013). Additionally, we requested the data from Mickes et al. (2012), from which we were able to extract the required cell frequencies for experiment 1a, but not experiments 1b or 2. This new corpus of eight studies (total N = 6453, simultaneous n = 2803, sequential n = 3650) provides more power to detect a simultaneous advantage in underlying discriminability than the Palmer and Brewer corpus (total N = 3871, simultaneous n = 1952, sequential n = 1919).

2.10.1 Method

As per our analysis of the Palmer and Brewer corpus, we estimated d_t, c and, where relevant, d_s for each study by fitting SDT-MAX to the simultaneous data and SDT-SEQ to the sequential data. We then calculated mean discriminability ($d_t - d_s$) and response bias (c) weighted by sample size for simultaneous and sequential presentation. For most of the studies, we estimated parameters separately for each experimental condition, rather than

collapsing over conditions other than presentation format. This led to thirteen simultaneous vs. sequential datasets from the eight studies. For Carlson and Carlson (2014) and Carlson et al. (2016), we collapsed the sequential target position two and target position five conditions, specifying that the target could only appear at these two positions when fitting SDT-SEQ. For Pozzulo et al. (2013) we collapsed the adolescent and adult age conditions because the original study reported no effect of age on decision performance.

2.10.2 Results

Model fit statistics and parameter values for each dataset are available in Appendix C, Table 2. SDT-MAX fit 12 of 13 simultaneous datasets at $\alpha = .05$, failing to fit the backloaded simultaneous condition of Carlson et al. (2016). SDT-SEQ fit 10 of 13 sequential datasets at $\alpha = .05$, failing to fit the sequential data from Sučić et al. (2015), the sequential weapon present plus distinctive feature condition from Carlson and Carlson (2014) and the sequential data from Pozzulo et al. (2013). Table 2.8 shows the mean and standard deviations for discriminability and response bias (*c*) for simultaneous and sequential presentation, weighted by sample size.

Table 2.8

Mean Parameter Values weighted by Sample Size from fits SDT-MAX to Simultaneous Lineup Data and SDT-SEQ to Sequential Lineup data from a Corpus of Eight Recent Studies

		Parameter			
		discriminability		*c*	
Format	Source	μ_w	σ_w	μ_w	σ_w
Simultaneous	SDT-MAX	1.23	.54	1.09	.21
Sequential	SDT-SEQ	1.02	.38	1.09	.32

Welch two-sample weighted *t*-tests indicated no significant difference in mean weighted discriminability, $t(21.43) = 1.14$, $p = .27$ or mean weighted response bias, $t(20.72) =$

0.08, $p = .94$, between presentation formats. As for the Palmer and Brewer corpus and our experiment, this does not support the hypothesis that underlying discriminability is greater for simultaneous presentation. Unlike our previous analyses, the hypothesis that responses are more conservative in the sequential procedure was not supported.

2.11 General Discussion

The present study sought to compare performance between the simultaneous lineup and sequential stopping-rule lineup in order to test the central prediction of the diagnostic feature detection hypothesis; that underlying discriminability is greater when lineups are administered simultaneously rather than sequentially (Wixted & Mickes, 2014). As structural differences between the procedures affect the shape of the corresponding ROCs, a difference in empirical discriminability between simultaneous and sequential presentation does not necessarily indicate a difference in underlying discriminability. In order to measure underlying discriminability, it is necessary to characterise the data in terms of an appropriate model. Accordingly, we developed a novel signal detection model that captures the structure of the sequential lineup task, SDT-SEQ, and contrasted this with models of the simultaneous lineup task, SDT-MAX and SDT-INT (as well as the Ensemble model).

We first fit SDT-MAX, SDT-INT and SDT-SEQ to the Palmer and Brewer (2012) database comprised of a set of earlier studies that directly compared simultaneous and sequential stopping-rule presentations. While we identified and corrected a number of methodological shortcomings in the original study, the conclusions that we reached were the same. First, we found no systematic difference in underlying discriminability between the two kinds of lineup (measured by the parameter, d_t, or $d_t - d_s$ where relevant). Second, we found a shift to a more conservative response bias in sequential lineups. As the studies in the database did not collect or report post-decision confidence estimates, we were unable to estimate all the parameters specified in our models, leaving more nuanced aspects of the

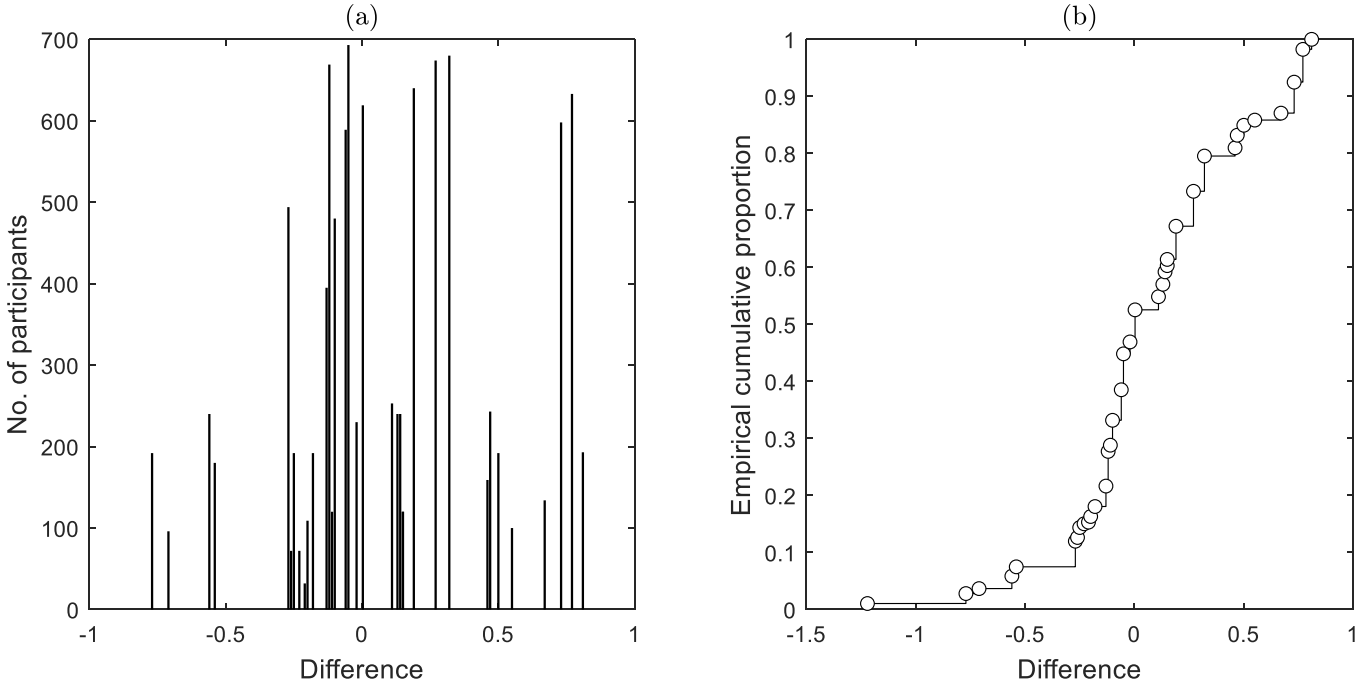
simultaneous vs. sequential presentation question unexplored. Most studies also had relatively small numbers of participants and so lacked statistical power to detect a small effect, selected designated innocent suspects designed to resemble the target, and did not randomise the position of the target in sequential lineups. For this reason, we conducted a more powerful experiment that elicited multiple confidence judgments, did not employ a designated innocent suspect and randomised the position of the target on sequential lineups. We found no significant difference in underlying discriminability and more conservative responding for the sequential lineup, consistent with the Palmer and Brewer reanalysis. Finally, we analysed a corpus of data containing eight recent lineup studies that compared simultaneous and sequential presentation. The results were consistent with the previous findings in that there was no significant difference in underlying discriminability, but we did not find more conservative response bias for sequential presentation.

Our analyses provide estimates of the difference in underlying discriminability between simultaneous and sequential lineups across a total of 36 separate studies or conditions within studies. While many features of these studies (e.g., lineup size, target position, presence of a designated suspect, backloading) vary considerably, each provides a point estimate of the difference in underlying discriminability. These estimates are plotted in

Figure 2.5 Panel A weighted by the number of participants and in Panel B as a cumulative proportion ogive.

Figure 2.5

Difference between Simultaneous and Sequential Discriminability plotted for each Dataset



Panel A can be viewed as a "group-based" histogram in which each participant is assigned the difference estimate calculated for their group as a whole. Each vertical bar is centered on a given estimate and the length of the bar corresponds to the total number of participants in the group. The total number of participants across all the studies is 10,913. According to these data, the overall weighted mean difference is 0.09, indicating a slight advantage for simultaneous lineups. The same data are plotted in Panel B as a cumulative proportion ogive. From this, it is possible to determine that the median difference is .003, the 5th percentile is -.56 and the 95th percentile is .77. Thus, in the studies we have analysed, approximately 50% of participants can be presumed to have shown a simultaneous advantage in underlying discriminability while the remaining 50% show the opposite. Overall, this means that although some more recent studies have observed a simultaneous advantage in

underlying discriminability, the evidence to date taken as a whole suggests that this effect is close to zero.

2.11.1 Diagnostic Feature Detection Theory

Our results are not consistent with a key prediction of diagnostic feature detection theory (DFDT), that the greater opportunity to compare lineup items in the simultaneous lineup should improve underlying discriminability compared to the sequential lineup. However, the lack of an easily detected difference in underlying discriminability between simultaneous and sequential lineups does not necessarily militate against the processes proposed by the DFDT. All things being equal, it is possible that the greater detectability of diagnostic features in simultaneous lineups may lead to a performance advantage. However, this is a critical caveat – there may be other differences between the procedures that serve to counteract this effect. One obvious difference is the size of the choice set. In a simultaneous lineup, the target (if present) is one of several alternatives while in a sequential lineup, on each trial only a single item is presented. It is well-known that the probability of correct target detection declines with the increasing size of the choice set (Swets, 1959). On the other hand, it is possible that sequential presentation may induce retroactive interference through re-encoding of lineup items into memory. This would be expected to have a greater impact on items appearing later in the sequence which is suggested by the finding reported by Wilson et al (2019) that underlying discriminability may increase over the course of the sequential lineup, at least after position one. The point is that because the two procedures have different characteristics, it is likely they induce a range of effects on memory which, in the cases we have so far examined, more or less cancel out. Diagnostic feature detection may well occur but its effects on memory may be counteracted by other differences.

The foregoing analysis suggests that if relevant differences between simultaneous and sequential lineups could be reduced then the effects of diagnostic feature detection may be

revealed. A recent study by Colloff and Wixted (2019) bears on this issue. They compared a standard showup in which only the suspect was presented with a novel *simultaneous showup* in which the suspect was presented along with five fillers none of which could be identified as the target. Based on ROC analysis, they found that the opportunity to compare the suspect to other similar faces in the "simultaneous showup" procedure improved empirical discriminability. Because the structural characteristics of the standard and simultaneous lineups are essentially the same – both require a decision to be made about a single item – the difference in empirical discriminability suggests a corresponding difference in underlying discriminability. If so, then the results reveal the kind of advantage predicted by the DFDT.

2.11.2 The UK Lineup Procedure

In a series of studies, Seale-Carlisle and colleagues have investigated the empirical and underlying discriminability of the UK (or PACE) lineup procedure (Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019; Wixted et al., 2018). This procedure is conducted in accordance with the UK Police and Criminal Evidence guidelines (Police and Criminal Evidence Act, Code D, 2017). It differs in important ways from the stopping rule sequential lineup. First, witnesses see short videos of each lineup member rotating through a head-and-shoulders profile rather than a static photo. Second, witnesses must view two full laps of the lineup procedure before making a decision, i.e. the lineup does not have a stopping rule, and may return to any item as many times as they wish before making their decision. In addition, the UK lineup contains nine items rather than six, as is common in other jurisdictions.

Seale-Carlisle and Mickes (2016) found that the UK lineup procedure had lower empirical discriminability based on ROC analysis than a comparable simultaneous lineup. Seale-Carlisle et al (2019) conducted a series of experiments to try to isolate which aspects of the UK procedure were responsible for this difference. They also examined underlying discriminability by fitting the ensemble model to different versions of the UK lineup. They

concluded that the crucial feature that impaired relative performance in the UK lineup was the sequential presentation format. This was identified in one experiment (Experiment 1) and partially verified in a second experiment (Experiment 5). That is, both experiments found a difference in empirical discriminability based on measure of area under the ROC curve, but although a significant difference in underlying discriminability was found in the first experiment, this was not replicated in the second.

The results found by Seale-Carlisle et al (2019) are, to our knowledge, the only example of a significant simultaneous lineup advantage in underlying discriminability. Because there is no stopping rule, witnesses make their decision after having viewed all the lineup items. Therefore, in terms of the task demands, the UK lineup functions as a kind of simultaneous lineup in which viewing of items is constrained to be sequential. The decrement in underlying discriminability identified by Seale-Carlisle et al. appears to be a consequence of this feature. However, our previous analyses suggest that it may not be a consequence of sequential presentation per se. These show that sequential presentation with a stopping rule does not significantly impair underlying discriminability. The difference must lie elsewhere. One possibility is that the UK procedure places additional memory demands on witnesses who must encode information about the members in the lineup, such as their facial features and lineup position, for a future identification decision. This may lead to the build-up of retroactive interference between test items and target memory (Dewar et al., 2007; Sosic-Vasic et al., 2018; Wickelgren, 1966). In contrast, the presence of a stopping-rule reduces memory demands because once a decision is made, the features of the current lineup item can be immediately forgotten.

Consistent with previous studies (Carlson et al., 2016; Clark, 2012a; Dobolyi & Dodson, 2013; Gronlund et al., 2009; Meissner et al., 2005; Palmer & Brewer, 2012), we

found that sequential presentation led to more conservative responding. This conforms to the original intention behind the introduction of sequential lineups, to reduce false alarms.

2.11.3 Conclusions

This study introduced a new model of the sequential lineup task, SDT-SEQ, and in conjunction with models of the simultaneous lineup task, SDT-MAX and SDT-INT, tested a key prediction of the diagnostic feature detection theory that underlying discriminability should be greater in a simultaneous lineup. In both our re-analysis of the Palmer-Brewer (2012) database and data from eight recently published studies, in addition to the results of a new experiment, we did not find evidence consistent with this prediction. This suggests that if the effect exists, it may be counteracted by other effects associated with differences between the two kinds of task. Further research is required to determine the conditions under which comparing features across lineup items improves memory, the limits of such an effect, and the extent to which it is affected by structural aspects of different lineup tasks.

Chapter 3

3.1 Preface to Study Two

Study one found that the sequential stopping rule lineup and simultaneous lineup did not significantly differ in discriminability. When aggregating results from a large number of previously published datasets and newly collected experimental data, the estimated effect size for this difference was close to zero. The sequential stopping rule lineup was associated with more conservative responding than the simultaneous lineup, consistent with previous research. The finding of no significant difference in discriminability was consistent with previous research that had compared these two tasks (Palmer & Brewer, 2012). However, previous research also found that the sequentially presented UK lineup task was associated with lower discriminability and more lenient – rather than more conservative – responding than the simultaneous lineup (Seale-Carlisle et al., 2016; Seale-Carlisle et al. 2019). These conflicting results for two sequentially presented lineup tasks raise the possibility that some other aspect of the UK lineup task, such as memory interference, or some aspect of the methodology of previous experiments, is responsible for the reported discriminability decrement and/or the lenient responding observed on the UK lineup.

Study two aimed to clarify the pattern of results observed in previous studies by comparing discriminability and response bias on the simultaneous lineup, the sequential stopping rule lineup, and the UK lineup. In a single large sample experiment (n = 2861), participants were allocated to see a six-item lineup with the target either present or absent, presented either simultaneously, sequentially with a stopping rule, or according to the UK presentation guidelines. As this study focused on presentation format, the UK lineup condition used still photos shown for five seconds each rather than 15 second long videos so as not to introduce added variability in to the experimental design, which would have reduced statistical power and complicated the interpretation of the results. However, the UK lineup

did include the option to revisit lineup items before identifying, which is a feature of the real-world task not included in previous studies. This also allowed the comparison of participants who revisited compared to those who did not revisit on the UK lineup. The method of lineup composition and pool of sixteen items were shared with study one.

3.2 Statement of Authorship

Title of Paper	Clarifying the Effects of Sequential Item Presentation on Discriminability and Response Bias
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	

Principal Author

Name of Principal Author (Candidate)	Matthew Kaesler		
Contribution to the Paper	Designed the study, programmed the UK lineup procedure for the web-based experiment, collected the data, fit mathematical models to data and wrote additional code for conducting hypothesis tests, wrote the manuscript.		
Overall percentage (%)	80		
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	13/07/2021

Co-Author Contributions

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	Professor John Dunn
Contribution to the Paper	Developed code for fitting the mathematical models to data (same code as for study one), assisted with interpreting the results, and provided feedback on manuscript draft.

Signature		Date	13/7/2021

Name of Co-Author	Associate Professor Carolyn Semmler		
Contribution to the Paper	Supervised all aspects of study design and execution and provided feedback on manuscript draft.		
Signature		Date	23/02/2021

Clarifying the Effects of Sequential Item Presentation on Discriminability and Response Bias

by

Matthew Kaesler[1], John C. Dunn[2,3], Carolyn Semmler[1]

[1] University of Adelaide

[2] University of Western Australia

[3] Edith Cowan University

Address for correspondence:

Mr Matthew Kaesler
University of Adelaide
North Terrace
Adelaide, SA, 5005
Email:

Publication Details

Unpublished manuscript

3.3 Abstract

Previous research has reported diverging patterns of results with respect to underlying discriminability and response bias when comparing the simultaneous lineup to two different lineup procedures on which items are presented sequentially, the sequential stopping rule[4] lineup and the UK lineup. As a result, the effect of sequential item presentation on these latent variables is unclear. In a single large sample experiment, we compared underlying discriminability and response bias on six-item photographic lineups presented either simultaneously, sequentially with a stopping rule, or sequentially according to UK presentation guidelines. Underlying discriminability was higher for the simultaneous lineup compared to the sequential stopping rule lineup, despite a non-significant difference in empirical discriminability between the procedures. There was little or no difference in underlying discriminability when comparing the simultaneous lineup to the UK lineup and the UK lineup the sequential stopping rule lineup. Responding was most lenient for the UK lineup, followed by the simultaneous lineup, followed by the sequential lineup. These results imply that sequential item presentation may not exert a large effect in isolation on the latent variables of interest. Rather, underlying discriminability and response bias on the sequential stopping rule lineup and UK lineup result from the interaction of sequential item presentation with other factors of these procedures.

[4]In contrast to study one, the terminology in this study distinguishes between the sequential lineup conducted with a stopping rule and the UK lineup, conducted without, as these two procedures are directly compared.

3.4 Introduction

The procedure for presenting a lineup to a witness differs across nations and jurisdictions. In the majority of United States (US) jurisdictions and in Australia, lineup items are presented to the witness simultaneously. In Canada, the United Kingdom (UK) and approximately 30% of US jurisdictions, lineup items are presented to the witness sequentially (Police Executive Research Forum, 2013). Recent research comparing simultaneous and sequential presentation has found that the sequential presentation of items may impair underlying discriminability compared to the simultaneous presentation of items (Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019; Wixted et al., 2018). However, these studies compared the simultaneous lineup to the UK lineup procedure, in which procedural aspects other than sequential presentation may impair underlying discriminability. In a recent study, we found no significant difference in underlying discriminability between the simultaneous lineup and a sequential lineup task conducted with a stopping rule (Lindsay & Wells, 1985) in both new experimental data and data from a corpus of previous studies (Kaesler et al., 2020). This raises the question of whether the sequential presentation of items itself impairs underlying discriminability or whether the decrement in underlying discriminability observed for the sequentially-presented UK lineup is due to some other aspect of that procedure. Our aim in this study is to clarify the effects of sequential presentation on underlying discriminability by comparing performance on the simultaneous lineup, the sequential stopping rule lineup and UK lineup in terms of underlying discriminability and response bias. We will estimate the underlying parameter values associated with each procedure by fitting appropriate signal detection models to the data.

3.4.1 Sequential Lineup Tasks

As originally described by Lindsay and Wells (1985), the sequential stopping rule lineup task involves presenting lineup items to the witness one at a time, with the witness

making a yes/no decision for each item. If the witness rejects an item, they are shown the next item. If the witness identifies an item, their decision is recorded and the procedure is terminated; any remaining items are not shown. We refer to this as the "stopping rule". If the witness rejects all items, their response is recorded as a non-identification. Like the simultaneous lineup, the sequential stopping rule lineup is conducted with photographic stimuli and is typically comprised of six items.

In the UK, eyewitness identification policy is governed by the Police and Criminal Evidence Act (1984) Code D (Revised 2017). The UK lineup procedure also presents items to the witness one at a time, but the procedure differs substantially in other ways. First, the witness does not make an identification decision for each sequentially presented item. Rather, the witness must view the sequence of lineup items twice through before making their identification decision. This viewing is not self-paced; the items in the UK lineup are 15 second long videos of a head-and-shoulders profile rotating through 180 degrees. Second, the UK lineup comprises nine items rather than the six items typical of simultaneous and sequential lineups. Finally, the witness is able to revisit items as many times as they wish before making their identification decision, although this feature has not been included in experimental versions of the task.

3.4.2 Procedural Aspects of the UK Lineup

Seale-Carlisle et al. (2019) conducted five experiments that aimed to understand which aspect of the UK lineup procedure was responsible for its decrement in underlying discriminability compared to the simultaneous lineup procedure (Seale-Carlise & Mickes, 2016). These experiments held other aspects of the lineup constant while manipulating presentation format (sequential vs simultaneous), stimulus modality (photo vs video), laps required before identification decision (one lap vs optional two laps vs required two laps) and lineup size (six vs nine items). There was a small but only marginally significant underlying

discriminability advantage for photo stimuli over video stimuli, while there were no differences in the lineup size or lap conditions. In two experiments, simultaneous presentation was associated with greater underlying discriminability that sequential presentation. This pattern of results implies that the sequential presentation of items is the main factor in the decrement in underlying discriminability observed for the UK lineup compared to the simultaneous lineup (Seale-Carlisle & Mickes, 2016).

3.4.3 Diagnostic Feature Detection Theory

These findings are consistent with the Diagnostic Feature Detection Theory (DFDT) of eyewitness memory. DFDT proposes that underlying discriminability is enhanced when witnesses are able to compare across lineup items in order to isolate features uniquely shared by the lineup items and their memory of the target. The greater opportunity to compare items afforded by a simultaneously presented lineup compared to a sequentially presentation lineup will therefore lead to greater discriminability. However, we recently failed to find support for this prediction (Kaesler et al., 2020) in an analysis of 36 datasets (n = 10,913) from studies that compared the simultaneous lineup and the sequential stopping rule lineup (not the UK lineup). The overall effect size for the difference in underlying discriminability between simultaneous and sequential presentation was close to zero. This indicates that comparing across lineup items may provide, at best, only a small benefit to underlying discriminability. This benefit may be counteracted by other procedural factors. The results of Kaesler et al. (2020) imply that sequential presentation may not be the main factor in the underlying discriminability decrement observed for UK lineup presentation compared to simultaneous presentation, contrary to the conclusions reached by Seale-Carlisle et al. (2019).

3.4.4 Possible Memory Interference in the UK Lineup

Another aspect of UK lineup presentation that might impair underlying discriminability is the requirement to defer the identification decision until each lineup item

has been seen twice. On the simultaneous lineup, a single identification decision is required from an array of items. The witness is required to assess whether any of the items are a suitably strong match to their memory of the target. On the UK lineup the witness may need to hold some or all items in memory until they make their identification decision, in addition to assessing the degree of match between each item and their memory of the target. The lineup items may therefore interfere with witness memory for the perpetrator (e.g. Criss et al., 2011).

Memory interference, rather than sequential presentation, as the main factor responsible for impairing underlying discriminability on the UK lineup would also explain the similar underlying discriminability observed in Kaesler et al. (2020) for simultaneous presentation and sequential stopping-rule presentation. On the sequential lineup, the witness assesses the degree of match between each item and their memory for the perpetrator, but are not required to hold rejected lineup items in memory because they cannot be revisited, mitigating the potential for interference.

If the beneficial effect of comparing across lineup members proposed by DFDT is small compared to the deleterious effect of interference on the UK lineup, then the simultaneous and sequential stopping-rule procedures would be expected to perform similarly despite a decrement in underlying discriminability for the UK lineup compared to the simultaneous lineup.

3.4.5 Present Study

In the present study, we sought to further investigate the extent to which the sequential presentation of items might impair underlying discriminability by comparing the simultaneous, sequential stopping rule and UK lineup procedures in terms of a critical feature; the method of presenting lineup items. To our knowledge, this is the first study to compare UK lineup presentation and sequential stopping rule presentation. This comparison

is also a test of the predictions of DFDT. To control for effects caused by differences other than presentation method, we conducted each lineup procedure with six items. This meant that the UK lineup was conducted with six photographic items, shown for five seconds each, rather than the typical nine 15-second long video items. Additionally, we included the ability to for witnesses to revisit items on the UK lineup in order to investigate whether this aspect of the task might affect performance. Finally, we sought to address the statistical power limitations in both the archival and experimental data in Kaesler et al. (2020) that may have resulted in us failing to detect a simultaneous advantage in underlying discriminability compared to sequential presentation. We therefore collected approximately three times as many responses for each presentation format condition compared to our experiment in Kaesler et al. (2020).

3.4.6 Model Selection

The signal detection models used in lineup research generally adopt the same underlying unequal variance signal detection framework (Kaesler et al., 2020; Kaesler et al., 2017; Wixted et al., 2018). The signal strength distributions from which the foil and target values are drawn are assumed to be Gaussian in form. The mean and standard deviation of the foil distribution are set to zero and one, respectively. Free parameters are estimated in standard deviation units relative to the mean of the foil distribution. The standard deviation of the target distribution (s_t) may be estimated as a free parameter given sufficient degrees of freedom. In lineup research, it is typical that $s_t < 1$, in contrast to the result typically observed in basic recognition memory research employing list learning designs (Wixted et al., 2018). The other free parameters typically estimated by the models are the mean of the target distribution, d_t, which captures underlying discriminability in conjunction with s_t, and one or more decision criteria, c, which determine the cut point for identification decisions at varying levels of confidence. The critical difference between signal detection models of the lineup

task is in the decision rule that specifies how the signal strength values are compared to the decision criteria (*c*). Specifying different decision rules allows the models to capture the structure of different lineup tasks.

We previously introduced a signal detection model called SDT-SEQ (Kaesler et al., 2020) that accounts for structure of sequential stopping rule lineup by specifying a "first above criterion" decision rule. This is necessary because the decision rules of all other models currently used in lineup research assume that all items are shown on every trial, which is not guaranteed when a stopping rule is in place. This model has provided a good fit to sequential stopping rule lineup data in two studies (Kaesler et al., 2020; Wilson et al., 2019). We employed SDT-SEQ to estimate the underlying signal detection parameters for the sequential lineup data in this study.

While the UK lineup procedure is also sequentially presented, the absence of a stopping rule constraint means that all lineup items are presented on each trial. This means that the UK lineup can be characterised by models on which the decision rules require all items to presented on each trial. Previous research comparing simultaneous and UK lineup presentation has fit the Ensemble model to data from both procedures (Seale-Carlisle et al., 2019; Wixted et al., 2018). We previously found SDT-MAX to provide a good fit to simultaneous lineup data (Kaesler et al., 2020; Wixted et al., 2018). We employed SDT-MAX here to estimate the underlying signal detection parameters for both the simultaneous and UK lineup data, although, as will be shown later, it fit both datasets poorly. We therefore employed the Ensemble model to estimate parameters for these tasks.

3.4.7 Hypotheses

The hypotheses for this study were preregistered on the Open Science Framework https://osf.io/xgzfu. Note that we revised our sampling plan for data collection contained in the original registration. The updated sampling plan is available at https://osf.io/f7nb6. In

accordance with Diagnostic Feature Detection Theory, we hypothesise that simultaneous presentation will be associated with the greatest underlying discriminability as it affords the greatest opportunity for comparison across items, followed by UK lineup presentation, followed by sequential stopping rule presentation.

Previous research indicates that UK lineup presentation leads to more lenient responding than simultaneous presentation (Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019), and that sequential stopping rule presentation leads to more conservative responding than simultaneous presentation (Kaesler et al., 2020; Palmer & Brewer, 2012). On this basis, we hypothesise that responding will be most conservative for sequential presentation, followed by simultaneous presentation, followed by UK lineup presentation.

3.5 Method

3.5.1 Design and Materials

This study employed a 3 x 2 between subjects design. We manipulated presentation format (simultaneous, sequential stopping rule or UK lineup) and target presence (target present or target absent). This study employed the stimuli from Kaesler et al. (2020), sixteen female faces selected on the basis of pairwise similarity ratings and feature-based criteria. See Kaesler et al. (2020) for full details of the stimuli selection process. All stimuli acted as both targets and foils and were randomly drawn on each trial.

3.5.2 Participants

Participants were n = 2982 Amazon Mechanical Turk workers. They were compensated 1 USD for completing the five- to ten-minute experiment. We excluded 120 participants for failing attention check questions presented after the stimulus video and one participants whose identification response did not save correctly, leaving n = 2861 (TP = 1925, TA = 936) participants for the eventual analysis. By presentation condition, simultaneous n = 913 (TP = 612, TA = 301), sequential stopping rule n = 948 (TP = 645, TA

= 303) and UK lineup *n* = 1000 (TP = 668, TA = 332). We used an approximately 2:1 ratio of TP to TA lineups because it is desirable to have more power in the target present conditions than the target absent conditions, as there are three possible target present decision outcomes (target identification, foil identification or miss) and two possible target absent decision outcomes (foil identification or correct rejection).

3.5.3 Procedure

The main structure of the experiment was the same as the experiment in Kaesler et al. (2020), study one of this thesis. The experiment was administered using Amazon Mechanical Turk (AMT) and took place entirely within participants' web browsers. Participants were allocated to conditions using a round robin. After viewing instructions and provided consent to participate, they answered validation questions about the instructions and were directed back to the previous screen if their answers were incorrect. Participants were then shown a video of a randomly selected target from the sixteen-member pool, before completing a visual search distractor task. Pre-lineup instructions corresponding to those in the U.S. National Department of Justice (1999) guidelines were then shown. For the sequential stopping rule condition, participants were instructed that the procedure would terminate if they made an identification. For the UK lineup condition participants were instructed that all lineup members would be sequentially presented, that all members would be shown twice before they could make their decision, then they would be able to revisit lineup members before making their decision. Participants were then shown a target present or target absent lineup, with the target (if any) and the appropriate number of foils (five for target present, six for target absent) randomly selected from the remaining fifteen members of the stimulus pool. The position of the target (if any) and foils was random.

3.5.3.1 Simultaneous Presentation

Participants were presented with six lineup items in a 3 x 2 grid. They could either select a lineup member or select a silhouette labelled “not present”, which indicated a rejection. They selected their desired option and clicked a continue button to confirm their choice.

3.5.3.2 Sequential Stopping Rule Presentation

Participants were presented with each lineup item individually with the option to either to identify or reject it. If an item was rejected, the next item was shown. If an item was identified, the procedure terminated. If all lineup items were rejected, participants were informed that the lineup was exhausted, indicating a rejection decision.

3.5.3.3 UK Lineup Presentation

This procedure is based on the Police and Criminal Evidence Act (1984) Code D (Revised 2017) guidelines. Participants saw each lineup item individually, looping through the set of six items twice in the same order. Each item was presented for five seconds and had a number corresponding to its serial position shown under the photograph. Numbering of the lineup items in this way is explicitly required by the PACE guidelines. Participants were then shown a decision screen which contained numbers one to six shown in a 3 x 2 grid and a silhouette labelled “not present”, which participants could select to indicate a rejection. Under each number were two clickable buttons, one labelled “See Again”, which showed the corresponding lineup item for five seconds before returning to the decision screen, and one labelled “Choose”, which showed a photo of the lineup item with two clickable arrow buttons. The left arrow button labelled “go back” returned participants to the decision screen and a right arrow button labelled “confirm” confirmed their choice. Showing a witness their chosen lineup item before confirming their identification is also explicitly required by the PACE guidelines.

3.5.3.4 Post-Decision Confidence

After making their lineup decision, participants provided a typed confidence estimate for their identification on a 0 – 100 scale and a written confidence estimate in words. Participants also provided a written justification for their identification or rejection decision. They then answered follow-up questions about instruction clarity and task difficulty, and were given the opportunity to provide feedback.

3.6 Results and Discussion

Table 3.1 shows the frequencies of each response type split across six confidence levels for simultaneous, sequential stopping rule and UK lineup presentation. The bin widths were set by dividing confidence ratings from all choosers into six even-as-possible categories.

Table 3.1

Decision Outcome Frequencies for each Presentation Format

Simultaneous Lineup							
Confidence	100	99 – 91	90 – 81	80 – 71	70 – 51	50 – 0	Reject
TP – Target ID	88	57	82	92	83	50	67
TP – Foil ID	1	6	2	16	26	42	
TA – Foil ID	9	12	17	43	38	58	124

Sequential Stopping Rule Lineup							
Confidence	100	99 – 91	90 – 81	80 – 71	70 – 51	50 – 0	Reject
TP – Target ID	69	39	76	114	61	31	102
TP – Foil ID	11	8	21	43	45	25	
TA – Foil ID	6	4	20	31	39	45	158

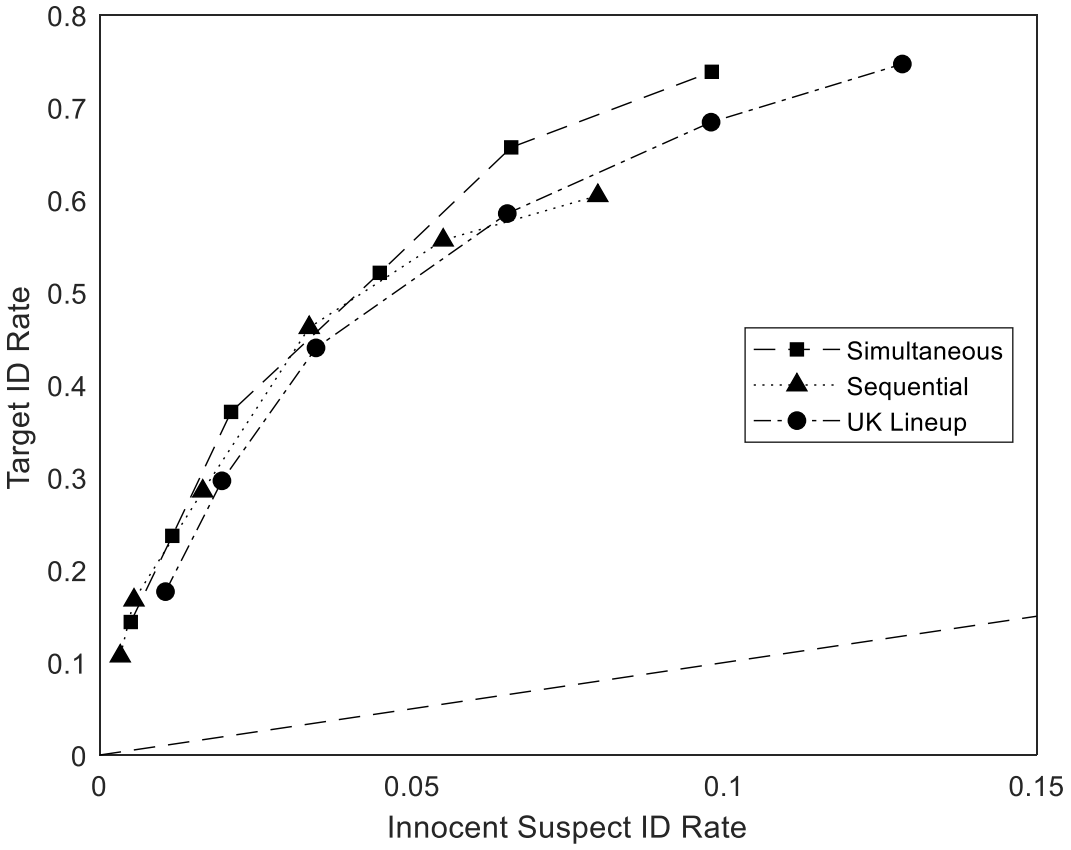
UK Lineup							
Confidence	100	99 – 91	90 – 81	80 – 71	70 – 51	50 – 0	Reject
TP – Target ID	118	80	96	97	66	42	34
TP – Foil ID	8	5	20	30	39	33	
TA – Foil ID	21	18	30	61	65	61	76

3.6.1 ROC Analysis

While our interest is primarily in underlying discriminability, this study is the first – to our knowledge – to directly compare sequential stopping rule and UK lineup presentation. This provides the opportunity to compare the procedures in terms of empirical discriminability. Figure 3.1 shows the target identification rate plotted against the innocent suspect identification rate, i.e. the ROC curve, for each presentation format.

Figure 3.1

Empirical ROC Curves for each Presentation Format



This study did not have a designated innocent suspect so we estimated the rate by dividing the target absent foil identification rate by the lineup size (Mickes et al., 2012). We used pROC (Robin et al., 2011) to conduct pairwise comparisons of partial Area Under the Curve (pAUC) between the procedures, following the tutorial in Gronlund et al. (2014). For

each comparison, the cut point was set at the innocent suspect identification rate of the rightmost point of the more conservative procedure.

First, pAUC did not significantly differ at a cut point of .080 between simultaneous (pAUC = .024) and sequential (pAUC = .027) presentation, D = -.47, p = .64. Second, pAUC did not significantly differ at a cut point of .098 between simultaneous (pAUC = .033) and UK lineup (pAUC = .023) presentation, D = 1.15, p = .25. Finally, pAUC did not significantly differ at a cut point of .080 between sequential (.027) and UK lineup (.016) presentation, D = 1.66, p = .10.

Empirical discriminability did not significantly differ between the simultaneous and sequential lineups, in contrast to some prior studies (Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Experiment 1a Mickes et al., 2012; Neuschatz et al., 2016) but in accordance with others (Flowe et al., 2016; Gronlund et al., 2012; Mickes et al., 2012; Sučić et al., 2015). There was also no significant difference in empirical discriminability between the sequential lineup and the UK lineup. The comparison of these two procedures highlights a known limitation of ROC analysis; comparing a relatively lenient and relatively conservative procedure may disadvantage the lenient procedure (Lampinen et al., 2019; Smith et al., 2020; Wixted & Mickes, 2018). This is because part of the range of the more lenient procedure is discarded when making the comparison. The UK lineup may have an empirical discriminability advantage over the sequential stopping rule lineup in the more lenient region of ROC space where it extends beyond the sequential lineup ROC, because the sequential lineup ROC has been shown to be non-monotonic (Rotello & Chen, 2016; Wilson et al., 2019). This highlights a potential difficulty in comparing pAUC from the sequential stopping rule task to other tasks with monotonic ROC curves. We now apply signal detection models to compare underlying discriminability between each procedure.

3.6.2 Model Fit Performance

The dashed lines in Figures 3.2, 3.3 and 3.4 show the model ROCs for each candidate model plotted against the empirical ROCs for each presentation format. We used the Matlab® fmincon function to fit each model to the data. SDT-SEQ fit the sequential stopping rule data well, $\chi^2(10) = 15.90$, $p = .10$. SDT-MAX did not fit the simultaneous data, $\chi^2(10) = 50.88$, $p < .001$, or the UK lineup data well, $\chi^2(10) = 42.00$, $p < .001$. We planned to use SDT-MAX as the simultaneous and UK lineup model in this study but it provided a poor fit to the data. As a result, we fit the Ensemble model to the simultaneous and UK lineup data, as it has been shown to outperform SDT-MAX in characterising data from these tasks, particularly the UK lineup (Wixted et al., 2018). The Ensemble model fit the UK lineup data well, $\chi^2(10) = 10.65$, $p = .39$, so we adopted it as the UK lineup model for the foregoing analyses. The Ensemble model fit the simultaneous data to a greater degree than SDT-MAX, but still failed to fit the data, $\chi^2(10) = 26.46$, $p < .01$.

Figure 3.2

Model ROC curves for the Simultaneous Lineup

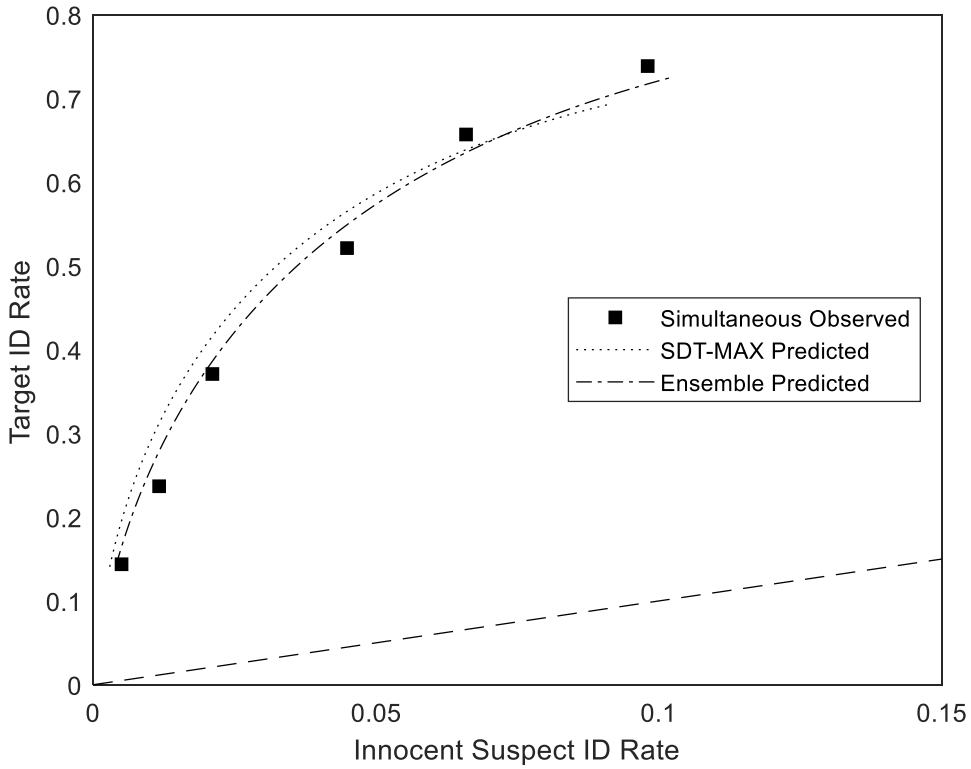


Figure 3.4

Model ROC Curves for the Sequential Lineup

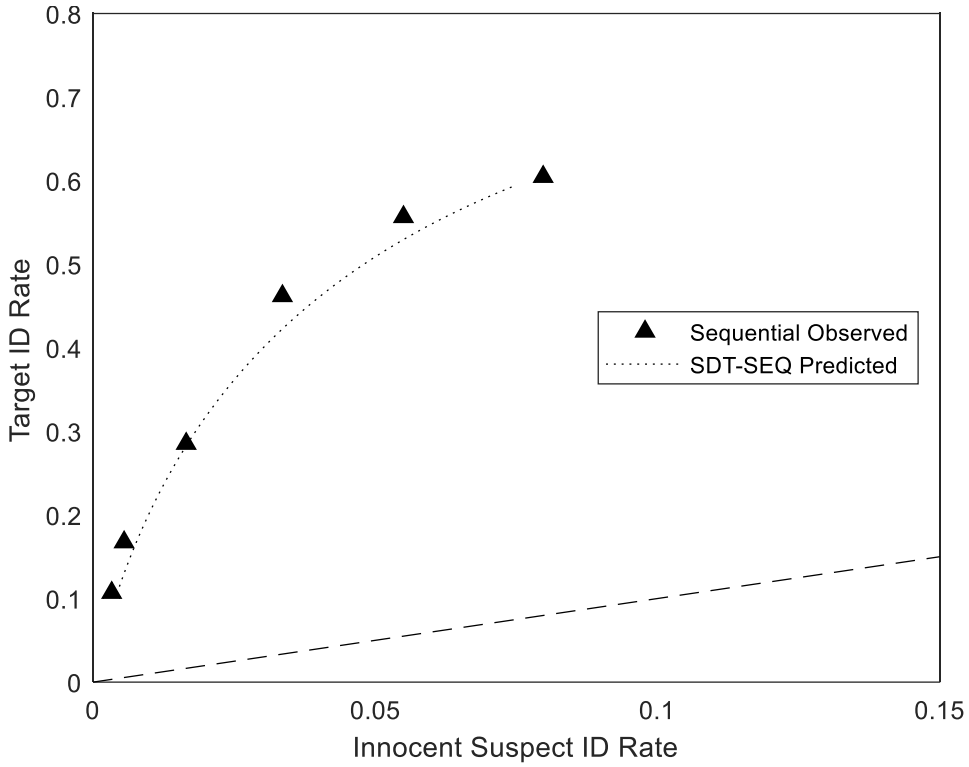
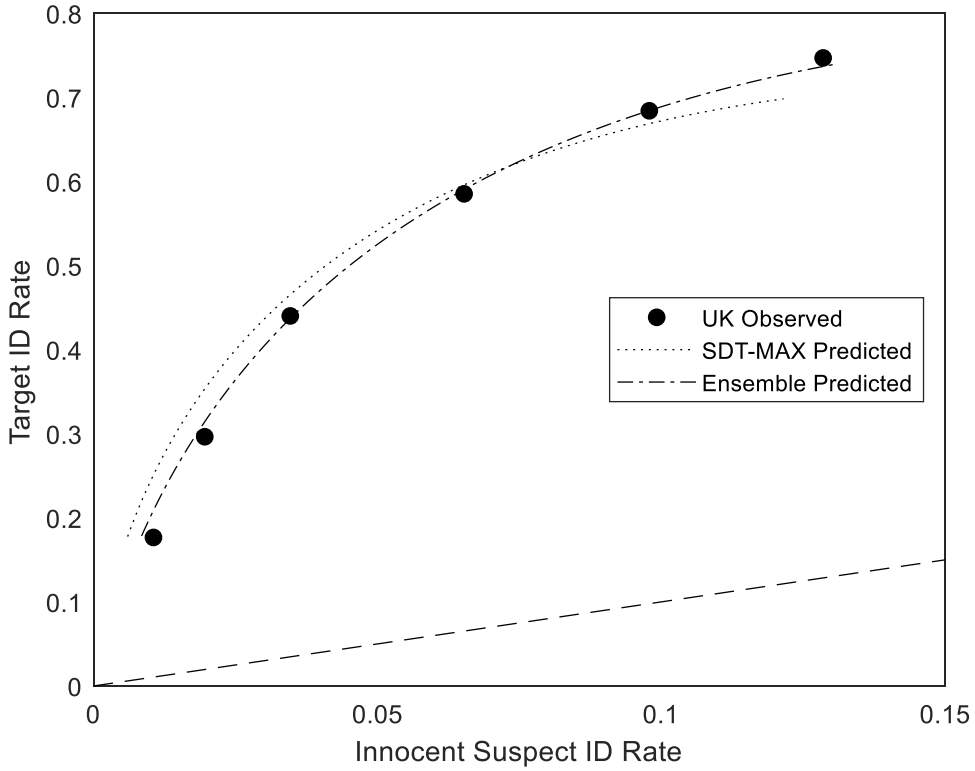


Figure 3.3

Model ROC Curves for the UK Lineup



3.6.2.1 Model Fit Issues

We planned to test our hypotheses using likelihood ratio tests, comparing the fit of a model with all parameters free to vary, fit simultaneously to two datasets, to the fit of a constrained model where one parameter was set to be even across two datasets. The advantage of likelihood ratio testing is that it offers a direct method for testing differences between model parameters. However, the poor fit of SDT-MAX and the Ensemble model to the simultaneous data means that the point estimates of underlying discriminability (d_t and s_t) and response bias (c) for the simultaneous condition recovered by the models may not be an accurate measure of the true parameter values. This complicates the interpretation of any significant differences in point value estimates of underlying discriminability and response bias between simultaneous presentation and the other presentation format conditions, because a significant difference in model parameter estimates may not indicate a significant difference in the true parameter values.

Figure 3.2 shows that SDT-MAX and the Ensemble model capture the general pattern of the simultaneous data, even though they fail to fit it $\alpha = .05$. Thus, the models may still prove useful from a measurement perspective by providing an indication of the region in which the true parameter values might lie. We therefore constructed bootstrap 95% confidence intervals (CIs) around each model parameter for both SDT-MAX and the Ensemble model fit to the simultaneous data, in addition to SDT-SEQ fit to the sequential stopping rule data and the Ensemble model fit to the UK lineup data. Another benefit of the bootstrap CI approach is that it provides some sense of the variability in the point estimate parameter values, although it is a less direct test of the differences in model parameters compared to a likelihood ratio test. When evaluating the predictions our hypotheses, we examined both the bootstrap 95% CIs around each parameter and the results of the likelihood ratio tests.

3.6.2.2 Bootstrap Procedure

We first created 10000 bootstrap datasets of the simultaneous, sequential stopping rule and UK lineup data by generating random vectors from a multinomial distribution using the decision outcome rates for target present and target absent trials and the total number of target present and target absent trials in presentation format condition. This is equivalent to randomly resampling from the data without replacement. We then fit the relevant model to each bootstrap dataset. For some bootstrap samples, the models found local minima in the parameter space and produced implausible parameter values as a result. We excluded these outlier datasets before calculating 95% CIs from the bootstrap distributions for each parameter. In the case with the greatest number of outliers, the Ensemble model fit to the simultaneous lineup data, 43 bootstrap samples out of 10000 were discarded. To clarify, this means that the bootstraps 95% CIs for model parameters reported here are calculated directly from percentiles of the empirical bootstrap distributions and are not bias-corrected and accelerated (BCa) CIs (Efron, 1993). This is because calculating BCa intervals involve an inner jacknife procedure to calculate the acceleration factor that is computationally intensive, which made the runtime for bootstraps unfeasible.

3.6.3 Parameter Estimates

3.6.3.1 Underlying Discriminability

Figure 3.5 shows the point estimates and bootstrap 95% CIs for d_t and s_t recovered by each model when fit to the relevant presentation format data. All 95% CIs for s_t overlap, indicating that s_t did not significantly differ between presentation formats, irrespective of which model was fit to the simultaneous data. None of the likelihood ratio tests for differences in s_t by presentation format and model shown in Table 3.2 were significant, aligning with the pattern of results in the 95% CIs. This indicates that the d_t estimates for

each presentation format are comparable for the purpose of estimating underlying discriminability.

Figure 3.5

Estimates of d_t and s_t with Bootstrap 95% CIs for each Presentation Format

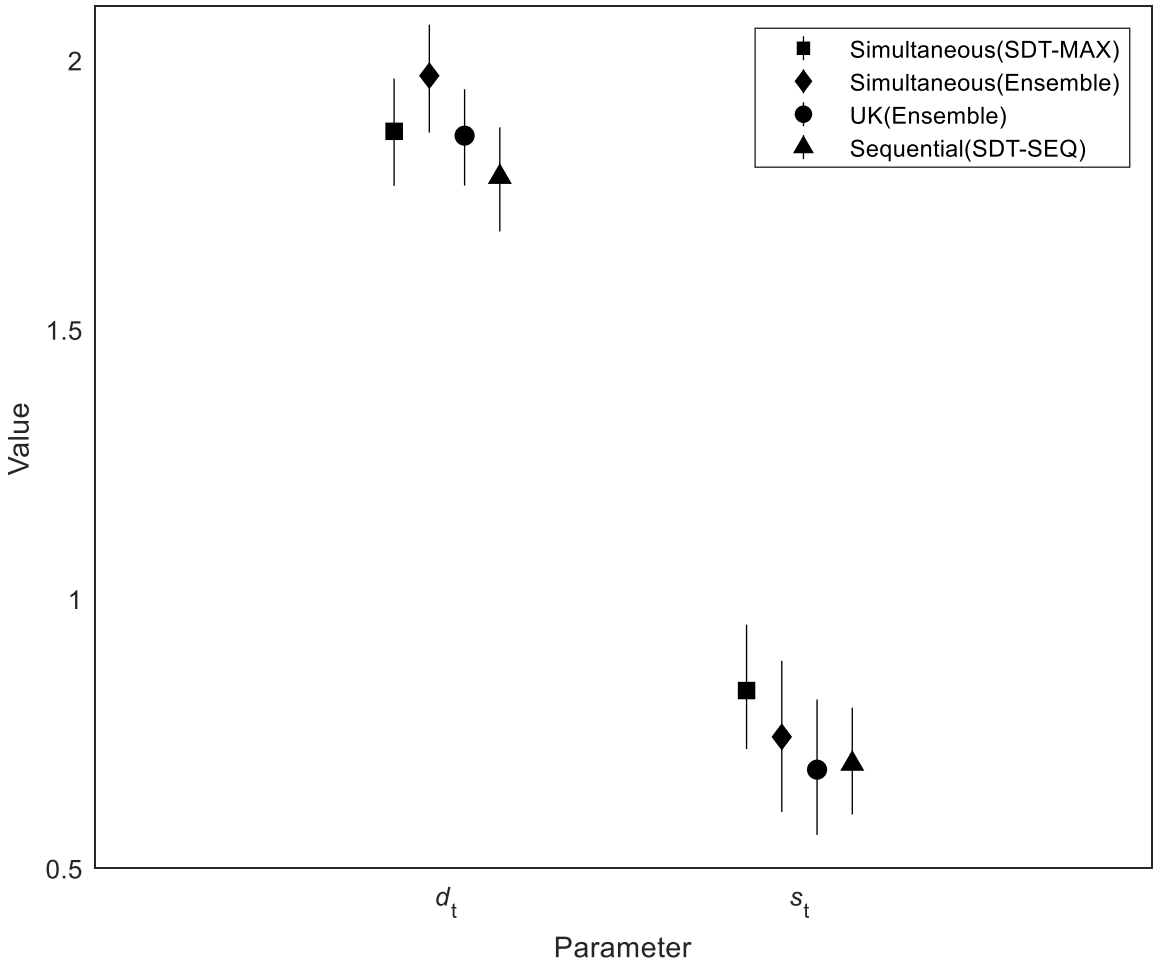


Figure 3.5 shows that d_t was greatest for simultaneous presentation (for both candidate models), followed by UK lineup presentation, followed by sequential stopping rule presentation. The bootstrap 95% CIs for simultaneous and UK d_t overlap, irrespective of which model was fit to the simultaneous data, indicating the d_t advantage for simultaneous presentation over UK lineup presentation is likely not significant. Whether d_t was greater for simultaneous presentation compared to sequential stopping rule presentation depended on which model was fit to the simultaneous data. When the Ensemble model was fit to the

simultaneous data, the bootstrap 95% CIs for simultaneous d_t and sequential stopping rule d_t overlapped, but only beyond the second decimal point, which implies that the difference may be significant. When SDT-MAX was fit to the simultaneous data, the bootstrap 95% CIs for simultaneous and sequential stopping rule d_t overlap, indicating no significant difference in d_t between the procedures.

Table 3.2

Likelihood Ratio Tests for d_t and s_t

	Sim(MAX) vs Seq		Sim(ENS) vs Seq	
	χ^2	p	χ^2	p
d_t	1.51	.22	7.08	<.05
s_t	3.56	.06	.34	.56
d_t and s_t	4.32	.12	7.10	<.05

	Sim(MAX) vs UK(ENS)		Sim(ENS) vs UK(ENS)	
	χ^2	p	χ^2	p
d_t	.01	.91	2.62	.11
s_t	3.40	.07	.44	.51
d_t and s_t	3.50	.17	2.64	.27

	UK(ENS) vs Seq	
	χ^2	p
d_t	1.33	.25
s_t	.02	.89
d_t and s_t	1.51	.47

The likelihood ratio tests for d_t by presentation format and model shown in Table 3.2 reveal a similar pattern of results. The difference between simultaneous and sequential stopping rule d_t was significant when the Ensemble model was fit to the simultaneous data. We calculated Hedge's g as a measure of the effect size, $g = .26$, which is a small effect. The difference in simultaneous and sequential stopping rule underlying discriminability was not significant when SDT-MAX was fit to the simultaneous data. There was no significant difference in d_t between simultaneous and UK lineup presentation, regardless of which model

was fit to the simultaneous data, or between UK lineup presentation and sequential stopping rule presentation.

These results partially support our hypothesis with respect to underlying discriminability. The prediction that underlying discriminability would be greater for simultaneous presentation than for sequential stopping rule presentation was supported when the Ensemble model was fit to the simultaneous data, in accordance with a key prediction of the diagnostic feature detection hypothesis (Wixted & Mickes, 2014). The prediction that underlying discriminability would be greater for simultaneous presentation compared to UK lineup presentation was not supported, in contrast to previous research (Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019), nor was the prediction that underlying discriminability would be greater for UK lineup presentation compared to sequential stopping rule presentation.

The underlying discriminability advantage for simultaneous presentation compared to sequential stopping rule presentation found in this study should be interpreted with caution for two reasons. First, neither the Ensemble model nor SDT-MAX provided a good fit to the simultaneous lineup data. We attempted to work around this issue by constructing bootstrap 95% CIs for differences in addition to testing for differences in the point estimates of the parameters, but we cannot rule out the possibility that point values were inaccurate and/or that the bootstrap CIs did not include the true parameter values. This failure of the models to fit the simultaneous data may be due to noise in the data or some unique feature of our sample not present in a previous simultaneous lineup experiment where we employed the same stimuli and procedure (Kaesler et al., 2020). Another possibility is that some simplifying assumptions made by the models are in conflict with reality, such as the Gaussian form of the distributions for signal strength or that every participant shares the same set of decision criteria (Macmillan & Creelman, 2005).

Second, whether underlying discriminability was greater for simultaneous compared to sequential stopping rule presentation depended on which model was fit to the simultaneous data. This highlights an inherent difficulty with the measurement model approach; measuring instruments sometimes disagree (Chang, 2004). We prefer the Ensemble model results here, as the model both quantitatively and visually provided a better fit to the data than SDT-MAX. The model ROC curves for SDT-MAX when fit to both simultaneous and UK lineup data in Figures 3.2 and 3.3 appear to show that SDT-MAX overestimates performance when responding is conservative and underestimates performance when response is lenient compared to the Ensemble model. The result of this is that SDT-MAX estimates a lesser value for simultaneous d_t compared to the Ensemble model. This pattern is also evident in the unanalysed SDT-MAX estimate of underlying discriminability for the UK lineup, $d_t = 1.77$, compared to the Ensemble model estimate, $d_t = 1.86$.

The uncertainty inherent in analysing simultaneous lineup data that was not well characterised by our candidate models means that we cannot state conclusively whether underlying discriminability was greater for simultaneous compared to sequential stopping rule presentation in this study, or indeed equivalent. However, we consider our results to favour the possibility of a small underlying discriminability advantage.

3.6.3.2 Response Bias

Examination of the empirical ROCs for each presentation format shown in Figure 3.1 imply that responding is most lenient for the UK lineup, followed by the simultaneous lineup, then the sequential stopping rule lineup. This is indicated by the UK lineup ROC curve extending further to the right in ROC space, i.e. achieving a higher overall innocent suspect identification rate, than the ROC curves for the other procedures.

Figure 3.6

Estimates of Decision Criteria with Bootstrap 95% CIs for each Presentation Format

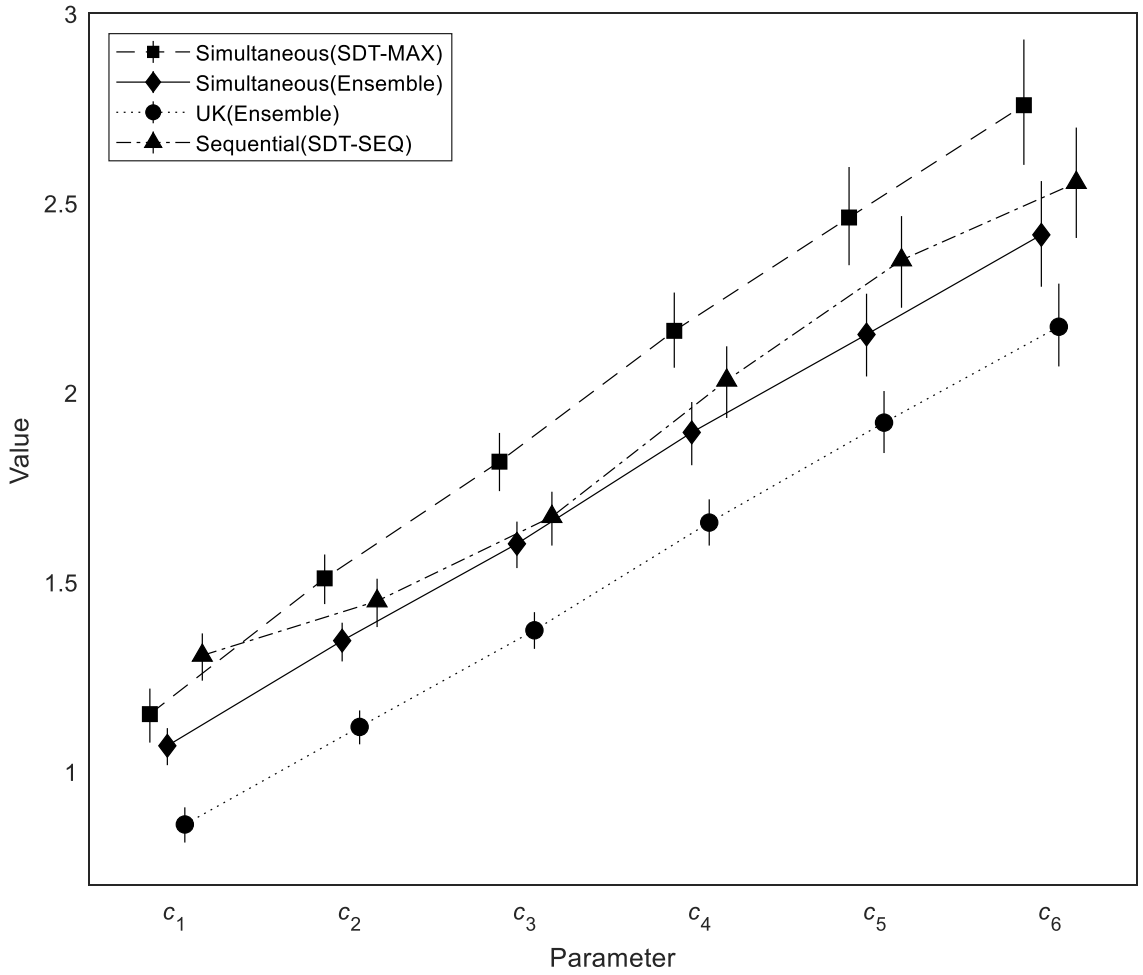


Figure 3.6 shows the estimated decision criteria (c_1 – c_6) and bootstrap 95% CIs for each model when fit to the relevant presentation format data. For c_1, i.e. the choose/no-choose threshold, responding was most conservative for sequential stopping rule presentation, followed by simultaneous presentation (either SDT-MAX or Ensemble estimates), followed by UK lineup presentation. This was indicated by no overlap in the 95% bootstrap CIs for each presentation format. Examination of Figure 3.6 beyond c_1 shows that the UK lineup was the most lenient procedure in general, with none of the bootstrap 95% CIs for UK lineup c_2 – c_6 overlapping with the those for sequential stopping rule presentation or simultaneous presentation as estimated by SDT-MAX. The c_6 bootstrap 95% CI for UK lineup presentation

overlapped with c_6 for simultaneous presentation as estimated by the Ensemble model. The bootstrap 95% CIs for sequential $c_2 – c_6$ are more lenient in general than those for simultaneous $c_1 – c_6$ as estimated by SDT-MAX, although only the difference at c_2 is marginally significant. The bootstrap 95% CIs for sequential stopping rule $c_2 – c_6$ are more conservative, but not significantly, than those for simultaneous $c_2 – c_6$ as estimated by the Ensemble model.

Table 3.3 shows the likelihood ratio tests for each criteria and presentation format. Using a Bonferroni corrected *p*-value of .05/6 = .008 to correct for multiple comparison, the results are largely in alignment with the bootstrap 95% CIs. Responding was significantly more conservative at c_1 for sequential stopping rule presentation, followed by simultaneous presentation, followed by UK lineup presentation, irrespective of which model was fit to the simultaneous data. The UK lineup was also significantly more lenient than the other procedures from $c_2 – c_6$, irrespective of the simultaneous lineup model. Some differences were significant according to the likelihood ratio tests even though their bootstrap 95% CIs did not overlap. In particular, $c_2 – c_6$ were significantly more conservative for sequential stopping rule presentation compared to simultaneous presentation as estimated by the Ensemble model, and c_2 and c_2 were significantly more lenient for sequential stopping rule presentation compared to simultaneous presentation as estimated by SDT-MAX.

These results support our response bias hypothesis; that responding would be most lenient for UK lineup presentation, followed by simultaneous presentation, then sequential stopping rule presentation. Our findings concur with previous model-based research comparing response bias between UK lineup presentation and simultaneous lineup presentation (Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019) and simultaneous presentation and sequential stopping rule presentation (Horry et al., 2012b; Kaesler et al., 2020).

The absence of a stopping rule is the most plausible explanation for more lenient responding on the UK lineup and simultaneous lineup compared to the sequential lineup (Kaesler et al., 2020; Palmer & Brewer, 2012; Seale-Carlisle et al., 2019). Under the stopping rule constraint, an identification decision terminates the lineup, so the optimal strategy to avoid errors is to set a higher criterion for identification compared to a procedure on which an identification decision has no effect on the possibility of viewing remaining lineup items.

Table 3.3

Likelihood Ratio Tests for Criteria between Presentation Formats

	Sim(MAX) vs Seq		Sim(ENS) vs Seq	
	χ^2	*p*	χ^2	*p*
c_6	4.18	<.05	1.86	.17
c_5	5.14	.08	9.73	<.05
c_4	8.02	.05	9.77	<.05
c_3	12.38	<.05	9.96	<.05
c_2	15.23	<.05	15.52	<.05
c_1	70.53	<.001	66.51	<.001

	Sim(MAX) vs UK(ENS)		Sim(ENS) vs UK(ENS)	
	χ^2	*p*	χ^2	*p*
c_6	45.98	<.001	7.70	<.05
c_5	63.20	<.001	12.41	<.01
c_4	86.83	<.001	22.07	<.001
c_3	108.30	<.001	33.57	<.001
c_2	119.84	<.001	45.18	<.001
c_1	121.17	<.001	51.81	<.001

	UK(ENS) vs Seq	
	χ^2	*p*
c_6	18.01	<.001
c_5	39.36	<.001
c_4	49.51	<.001
c_3	54.39	<.001
c_2	79.99	<.001
c_1	147.15	<.001

It is less obvious why the UK lineup should lead to more lenient responding than the simultaneous lineup. One possibility is that allowing participants to revisit lineup members before identifying produces a demand characteristic that encourages choosing. Perhaps participants interpret the option to revisit as feedback that the target is present in the lineup (Quinlivan et al., 2017). If this leads them to revise upward their belief in the prior probability that the target was present in the lineup, the optimal strategy is to choose more readily. However, this explanation alone cannot account for the pattern of results observed here because Seale-Carlisle et al. (2019) also observed lenient responding for UK lineup presentation without the option to revisit compared to simultaneous presentation. Another possibility is that the requirement to lap twice leads to lenient responding, consistent with research examining shifts in response bias from lap one to lap two in a sequential lineup procedure conducted without a stopping rule (Horry et al., 2015; Steblay et al., 2011a). However, it does not appear that responding was more lenient for the two lap compared to the one lap condition in Seale-Carlisle et al. (2019).

3.6.4 Revisiting Items on the UK Lineup

The opportunity to revisit lineup members at will before identifying is a unique aspect of the UK lineup procedure compared to the simultaneous and sequential procedure and reflects the way real police lineups are conducted in the United Kingdom (Police and Criminal Evidence Act, 1984). We sought to understand whether there are differences in underlying discriminability and response bias between those participants who revisited lineup members before identifying and those who did not. We partitioned the UK lineup data into two groups based on whether participants did or did not revisit lineup members before identifying. The data, with three confidence categories are shown in Table 3.4. We then fit the Ensemble model to each dataset, estimating d_t, s_t, and three decision criteria ($c_1 - c_3$).

Table 3.5 shows the fit statistics and parameter estimates for each condition. The Ensemble model fit the "revisited" data well, but failed to fit the "did not revisit" data. Despite the failure of the Ensemble model to fit one condition, we proceed with the parameter estimates from this model because this facilitates straightforward comparison between each group. Additionally, the Ensemble model provided a superior fit to the original UK lineup dataset compared to SDT-MAX. It appears that underlying discriminability (d_t) is greater for participants who did not revisit items compared to participants who did, but response bias (c_1 – c_3) appears similar between the two groups.

Table 3.4

Decision Outcome Frequencies for Revisit vs Not Participants on the UK Lineup

UK Lineup – revisited items before identifying				
Confidence	100 – 91	90 – 71	70 – 0	Reject
TP – Target ID	29	65	52	16
TP – Foil ID	9	12	36	
TA – Foil ID	15	41	68	25

UK Lineup – did not revisit items before identifying				
Confidence	100 – 91	90 – 71	70 – 0	Reject
TP – Target ID	169	128	56	18
TP – Foil ID	4	38	36	
TA – Foil ID	24	50	58	51

Table 3.5

Parameter Values for Revisit vs Not Participants on the UK Lineup

	Revisited	Did Not Revisit
d_t	1.56	2.01
s_t	.54	.75
c_3	1.94	1.92
c_2	1.40	1.34
c_1	.81	.90
χ^2	5.74	14.54
df	4	4
p	.22	.01

We compared the parameter values using likelihood ratio tests, comparing the fit of an unconstrained model fit to both datasets simultaneously to a number of constrained models where one parameter was fixed to be equal across the datasets. Table 3.6 shows the likelihood ratio tests. The difference in underlying discriminability between the groups was significant, the effect size was medium, g = .66. We note that the effect size of the d_t advantage for those who did not revisit items compared to those who did is larger than the effect size of the difference between simultaneous and sequential presentation observed here. It is unclear whether some participants revisit items due to poor memory caused by pre-test factors, such as poor encoding of the target or memory decay, or whether the act of revisiting lineup members itself impairs memory through an interference-type mechanism.

Table 3.6 shows that responding was significantly more lenient at the choose/no-choose threshold (c_1) for participants who revisited items before identifying, although the difference in the parameter estimates between each group is small. This indicates that a subset of participants with poor memory also chose more readily, an undesirable outcome.

Table 3.6

Likelihood Ratio Tests for Revisit vs Not on the UK lineup

	χ^2	*df*	*p*
d_t	23.72	1	<.001
s_t	2.46	1	.12
d_t and s_t	23.76	2	<.001
c_3	.05	1	.81
c_2	1.88	1	.39
c_1	8.70	1	<.05

There were no significant differences for c_2 and c_3. The estimates of c_1 for each group are similar to the estimate for UK presentation overall (c_1 = .86). This implies that act of revisiting itself may not cause more lenient responding. It could be that the opportunity to

revisit prompts choosing, or that some other aspect of the procedure for lenient responding compared to simultaneous and sequential presentation.

To disentangle these effects, an experimental test is required to understand the effects of allowing witnesses to revisit items. A future experiment could compare a UK lineup on which participants are able to revisit items to one on which they are not. With sufficient power, it would also be possible to investigate whether the nature of the relationship between underlying discriminability and response bias change with the number of items revisited.

3.7 General Discussion

This study compared simultaneous, sequential stopping rule and UK lineup presentation in terms of underlying discriminability and response bias by fitting signal-detection based measurement models to experimental data. We sought to investigate the possibility that some aspect of the UK lineup task other than the sequential presentation of items may impair memory, in addition to providing a test of Diagnostic Feature Detection Theory (DFDT). To our knowledge, this is the first study to compare the sequential stopping-rule lineup procedure and the UK lineup procedure. We found that simultaneous presentation was associated with greater underlying discriminability than sequential stopping rule presentation when the Ensemble model was fit to the simultaneous and UK lineup data, consistent with the predictions of DFDT. There were no significant differences in underlying discriminability between the simultaneous lineup and the UK lineup or the UK lineup and the sequential stopping rule lineup. Sequential stopping rule presentation was associated with the most conservative responding, followed by simultaneous presentation, followed by UK lineup presentation. The UK lineup was more lenient than the other presentation formats at all decision criteria, while simultaneous and sequential stopping rule presentation differed largely in the setting of the choose/no-choose threshold.

3.7.1 Diagnostic Feature Detection Hypothesis

Our results were consistent with a key prediction of the diagnostic feature detection hypothesis (DFDT), that simultaneous presentation is associated with greater underlying discriminability than sequential stopping rule presentation. However, our results failed to conform to the DFDT predictions that the simultaneous lineup would be associated with greater underlying discriminability than the UK lineup, and that the UK lineup would be associated with greater discriminability than the sequential stopping rule lineup. The pattern of results can be interpreted as providing modest evidence in favour of DFDT. The additional opportunity afforded to compare items on the UK lineup meant that its decrement in underlying discriminability compared to the simultaneous lineup was smaller than the decrement for the sequential stopping rule lineup.

We are somewhat constrained in evaluating DFDT by the limitations of a measurement model approach. As in our previous study (Kaesler et al., 2020), it is possible that there is a small beneficial effect of comparing across lineup members on underlying discriminability, which was counteracted by other aspects of the tasks under investigation. We may have also lacked power to detect a small effect. Signal detection models measure underlying discriminability for a given lineup procedure, but cannot separately quantify the benefit to underlying discriminability of comparing across lineup members and the harmful effect of, for example, memory interference that might increase with the number of items viewed. Teasing out these differences requires careful experimentation. As we did not find patterns in the data that clearly contradict DFDT, such as a robust underlying discriminability advantage for sequential stopping rule presentation over simultaneous presentation, our results present little problem for the process account specified by the theory. Further experimental research is required to understand whether there are empirical results at this

level of explanation that are inconsistent, rather than inconclusive, with respect to the predictions of DFDT.

The utility of the feature detection explanation is also difficult to test using signal detection models because this class of model does not formalise the exact processes that give rise to familiarity. That, is signal detection models are agnostic about what a feature is, what it means to detect one, and how that might enhance underlying discriminability. If the predictions of DFDT prove to be robust at the measurement level, then a model framework that does formalise the feature detection process and its relationship to underlying discriminability may be required to further test the theory.

3.7.2 The UK Lineup Procedure

Our results are consistent with the notion that the sequential presentation of items is the main factor responsible for the underlying discriminability decrement previously observed on the UK lineup compared to the simultaneous lineup, although we cannot draw strong conclusions as we failed to replicate this decrement in this study (Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019). In addition to sequential presentation, we isolated two critical aspects of the UK lineup procedure, the requirement for witnesses to lap through the lineup items twice before identifying and the ability for witnesses to revisit lineup members before making their identification. We found no evidence that the requirement to defer the identification decision, lapping twice through the lineup items, impaired underlying discriminability. This weighs against the possibility that the decrement observed in other studies (Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019) was due to the UK lineup interfering with memory for the target, rather than the sequential presentation of items (Kaesler et al., 2020). However, the duration of our UK lineup procedure was also much shorter than the more externally valid procedure used in Seale-Carlisle et al. (2019) and Seale-Carlisle and Mickes (2016). Each of our six photographic items was shown for five

seconds each while their nine photographic items were shown for fifteen seconds each. The longer duration for item presentation in Seale-Carlisle et al. (2019) and the greater number of lineup items may have led to more interference than in this study, which might have contributed to the larger decrement in underlying discriminability for UK lineup presentation in that study.

The UK lineup procedure is complex to conduct relative to the simultaneous procedure, involving a presentation phase and a decision phase. It also requires specialised software (VIPER and PROMAT) and the maintenance of a library of appropriate foil videos. The evidence presented in this study and others suggests that this increased complexity is not associated with a corresponding benefit to the memory strength of witnesses. This demonstrates the potential for unintended consequences when memory test procedures are developed without reference to relevant research in recognition memory (Pike et al., 2002).

3.7.3 The Sequential Lineup as applied in the United States

While the stopping rule is in integral part of the sequential procedure described by Lindsay and Wells (1985), surveys of US police jurisdictions reveal that the stopping rule is not enforced (Wells et al., 2015a). The lineup does not terminate when an identification is made and witnesses are offered the opportunity of a second lap. They are also able to change their identification decision, which is recorded when their evidence is presented in court. Our results conform to the previous research (Wells et al., 2015a) finding that the omission of the stopping rule in US sequential lineups may lead witnesses to choose more readily than might be expected based on results of experimental studies, which typically do employ the stopping rule (Kaesler et al., 2020; Palmer & Brewer, 2012).

The sequential lineup was adopted by jurisdictions based on the finding that it produced a higher diagnosticity ratio, the ratio of the target identification rate to the innocent suspect identification rate, than the simultaneous lineup (Steblay et al., 2011b). Subsequent

research demonstrated that increases in this ratio largely reflect increasingly conservative responding rather than an increase in memory strength (Mickes et al., 2012). Some researchers have argued that the sequential procedure may still be preferred because conservative responding necessarily reduces innocent suspect identifications (Wells, 2014). The promised reduction of innocent suspect identifications relative to simultaneous presentation may not be realised when sequential lineups are conducted without a stopping rule (Wells et al., 2015a), which may be of concern to policy makers in light of a lack of enforcement of this aspect of the lineup procedure.

3.7.4 Empirical and Underlying Discriminability

Previous research has demonstrated that empirical discriminability, measured by ROC analysis, and underlying discriminability, measured by signal detection models, can be dissociated (Wilson et al., 2019; Wixted & Mickes, 2018). We found such a dissociation here; there was no significant difference in empirical discriminability between simultaneous and sequential stopping rule presentation but there was a small underlying discriminability advantage for simultaneous presentation. This further reinforces the necessity of adopting a measurement model approach when comparing the underlying signal detection parameters between lineup tasks that differ in structure (Kaesler et al., 2020).

3.7.5 Model Selection

Our results with respect to simultaneous presentation highlight important issues regarding model selection. First, SDT-MAX failed to fit both the simultaneous and UK lineup tasks and was outperformed by the Ensemble model, consistent with prior research (Wixted et al., 2018). This implies that SDT-MAX may not be a suitable candidate model for characterising data from these tasks. However, SDT-MAX successfully fit data from a wide range of simultaneous lineup experimental data in a prior study we conducted (Kaesler et al., 2020), although some of the older datasets analysed in that study may have had insufficient

power to reject the model. Another issue is that underlying discriminability was greater in this study than in some previous model-based studies (Seale-Carlisle et al., 2019; Wixted et al., 2018). This led to relatively few foil identifications at high confidence. SDT-MAX may have been more affected by low power in these cells than the Ensemble model, which might explain its apparent overestimation of performance when responding was conservative. SDT-MAX may provide an adequate fit to data when performance is poorer, whereas the Ensemble model may provide an adequate fit over a greater range of possible performance. In all, our results provide further evidence that the Ensemble model is superior to SDT-MAX for characterising data from simultaneous and UK lineups (Wixted et al., 2018).

Second, the underlying discriminability advantage for simultaneous compared to sequential presentation was evident when the Ensemble model was fit to the simultaneous data but not when SDT-MAX was fit to the simultaneous data. In this study, we adjudicated between the models based on goodness-of-fit. For our data, it is principled to prefer the results when the Ensemble model is the simultaneous and UK lineup model because its decision rule is a plausible characterisation of both procedures and it provides a better fit to the data than SDT-MAX. Our results depended on the model fit to the data, which highlights the possibility that the conclusions of future lineup studies could be influenced by model selection, particularly when more than one model fits the data. In the case where model selection leads to different results, it would be prudent to report both versions of an analysis in order to present the most complete account.

3.7.6 Conclusion

This study compared simultaneous, sequential stopping rule and UK lineup presentation in terms of underlying discriminability and response bias. Consistent with Diagnostic Feature Detection Theory, we found that simultaneous presentation was associated with greater underlying discriminability than sequential stopping rule presentation.

The pattern of results with respect to the UK lineup implied that the performance decrement observed in previous research (Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019) may be due to the sequential presentation of items rather than an interference-type effect resulting from other aspects of the procedure. Further work is required to understand how allowing witnesses to revisit items on sequentially presented lineup tasks affects their decision behaviour.

Chapter 4

4.1 Preface to Study Three

Study one developed and introduced a model for measuring discriminability and response bias on the sequential stopping rule lineup, SDT-SEQ. This model was used in studies one and two to compare performance on the simultaneous lineup to the sequential stopping rule lineup and the UK lineup. One aspect of the sequential stopping rule lineup unexplored in studies one and two is the extent to which discriminability and response bias change over the course of the task. Understanding this aspect of the task may prove useful when developing policy about where the suspect should be placed in the lineup.

In order to address the question it is necessary once again to consider the match between the task, measurement model and research question. In study three, the sequential stopping rule lineup is conceptualised as a series of yes/no decision tasks, on which a "yes" response terminates the procedure. Thus, the probability of identifying an item from the sequential lineup is joint probability of not identifying the previous items and the probability of identifying the current item. This means that it is necessary to consider not only the position of the target (or innocent suspect) in the lineup, which is accounted for by SDT-SEQ, but also the position at which identifications are made, which is not accounted for by SDT-SEQ. Addressing the issue of position effects in the sequential stopping rule lineup therefore requires the development of a new model. Study three develops the Independent Sequential Lineup (ISL) model, which gives the probability of each lineup decision outcome at each target and identification position, and a signal detection implementation, SDT-ISL, which converts these probabilities into discriminability and response bias. The ISL model and SDT-ISL are used to reanalyse data from a recent study that investigated position effects on this task (Wilson et al., 2019), in addition to newly collected experimental data. This experiment followed the same procedure for administering the sequential stopping rule lineup

and used the same stimuli as the experiments in studies one and two. The data structure of the ISL model required a large sample size to ensure adequate counts in all cells for model fitting (n = 7,204)

4.2 Statement of Authorship

Title of Paper	A Model of Position Effects in the Sequential Lineup
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	

Principal Author

Name of Principal Author (Candidate)	Matthew Kaesler		
Contribution to the Paper	Designed the study, collected the experimental data, fit models to data and wrote additional code for conducting hypothesis tests, conducted hypothesis tests on the Wilson et al. (2019) data, and wrote the manuscript.		
Overall percentage (%)	70		
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	13/7/21

Co-Author Contributions

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	Professor John Dunn
Contribution to the Paper	Developed mathematical model of sequential position effects used to analyse the data, wrote code for fitting it to data, extracted and conducted preliminary analysis of data from Wilson et al. (2019), and provided feedback on manuscript drafts.

Signature		Date	13/7/2021

Name of Co-Author	A/Professor Carolyn Semmler		
Contribution to the Paper	Supervised all aspects of study design and execution and provided feedback on manuscript drafts		
Signature		Date	23/02/2021

A Model of Position Effects in the Sequential Lineup

by

Matthew Kaesler[1], John C. Dunn[2,3], Carolyn Semmler[1]

[1] University of Adelaide

[2] University of Western Australia

[3] Edith Cowan University

Address for correspondence:

Mr Matthew Kaesler
University of Adelaide
North Terrace
Adelaide, SA, 5005
Email:

Publication Details

Unpublished manuscript

4.3 Abstract

How do underlying discriminability and response bias change over the course of the sequential lineup[5]? To explore this question, we developed and applied the Independent Sequential Lineup (ISL) model, which accounts for both identification position, the position at which the witness identifies a lineup item as the target, and target position, the position at which the target or suspect appears. We conducted a large sample sequential lineup experiment (n = 7,204) and reanalysed data from two similar experiments recently conducted by Wilson, Donnelly, Christenfeld and Wixted (2019; *Journal of Memory and Language, 104*, 108-125). There was a small increase in underlying discriminability from serial position one to position two in our experiment, but not in the experiments from Wilson et al., contrary to their original analysis. Responding became more conservative with serial position when a foil was identified prior to the presentation of the target. Failing to identify a target led to a conservative shift in response bias for the next item, although this effect weakened as the lineup progressed. Results are discussed in terms of their implications for diagnostic feature detection theory and for the placement of the suspect in real police lineups conducted using the sequential procedure.

[5]As for study one, the term "sequential lineup" in this study refers to the procedure described by Lindsay and Wells (1985), conducted with a stopping rule.

4.4 Introduction

The sequential lineup (Lindsay & Wells, 1985) was developed to reduce the risk of mistaken identification and has generated much research since its introduction. Of critical interest in this research has been the question of where to place the suspect in the sequence of lineup members, to minimise false alarms and maximise correct detections. Early sequential lineup research concluded that choosing rates did not differ by position if the number of lineup members was concealed from the witness (Lindsay & Wells, 1985; Sporer, 1993). However, subsequent studies employing choosing-rate based analyses (e.g. Gronlund et al., 2009), Receiver Operating Characteristic (ROC) analysis (e.g. Meisters et al., 2018) and signal detection modelling (Wilson et al., 2019) have reported results consistent with both differences in overall choosing by position and differential effects on target and non-target identification rates by position. Aside from practical concerns, these results raise important theoretical questions about why such effects might occur. According to signal detection theory, the different choosing rate patterns reported in previous literature imply two different explanations of decision behaviour: A change in overall choosing rates by position corresponds to a change in response bias, while a differential effect on target and non-target identifications by position corresponds to a change in underlying discriminability (Wixted & Mickes, 2018). Given the variety of analysis techniques used in existing research and the mixed pattern of results, the extent to which response bias and underlying discriminability change over the course of the sequential lineup remains unclear. In this paper, we aim to develop a model of the sequential lineup that can be used to estimate the signal detection parameters at each serial position. This will describe the nature of the effects that occur. We fit the model of the sequential lineup to new experimental data and the data from Wilson, Donnelly, Christenfeld and Wixted (2019). This will clarify the conclusions reached about

position effects in previous studies and highlight the areas where models need to be developed to understand why they occur.

4.4.1 Choosing Rate-Based Sequential Position Studies

Prior to the introduction of ROC analysis to lineup research (Mickes et al., 2012), position effects were investigated using choosing rate based measures. Some studies compared the rates of target identification and innocent suspect identification at each position (Clark & Davey, 2005; Memon & Gabbert, 2003), while others computed the "diagnosticity ratio", the ratio of the target identification rate to the innocent suspect identification rate, at each position (Carlson et al., 2008; Gronlund et al., 2009). Memon and Gabbert (2003) and Clark and Davey (2005) found that the target identification rate was lower when a "next-best" familiar foil appeared before the target, as opposed to after the target. This highlights the effect of the stopping-rule constraint; witnesses see no more members once they have made an identification and thus a familiar foil place before the target may cause them to "spend" their identification choice. Clark and Davey (2005) also found evidence of a strict-to-lenient criterion shift in terms of overall choosing rates in their second experiment.

Carlson et al. (2008) experiment two found that both overall choosing and target identifications increased on target present sequential lineups in late positions compared to earlier positions. There was no difference in overall choosing by position on target absent lineups but innocent suspect identifications decreased in late positions compared to earlier positions. Overall, the diagnosticity ratio increased with serial position. Gronlund et al. (2009) also reported a diagnosticity increase for late positions in the sequential lineup.

These findings tell us little about underlying discriminability and response bias as changes in choosing rates by position reflect some combination of a change in willingness to choose and a change in ability to discriminate target and non-target items. That is, analyses

based on choosing rates confound discriminability and response bias (Mickes & Wixted, 2012). This led to a different approach to describing position effects.

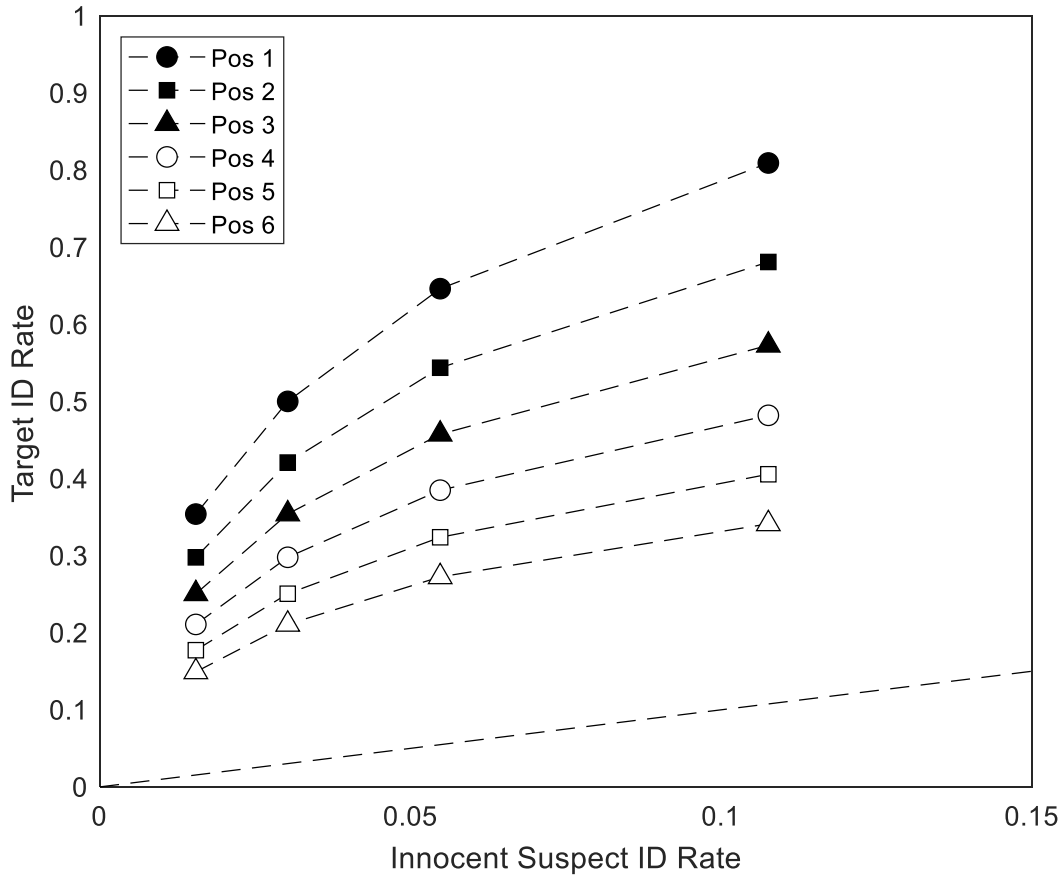
4.4.2 Studies Employing ROC Analysis

Subsequent research has investigated sequential position effects using Receiver Operating Curve (ROC) analysis, which provides a method for quantifying discriminability independent of response bias (Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al., 2012; Meisters et al., 2018; Neuschatz et al., 2016). To perform ROC analysis, the target identification rate is plotted against the innocent suspect identification, i.e. the false identification rate, for the subset of trials where the target appeared in a particular position. For studies without a designated innocent suspect, the false identification rate for a position can be estimated by dividing the target absent foil identification rate for that position by the lineup size. Empirical discriminability (Wixted & Mickes, 2018) at each position can then compared based on area under the ROC curve (AUC). Three of these studies reported greater empirical discriminability for late positions in the lineup compared to early positions (Gronlund et al., 2012; Meisters et al., 2018; Neuschatz et al., 2016). The remainder reported no difference in empirical discriminability between early and late positions (Carlson & Carlson, 2014; Dobolyi & Dodson, 2013).

However, it is important to distinguish between empirical discriminability, measured by ROC analysis, and underlying discriminability, which is measured with respect to a particular signal detection model (Wixted & Mickes, 2018). These two measures may be dissociated when structural features of a memory test affect ROC shape. Figure 4.1 shows ROC curves from data simulated according to a model of the stopping-rule sequential lineup called SDT-SEQ, described in detail in Kaesler et al. (2020). In each simulated dataset, n = 4000 (TP = 2000, TA = 2000) and the target was fixed in a single position. SDT-SEQ instantiates the stopping rule constraint through a “first above criterion” decision rule.

Figure 4.1

ROC Curves Simulated from SDT-SEQ with constant Discriminability



The first lineup item that exceeds the choose/no-choose threshold will be identified, even if a more familiar item appears later in the lineup. If no lineup member exceeds the criterion, then the lineup will be rejected. Under the assumption that all foils are drawn from one distribution, underlying discriminability is summarised by free parameters d_t and s_t, the mean and standard deviation of the target distribution, respectively. Response bias is summarised by a number of decision criteria denoted $c_1 \ldots c_n$, where c_1 is threshold separating an identification from a rejection. For the six simulated datasets in Figure 4.1, $d_t = 1.7$, $s_t = 0.8$ and $c = 1.0$, 1.4, 1.7 and 2.0. It is evident from Figure 4.1 that empirical discriminability as measured by AUC decreases as the target appears later in the lineup, even though

underlying discriminability (d_t) is the same across all datasets. This is attributable to the stopping rule constraint of the sequential lineup task (Rotello & Chen, 2016; Wilson et al., 2019) which reduces AUC as the lineup progresses because witnesses who identify early lose the opportunity to identify later in the lineup, but are still counted in the total number of trials for calculating target and innocent suspect identification rates. This implies that target position affects the shape of the sequential ROC and resulting empirical discriminability independent of underlying discriminability. Thus, the results of studies employing ROC analysis alone cannot clearly establish whether memory strength changes over the course of the sequential lineup. Consequently, it is necessary to employ signal detection measurement models to measure underlying discriminability and response bias on the sequential lineup.

4.4.3 Model-Based Studies

Two studies have adopted a model-based approach to understanding sequential position effects, employing the integration model (SDT-INT) and the independent observations model (SDT-MAX), which are described in Kaesler et al. (2020). Carlson et al. (2016) reported greater underlying discriminability for position two than position five based on fits of SDT-MAX when participants were not told how many lineup members to expect before commencing the procedure. Horry et al. (2012b) reported that underlying discriminability was greater at position two than position six based on fits of SDT-INT when participants were told how many photos to expect, either 6, 12 or 30. Both Horry et al. (2012) and Carlson et al. (2016) reported a strict-to-lenient criterion shift from early to late positions based on choosing rates, although this was not reflected in the estimated model parameters.

These results are called into question because SDT-INT and SDT-MAX are not appropriate models for characterising data from a sequential lineup (Kaesler et al., 2020). This is because the model's decision rules depend on access to the familiarity for all lineup items in order to isolate the most familiar item (SDT-MAX) or calculate the summed

familiarity of the items (SDT-INT). This is not compatible with the structure of the sequential lineup, on which an identification can be made before all items are shown. Relatedly, the position of the target is not represented within these models' decision rules, because it does not affect the calculation of the summed familiarity and maximum familiar decision variables. This means that these models may not recover accurate parameter estimates when fit to sequential lineup data in the aggregate and may not capture effects of target position on discriminability and response bias.

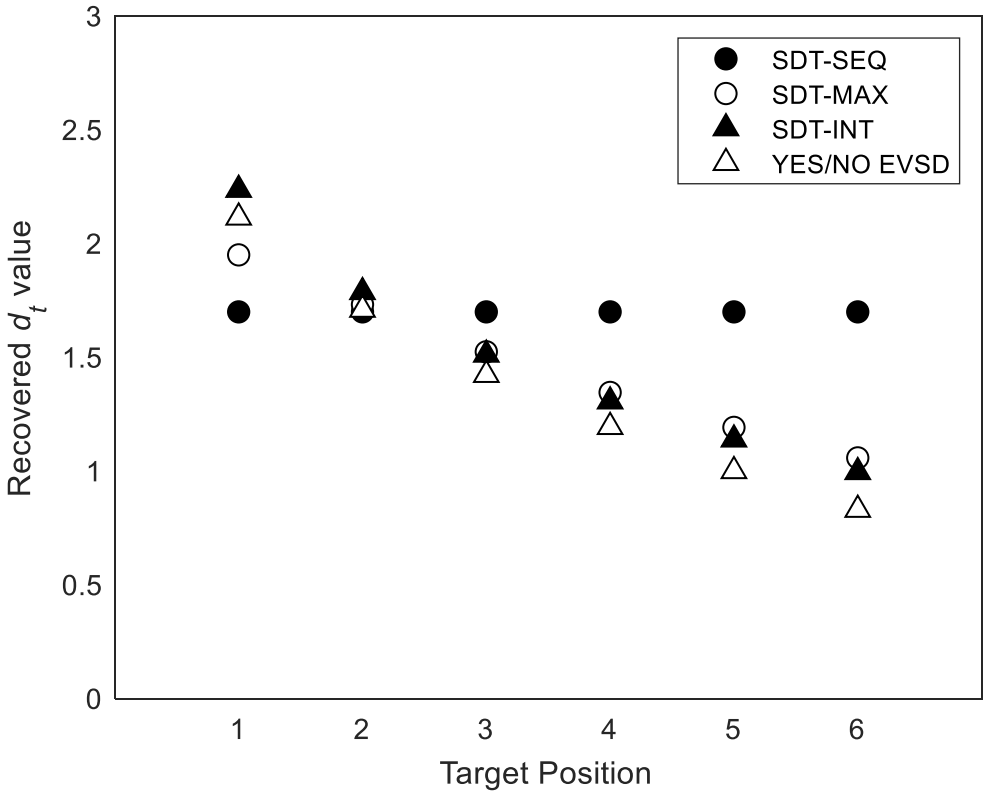
Following Mickes et al. (2014), Meisters et al. (2018) calculated underlying discriminability by target position according to equations from a signal detection model of the yes-no task, $d' = \Phi^{-1}(\text{Target Identification Rate}) - \Phi^{-1}(\text{False Identification Rate})$, where Φ^{-1} is the inverse normal cumulative distribution function. They found that underlying discriminability increased over the course of a four-item sequential lineup. This estimate of underlying discriminability may not be accurate for the same reason as for SDT-MAX and SDT-INT; the yes-no involves the presentation of a single item at test rather than a set of items and therefore does not account for the position of the target or specify a decision rule that captures the stopping rule constraint.

Figure 4.2 shows d_t, i.e. the mean of the target distribution, recovered by SDT-SEQ, SDT-MAX and SDT-INT when fit to the fixed target position datasets generated by SDT-SEQ for Figure 4.1, in addition to the yes-no d' calculation used by Meisters et al. (2018). Recall that the data were generated from SDT-SEQ with $d_t = 1.7$ and $s_t = 0.8$ at all target positions, so we fixed s_t to 0.8 when fitting SDT-MAX and SDT-INT to ensure that estimates of d_t at each position were comparable between the models. The yes-no calculation assumes an underlying equal variance model where $s_t = 1$. When fit to its own data, SDT-SEQ unsurprisingly recovers $d_t = 1.7$ for each dataset because the possible positions of the target in the lineup are specified when the model is fit to the data. SDT-MAX, SDT-INT and the yes-

no equal variance signal detection model do not recover $d_t = 1.7$ for each dataset, likely because these models' decision rules do not account for position of the target and its interaction with the stopping rule constraint.

Figure 4.2

Estimates of d_t from fitting Various Lineup Models to the Simulated Data from Figure 4.1



4.4.4 SDT-SEQ and Wilson et al. (2019)

Wilson et al. (2019) recently conducted an analysis of sequential position effects, employing ROC analysis to quantify empirical discriminability and SDT-SEQ to estimate underlying discriminability. They sought to address the shortcoming of previous work by fully randomising target position, rather than fixing the target in one early and one late position, as in some previous studies (e.g. Gronlund et al., 2009; Memon & Gabbert, 2003), which required a large sample size (Experiment 1 $n = 6530$, Experiment 2 $n = 1966$ lineup trials). The results of experiment one showed a dissociation between empirical and

underlying discriminability: Empirical discriminability decreased with target position, while underlying discriminability increased from position one to position two, but changed little beyond. Empirical discriminability also decreased by position in experiment two but underlying discriminability did not vary. This is consistent with the simulated ROC data from SDT-SEQ shown in Figure 4.1. Response bias varied by target position across both experiments, but there was no obvious pattern of increase or decrease.

The results of Wilson et al. (2019) experiment one with respect to underlying discriminability support a key prediction of Diagnostic Feature Detection Theory (DFDT Wixted & Mickes, 2014); that underlying discriminability should increase over the course of the sequential lineup. DFDT proposes that the opportunity to compare across lineup items increases underlying discriminability because witnesses are able to isolate the unique, diagnostic features shared by their memory of the perpetrator and the lineup items. As a result, underlying discriminability should increase over the course of the sequential lineup because the presentation of each successive lineup item provides additional opportunity to isolate diagnostic features.

Wilson et al. (2019) provided a valuable test of DFDT and empirically demonstrated the dissociation between underlying and empirical discriminability on the sequential lineup task. However, two aspects of Wilson et al.'s (2019) research may have affected their results. First, they did not employ a stopping rule when administering their sequential lineup task, which allowed participants to make multiple identifications. Second, SDT-SEQ is not an ideal model for investigating sequential position effects, as outlined below.

4.4.4.1 Wilson et al.'s (2019) Sequential Lineup Task.

Wilson et al. (2019) showed that the non-monotonic sequential lineup ROC curve simulated in Rotello and Chen (2016) could be recovered empirically by changing the confidence threshold for counting identifications. Conducting this analysis required the

sequential lineup task to be conducted without a stopping rule. Participants provided a yes/no identification decision and confidence rating from -100 (sure item is not the target) to +100 (sure item is the target) for each lineup item. That is, they were able to identify multiple items, in contrast to the stopping rule sequential lineup, where the first identification terminates the procedure. For experiment one, the "neutral" cut point was to count as an identification the first item with a confidence rating > 0, the "conservative" cut point was to as an identification the first item with a confidence rating > 80 and the "liberal" cut point was to count as an identification the item response > -80. Wilson et al. (2019) Figure 7A shows the resulting ROC curve, which is qualitatively similar to the "first-above-criterion" decision rule ROC curve shown in Rotello and Chen (2016) Figure 6.

For the position effect analyses, Wilson et al. (2019) used the neutral cut point to count identifications. Implicit in these analyses is the notion that imposing a post-hoc stopping rule on the data is equivalent to conducting a sequential lineup with a stopping rule. This assumption may be unsound because participants may behave differently when their task is to provide a binary decision and confidence rating for sequentially presented items, compared to when their task is to decide whether the current sequentially presented item is a strong enough match to memory that they are willing to forgo viewing the remaining items. On this basis, we might expect that conducting a sequential lineup task with a stopping rule would lead to more conservative responding than the Wilson et al. (2019) task, because identifying an item carries a higher cost in the stopping rule task. Additionally, allowing participants to make multiple identifications from a lineup is known to reduce accuracy compared to allowing participants only a single identification (Horry et al., 2015; Steblay et al., 2011a), which may indicate reduced underlying discriminability for procedures that allow multiple identifications.

4.4.4.2 Issues with the application of SDT-SEQ

To estimate signal detection parameters at each position, Wilson et al. (2019) constructed subsets of their data, i.e. trials on which the target appeared in position *j*, and fit SDT-SEQ to each subset, specifying that the target appeared in only one position. This approach has two major limitations, outline below.

Many lineup studies that employ fair target absent (TA) lineups, including this study and Wilson et al. (2019), do not designate an innocent suspect on TA trials. It is not obvious how to extract the subset of TA trials where the target was in position *j* from experiments with this design, because there is no designated innocent suspect to stand in for the target.

One possibility is to count the TA foil identifications that occur at each serial position. However, this subsets TA identification by *identification* position *i* rather than target position *j*. Under this rule, it is also unclear how correct rejections should be allocated to a subset where the target appeared in position *j*, as is possible for misses on TP lineups. As an approximation, the total number of correct rejections could be divided by the lineup size (Wixted & Mickes, 2012). In this case, the correct rejection frequencies in each subset are no longer "observed", which compromises the χ^2 test used to assess model fit.

Another possible method is to conceive of each subset as a separate experiment where the target was fixed in position *j*. TA foil identifications at position *j* are treated as false identifications, with TA foil identifications at positions $\neq j$ making up the remainder of TA identification responses. Along with the total number of correct rejections, this method allows for the calculation of false identification rates for plotting ROCs but it is not ideal for model fitting analyses. Using this method, TA foil identifications and correct rejections are repeated in each position subset, violating the assumption of independence and compromising the χ^2 goodness-of-fit test. Our understanding is that Wilson et al. (2019) employed this method of

subsetting when analysing their data by position (B. M. Wilson, personal communication, July 2, 2019).

It seems that the sub-setting issue could be solved by designating an innocent suspect on TA trials, allowing false identifications, TA foil identifications and correct rejections to be counted when an innocent suspect position appears in *j* in the same way as target identifications, TP foil identifications and misses are counted on TP lineups. SDT-SEQ could then be fit to each subset with the target in a fixed position. However, this approach collapses over *identification* position.

SDT-SEQ assumes that an identification is equiprobable at each serial position for tractability. For example, a subset of lineup data when the target appeared in position 1 contains target identifications from position 1, but TP foil identifications from positions 2 to 6. As a result, differences in underlying discriminability between subsets of the data where the target appeared in position *j* may also reflect changes in response criteria across identification positions. It is therefore unclear whether Wilson et al.'s (2019) results for experiment one reflect a genuine increase in underlying discriminability with serial position, as predicted by the DFDT. To address these issues, an alternative modelling approach is required.

4.4.5 The Independent Sequential Lineup Model

We propose a probabilistic model, which we call the Independent Sequential Lineup (ISL) model, that accounts for both the position of the target *and* the position at which identifications are made. Table 4.1, presented on the following page, shows the data structure for this model.

Table 4.1

The Structure of the ISL Model

	Target Position						
	1	2	3	4	5	6	None
ID Position	$(1-G_{1k})$	$(1-F_{1k})$	$(1-F_{1k})$	$(1-F_{1k})$	$(1-F_{1k})$	$(1-F_{1k})$	$(1-F_{1k})$
	$G_1(1-F^*_{2k})$	$F_1(1-G_{2k})$	$F_1(1-F_{2k})$	$F_1(1-F_{2k})$	$F_1(1-F_{2k})$	$F_1(1-F_{2k})$	$F_1(1-F_{2k})$
	$G_1F^*_2(1-F^*_{3k})$	$F_1G_2(1-F^*_{3k})$	$F_1F_2(1-G_{3k})$	$F_1F_2(1-F_{3k})$	$F_1F_2(1-F_{3k})$	$F_1F_2(1-F_{3k})$	$F_1F_2(1-F_{3k})$
	$G_1F^*_2F^*_3(1-F^*_{4k})$	$F_1G_2F^*_3(1-F^*_{4k})$	$F_1F_2G_3(1-F^*_{4k})$	$F_1F_2F_3(1-G_{4k})$	$F_1F_2F_3(1-G_{4k})$	$F_1F_2F_3(1-G_{4k})$	$F_1F_2F_3(1-G_{4k})$
	$G_1F^*_2F^*_3F^*_4(1-F^*_{5k})$	$F_1G_2F^*_3F^*_4(1-F^*_{5k})$	$F_1G_2F^*_3F^*_4(1-F^*_{5k})$	$F_1F_2F_3G_4(1-F^*_{5k})$	$F_1F_2F_3F_4(1-G_{5k})$	$F_1F_2F_3F_4(1-F_{5k})$	$F_1F_2F_3F_4(1-F_{5k})$
	$G_1F^*_2F^*_3F^*_4F^*_5(1-F^*_{6k})$	$F_1G_2F^*_3F^*_4F^*_5(1-F^*_{6k})$	$F_1G_2F^*_3F^*_4F^*_5(1-F^*_{6k})$	$F_1F_2F_3G_4F^*_5(1-F^*_{6k})$	$F_1F_2F_3F_4G_5(1-F^*_{6k})$	$F_1F_2F_3F_4F_5(1-G_{6k})$	$F_1F_2F_3F_4F_5(1-F_{6k})$
Reject	$G_1F^*_2F^*_3F^*_4F^*_5F^*_6$	$F_1G_2F^*_3F^*_4F^*_5F^*_6$	$F_1F_2G_3F^*_4F^*_5F^*_6$	$F_1F_2F_3G_4F^*_5F^*_6$	$F_1F_2F_3F_4G_5F^*_6$	$F_1F_2F_3F_4F_5G_6$	$F_1F_2F_3F_4F_5F_6$

The columns correspond to each target position on TP lineups, with the final column corresponding to the "no target" case, i.e. TA lineups. The rows correspond to the position at which an identification was made, with the last row corresponding to no identification, i.e. a lineup rejection. It parameterises three decision outcomes; the probability of not identifying a target at identification position *i* = target position *j*, G_i, the probability of not identifying a foil at position $i < j$, F_i and the probability of not identifying a foil at position $i > j$, F_i^*. If there were a designated innocent suspect on TA trials, this would lead to the same data structure as the TP data for TA lineups, with an additional set of *G* parameters for the probability of not identifying the innocent suspect at position $i = j$. At each identification position, the probability of identifying an item is equal to the probability of not identifying the previous items and the probability of identifying the present item. Consider the example of a lineup with the target in position two. There is some probability $(1 - F_{1k})$ of incorrectly identifying the foil in position one as the target with confidence level *k*. The probability of correctly identifying the target in position two with confidence level *k* is equal to $F_1(1 - G_{2k})$, i.e. the probability of not identifying the first foil and the probability of identifying the target in position two. Following on, the probability of incorrectly identifying the foil in position three at confidence level *k* is equal to $F_1 G_2(1 - F_{3k}^*)$. We distinguish between foil identifications before and after the presentation of the target because presenting the target is a singular event in the lineup. Experiencing a target and not identifying may lead participants to reset their response criteria. For a lineup of *n* items, there are G_i and F_i parameters for each confidence category at all identification positions and F_i^* parameters for each confidence category at identification positions two to *n*. The model assumes that the false alarm rates, i.e. F_i and F_i^*, vary across identification position, but not across target position.

Each column in Table 4.1 contains conditional probabilities for a target position, i.e. each column sums to one. To convert these to absolute probabilities, each column entry can be multiplied by p_j, the probability that the target appears in position j.

4.4.5.1 The ISL Model reduces to SDT-SEQ

When there are no position effects, i.e. all $G_{ik} = G_k$, all $F^*_{ik} = F_{ik}$ and all $F_{ik} = F_k$ for all identification positions *i*, the ISL model reduces to SDT-SEQ (see Kaesler et al., 2020). Recalling that *pj* is the probability of the target appearing in position *j*, we can show that the probability of a target identification when no position effects are present is equal to:

$$P_{TID}(k) = \sum_{j=i}^{n} p_j P_{TID}(k)_j,$$

where *n* is the lineup size and,

$P_{TID}(k)_j = (1 - G_k)F^{j-1}$.

The probability of a target detection, i.e. any identification at position *j* is given by:

$$P_{TD}(k) = \sum_{j=i}^{n} p_j P_{TD}(k)_j,$$

where,

$$P_{TD}(k)_j = P_{TD}(k)_{i<j} + P_{TD}(k)_{i=j} + P_{TD}(k)_{i>j},$$

and,

$$P_{TD}(k)_{i<j} = (1 - F_k)\sum_{i=1}^{j-1} F^{i-1},$$

$$P_{TD}(k)_{i=j} = (1 - G_k)F^{i-1},$$

$$P_{TD}(k)_{i>j} = (1 - F_k)G\sum_{i=j+1}^{n} F^{i-2}.$$

The probability of a false alarm, i.e. any identification on a TA lineup (with no designated innocent suspect) is:

$$P_{FA}(k)_j = \sum_{i}^{n} F^{i-1},$$

where,

$$P_{FA}(k)_i = (1 - F_k) \sum_{i=1}^{n} F^{i-1}.$$

The probability of a miss, i.e. incorrectly rejecting a target present lineup is:

$$P_M = \sum_{j=1}^{n} P_{M_j},$$

Where,

$$P_{M_j} = p_j G F^{n-1}.$$

The probability of a correct rejection on a target absent lineup is:

$$P_{CR} = F^n.$$

4.4.5.2 The SDT-ISL Model

In the form specified above, the ISL model is not a signal detection model. It provides point estimates of the probabilities for different decision outcomes but makes no assumptions about the form of underlying distributions for signal strength. It can be recast as a signal detection model by treating $1 - G_i$ as the hit rate and $1 - F_i$ and $1 - F_i^*$ as pre-target and post-target false alarm rates, respectively, for each identification position *i*, and confidence level *k*. If the underlying signal strength distributions are assumed to be Gaussian then the model can be parameterised in the following way:

$$1 - F_{ik} = 1 - \Phi(c_{ik})$$

$$1 - F^*{}_{ik} = 1 - \Phi(c^*_{ik})$$

$$1 - G_{ik} = 1 - \Phi\left(\frac{c_{ik} - d_i}{s_i}\right),$$

Where c_{ik} is a decision criterion at identification position *i* and confidence level *k*, d_i is the mean of the target distribution at identification position *i*, s_i is the standard deviation of

the target distribution at identification position i, and Φ is the normal cumulative distribution function. In this version of the model, the s_i parameter is not identifiable. We fixed this parameter to 0.8 across all positions in our analyses to ensure that our d_i values for each identification position were interpretable.

4.4.6 Aims and Hypotheses

The aim of this paper is to investigate the effects of target and identification position on the probabilities of target and foil identification rates using the ISL model, and on underlying discriminability (d_i) and response bias (c_i and c_i^*) using the SDT-ISL model. Building on Wilson et al. (2019), we conducted a stopping rule sequential lineup experiment, in contrast to their experiments in which there was no stopping rule and from which participants were able to make multiple identifications. We fit the ISL and SDT-ISL models to our data and data from Wilson et al. (2019).

Hypotheses were pre-registered on the Open Science Framework, and are available at https://osf.io/3v2p6. Based on the predictions of diagnostic feature detection theory, we hypothesise that underlying discriminability will increase over identification position. Based on suggestions from previous research (Carlson et al., 2016; Carlson et al., 2008; Clark & Davey, 2005; Horry et al., 2012b), we also hypothesise that response bias will become more lenient with identification position.

4.5 Experiment

This experiment presented a stopping-rule sequential lineup task where the position of the target on target present (TP) trials was randomised. The main outcome variables for each participant were the identification decision, the position serial position of the identification (if one was made) and post-decision confidence.

4.5.1 Design

The ISL model approach required us to conceptualise this experiment as containing seven between-participants conditions, corresponding to the position of the target in the lineup. The target appeared in either position one, two, three, four, five or six on a target present lineup, or was not present on a target absent lineup.

As there are six target present conditions and one target absent condition, it is necessary to adjust the ratio of target present to target absent lineups accordingly. A 1:1 ratio of target present to target absent trials would require a prohibitively large sample size to ensure adequate cell counts in the six target present conditions and would also lead to more responses than necessary in the target absent condition. As a result, we collected approximately four times as many target present responses as target absent responses.

4.5.2 Participants

Participants were n = 7525 Amazon Mechanical Turk workers who were compensated $1 USD for the five- to ten-minute experiment. We excluded 321 participants for failing the attention check questions relating to the content of the stimulus video, leaving n = 7204 (TP = 5784, TA = 1420) participants for the analysis. This dataset comprises data from two separate collection batches of the same experiment. In the first batch (n = 1846; TP = 952, TA = 894), collected in May 2019, we used a 1:1 ratio of target present to target absent trials. These data were not analysed using the ISL model. We subsequently came to appreciate that more power was required to ensure adequate cell counts for fitting the ISL model and conducted a second batch of data collection (n = 5358; TP = 4832, TA = 526) in January 2020 with a 9:1 ratio of target present to target absent trials.

4.5.3 Materials

This study used the stimuli from Kaesler et al. (2020); a pool of sixteen headshots of females and accompanying videos that were selected from a larger corpus of ninety

candidates using pairwise similarity ratings and feature-based exclusion criteria. See Kaesler et al. (2020) for a full description of the stimulus selection process. All stimuli acted as both targets and foils. This pool approach was intended to mitigate the possibility of stimulus effects that may result from using a single target and set of accompanying foils.

4.5.4 Procedure

The procedure was largely the same as for Kaesler et al. (2020) but with a sequential lineup condition only. The entire procedure took place within Amazon Mechanical Turk (AMT), within the participants' web browsers. Participants were allocated to either a target present or target absent lineup on a round robin basis. After seeing instructions and providing their consent to participate, participants had to correctly answer comprehension questions, or view the instructions again, before proceeding to the experiment. They were then shown a video of a target randomly selected from the sixteen-member pool, before completing a visual search distractor task. Participants were shown pre-lineup instructions corresponding to those in the U.S. National Department of Justice (1999) guidelines, including instructions that the procedure would terminate if they made an identification. They were then shown a sequentially presented target present or target absent lineup, with the target (if any) and the appropriate number of foils (five for target present, six for target absent) randomly selected from the remaining fifteen members of the stimulus pool. The position of the target (if any) and foils was random.

Participants saw each lineup item individually with an option either to identify or to reject it. If an item was rejected, the next item was shown. If an item was identified, the procedure terminated and the participant provided a typed confidence estimate for their identification on a 0 – 100 scale and a written confidence estimate in words. If all lineup items were rejected, participants were informed that the lineup was exhausted, indicating a rejection decision, and the provided a typed confidence rating on a 0 – 100 scale and a written

confidence estimate in words. Participants also provided a written justification for their identification or rejection decision. They then answered follow-up questions about instruction clarity and task difficulty, and were given the opportunity to provide feedback.

4.5.5 Analyses

We fit models using the fmincon function in Matlab® and quantified model fit using a χ^2 test that compared the predicted data generated by the most optimal parameter estimates to the observed data.

4.5.5.1 Bootstrapping Procedure

We employed a bootstrapping procedure for selected ISL and SDT-ISL model analyses. On each bootstrap replication, we first generated a bootstrap dataset by resampling with replacement by row from the observed data. We then fit the model to the bootstrap dataset. We bootstrapped the predicted data obtained from this first fit by generating random vectors from a multinomial for each subset of trials where the target appeared in position *j*. In practice, this involves converting each column of the observed data structure shown in Table 4.1 to proportions by dividing the cell counts by the total number of observations in each column. Then, these probabilities and the total *n* in each column can be used to draw a random sample from a multinomial distribution. We then fit the model to this second bootstrap dataset and recorded the model fit statistics and parameter values.

4.5.5.2 Likelihood Ratio Tests

We employed likelihood ratio tests to test for differences in parameters. We compared the χ^2 value of an unconstrained model to the χ^2 value of a model with one or more parameters constrained to be equal to one another. Adding a constraint adds a degree of freedom to the model. The difference between the χ^2 values for each model is distributed as χ^2, with the degrees of freedom given by the number of constraints.

4.5.6 Results and Discussion

Table 4.2 shows the observed data in the format used by the ISL model. The data at each position are collapsed over confidence, i.e. there are $k = 1$ confidence categories. The columns correspond to the subset of trials where the target was present at position *j*, with the rightmost column corresponding to the subset of trials on which the target was not present. The rows correspond to the subset of trials where an identification was made at position *i*, with the bottom row corresponding to trials on which the lineup was rejected.

Table 4.2

Observed Frequencies for each Target and Identification Position in Experiment One

ID Position	Target Position						
	1	2	3	4	5	6	None
1	701	119	104	122	90	106	185
2	8	626	68	75	68	62	94
3	7	10	631	72	47	91	99
4	11	10	12	532	63	63	86
5	17	12	6	6	485	53	87
6	15	12	11	9	10	421	79
None	224	185	175	154	148	143	790

4.5.6.1 Model Fit Performance

Recall that G_i is the probability of not identifying a target at position $i = j$, F_i is the probability of not identifying a foil in position *i* where $i < j$ and F_i^* is the probability of not identifying a foil in position *i* where $i > j$. We estimated *G* and *F* for each identification position and *F** for positions two to six, making 17 model parameters in all. The ISL model fit the data well, $\chi^2(25) = 31.46$, $p = .17$.

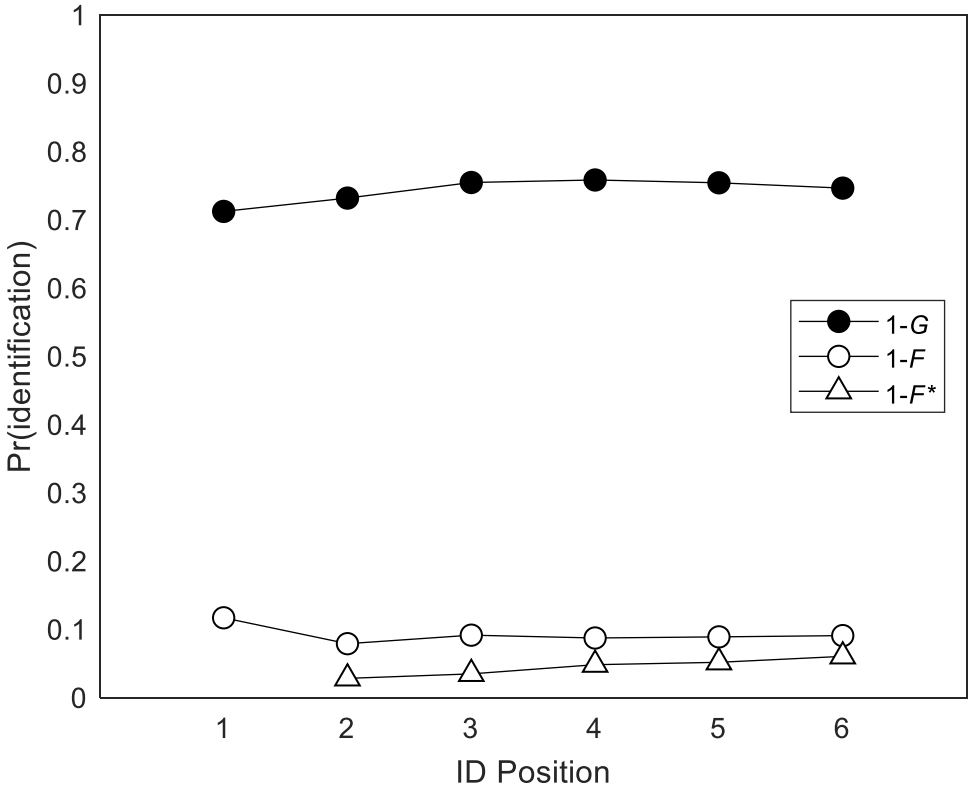
4.5.6.2 ISL Model Parameters

Figure 4.3 plots identification rates, $1 - G_i$, $1 - F_i$ and $1 - F_i^*$, by identification position *i*. Visual examination of Figure 4.3 reveals a robust target effect; the probability of

identifying the target is greater than the probability of identifying a foil at all identification positions.

Figure 4.3

ISL Model Parameters by Identification Position for Experiment One



It appears that the probability of a target identification ($1 - G_i$) changes little with identification position, peaking slightly in the middle positions. A likelihood ratio test indicated that there were no differences in G_i by position, $\chi^2(5) = 7.18$, $p = .21$. The probability of identifying a foil prior to the presentation of a target ($1 - F_i$) appears to decrease with identification position, particularly from position one to position two, as indicated by a significant likelihood ratio test, $\chi^2(5) = 50.54$, $p < .001$. A model with equality constraints for F_2 to F_6 was not significant, $\chi^2(4) = 4.72$, $p = .32$, indicating that the decrease is from position one to position two. The probability of identifying a foil after the presentation of a target ($1 - F_i^*$) appears to increase with identification position but the

likelihood ratio test was not significant, $\chi^2(4) = 7.66$, $p = .10$. It also appears that the probability of identifying a foil is greater prior to the presentation of the target than after the presentation of the target at all positions, although this effect weakens with identification position. A likelihood ratio test constraining F_i and F_i^* to be equal from positions two to six indicated that this difference was significant, $\chi^2(5) = 56.72$, $p < .001$. This demonstrates the necessity of distinguishing between pre and post-target foil identification rates, as the model fits poorly when the two are equated.

We now fit SDT-ISL to the data in order to investigate how underlying discriminability and pre- and post-target response bias differ by identification position.

4.5.6.3 SDT-ISL Model Parameters

One limitation of SDT-ISL is that the s_i parameter is unidentifiable, so fitting the unconstrained SDT-ISL model to the data produces d_i and s_i estimates that are not interpretable. Consequently, we employed a two-step procedure to test for differences in underlying discriminability by position. We first examined a plot of the ISL model parameters shown in Figure 4.3 projected in to zROC space for evidence of any differences in underlying discriminability by position. We then estimated d_i at each position by fitting a constrained SDT implementation of the ISL model with s_i fixed to be equal across all identification positions.

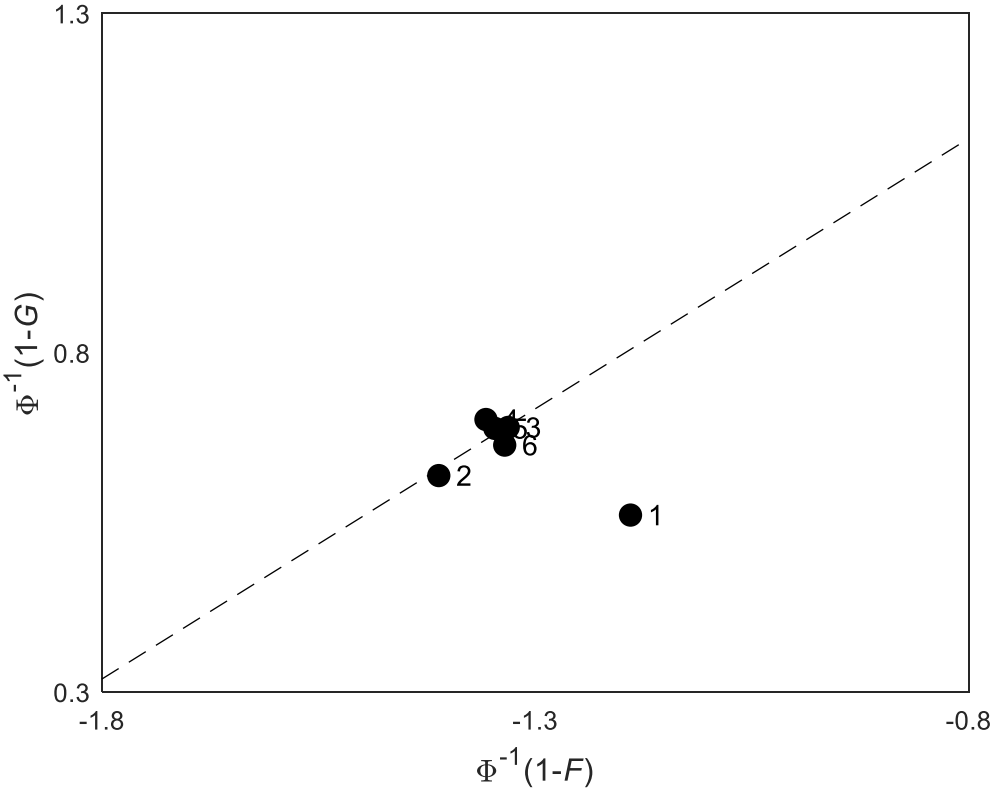
4.5.6.4 Examining the zROC Plot

We transformed the ISL parameters in to zROC space by calculating $\Phi^{-1}(1 - G_i)$ and $\Phi^{-1}(1 - F_i)$ at each identification position *i* where Φ^{-1} is the inverse normal cumulative distribution function. In zROC space, any point lying on a straight line with a slope of $1/s$ and an intercept of d/s represents a constant level of underlying discriminability, where d and s are the mean and standard deviation of the target distribution, respectively (Macmillan & Creelman, 2005). Thus, if $d_i = d$ and $s_i = s$, i.e. there is no effect of identification

position on underlying discriminability, the points of $\Phi^{-1}(1 - G_i)$ plotted against $\Phi^{-1}(1 - F_i)$ for all i will fall on a straight line. Figure 4.4 shows $\Phi^{-1}(1 - G_i)$ plotted against $\Phi^{-1}(1 - F_i)$ for each identification position i.

Figure 4.4

zROC plot for Experiment One



In zROC space, performance is best at the top left corner and responding is most conservative at the bottom left corner. It is evident that the position one point is not captured by the dashed line of best fit through positions two to six. From this line, we estimated an aggregate d_t = 2.20 and s_t = 1.26 for positions two to six.

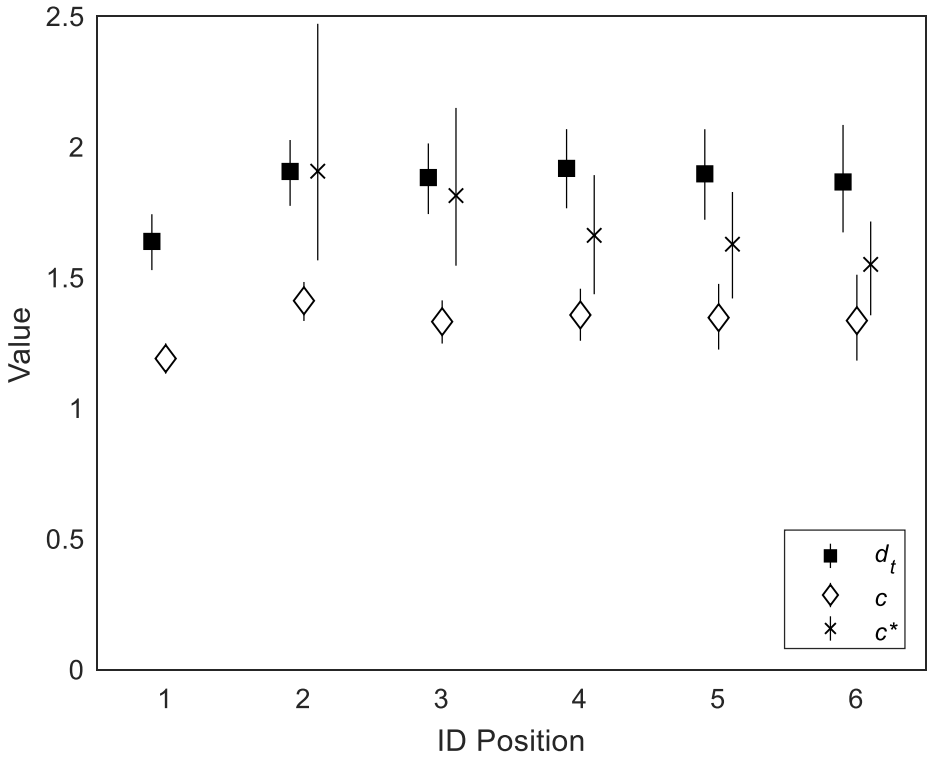
The position one point lies further away from the top left corner of Figure 4.4 than the remaining positions. This indicates that discriminability may be lower at position one compared to position two to six.

4.5.6.5 Likelihood Ratio Tests for SDT-ISL

To make SDT-ISL identifiable, we fixed s_i to 0.8 across all positions, as this is a plausible value for aggregate data based on previous research that has fit signal detection models to estimate s (e.g. Wilson et al., 2019; Wixted et al., 2018). This also ensures that differences in d_i between identification positions are interpretable as differences in underlying discriminability. As SDT-ISL is a reparameterised version of the ISL model, it produces an identical fit to the data, $\chi^2(25) = 31.46$, $p = .17$. Figure 4.5 shows the parameter estimates for d_i, c_i and c_i^* at each identification position, along with 95% bootstrap confidence intervals (CIs) around each parameter, generated from 10,000 bootstrap resamples using the procedure described earlier.

Figure 4.5

SDT-ISL Parameters for Experiment One



We excluded any outliers in the bootstrap distributions before calculating the 95% CIs. Some of these outliers were likely caused by the model finding local minima. In the case of c_2^* and c_3^*, there are likely ceiling effects due to the relatively small number of post-target foil identifications when the target appears in early positions. That is, F_2^* and/or F_3^* equal 1 on some bootstrap resamples. It is evident from the bootstrap 95% CIs that the lack of power in these cells means that the c_i^* estimates are more variable than the d_i and c_i estimates.

On examining Figure 4.5, it appears that d_i increases from position one to position two, but changes little over subsequent positions. A likelihood ratio test indicated that this difference was significant, $\chi^2(5) = 31.14$, $p < .01$, while a likelihood ratio test constraining d_i from positions two to six was not significant, $\chi^2(4) = .42$, $p = .98$. This partially supports our hypothesis that underlying discriminability would increase with position. The difference between d_t at identification position one and identification position two is 1.64 – 1.91 = -.27. If we interpret this in a similar way to Cohen's *d*, the effect size of the differences is small according to the interpretation suggested by Cohen (1988).

As the SDT implementation of the ISL model specifies that $F = \Phi(c_i)$, it follows that $c_i = \Phi^{-1}(F_i)$ and $c_i^* = \Phi^{-1}(F_i^*)$. That is, the c_i and c_i^* values shown for each identification position in Figure 4.5 are the F_i and F_i^* values for each identification position from Figure 4.1, expressed on a different scale. Accordingly, the results of the likelihood ratio tests are the same as those reported for the ISL model parameters. They indicate that, while responding after the presentation of the target (c^*) appears to decrease sharply from positions two to six, the effect did not reach significance, $\chi^2(4) = 7.66$, $p = .10$, likely due to the uncertainty of the c^* estimates. There was no significant difference in responding prior to the presentation of the target (c) from positions two to six, $\chi^2(4) = 7.66$, $p = .10$, but there was a difference in c across all positions, $\chi^2(5) = 50.54$, $p < .001$. This indicates that responding prior to the presentation of the target became more conservative from position one to position two but did

not change beyond position two. Responding was also more conservative after the presentation of the target than prior to the presentation of the target at positions two through to six, $\chi^2(5) = 56.72$, $p < .001$. As for results for F_i and F_i^*, this demonstrates the importance of separately modelling pre- and post-target identification rates.

4.6 Reanalysis of Wilson et al. (2019)

We now employ ISL and SDT-ISL models to reanalyse the data from both experiments in Wilson et al. (2019).

4.6.1 Data Considerations and Model Fit Performance

As Wilson et al. (2019) employed a 1:1 ratio of target present to target absent lineups, there were fewer target present responses in both experiment one (n = 6530; TP = 3258, TA = 3272) and the lineup trials of experiment two (n = 1966; TP = 999, TA = 967) than in our experiment (n = 7204; TP = 5784, TA = 1420). Recall that Wilson et al.'s (2019) participants provided an identification decision and a confidence rating for all items. Following Wilson et al. (2019), we imposed a stopping rule on participant responses in order to compare the data from their task to the stopping rule sequential lineup. We counted the first binary identification response from each participant, as this is the decision that would have been recorded for each participant if the lineup were conducted with a stopping rule. For example, if a participant made identifications at position two and another at position five, we counted the identification at position two. Table 4.3 and Table 4.4 show the data from Wilson et al. (2019) experiment one and the lineup data from experiment two, respectively, in the ISL model format.

The ISL model fit the experiment one data well, $\chi^2(25) = 31.46$, $p = .17$. For experiment two, there are a large number of cells with counts less than five, compromising the χ^2 test used to assess model fit. As a result, we employed the bootstrapping procedure described earlier to assess model fit for experiment two. We first generated a bootstrap

distribution of χ^2 values from 10,000 bootstrap resamples. We calculated a *p*-value for the fit of the ISL model to the experiment two data by calculating the proportion of the bootstrap distribution of χ^2 values that was greater than or equal to the observed χ^2 value. The ISL fit the experiment two data well according to the bootstrap fit test, $\chi^2(25) = 27.90$, $p = .25$.

Table 4.3

Observed Data for Wilson et al. (2019) Experiment One

ID Position	Target Position						
	1	2	3	4	5	6	None
1	448	103	103	104	107	98	602
2	2	350	77	61	73	69	414
3	6	7	296	77	83	76	439
4	12	13	9	255	45	48	375
5	16	5	3	2	183	41	239
6	7	4	5	6	3	160	170
None	54	59	43	38	53	54	1033

Table 4.4

Observed Data for Wilson et al. (2019) Experiment Two

ID Position	Target Position						
	1	2	3	4	5	6	None
1	135	34	24	19	25	27	164
2	0	111	21	21	18	24	125
3	3	0	100	31	26	13	115
4	2	2	0	85	22	14	113
5	5	2	2	0	55	12	66
6	2	2	4	2	2	62	62
None	23	17	12	13	12	15	322

4.6.2 Estimating Signal Detection Parameters at each Identification Position

For each experiment, we employed the same procedure as for our data. We first examined a zROC plot of $\Phi^{-1}(1-G_i)$ plotted against $\Phi^{-1}(1-F_i)$ at all *i* for evidence of differences in underlying discriminability by position. We then conducted likelihood ratio tests using the SDT implementation of the ISL model with s_t fixed to 0.8 across all identification positions.

4.6.2.1 Examining the zROC

Figure 4.6 and Figure 4.7 show $\Phi^{-1}(1-G_i)$ plotted against $\Phi^{-1}(1-F_i)$ at each identification position and a line of best fit for Wilson et al. (2019) experiment one and two, respectively. For both experiments, the points for each identification position seem to deviate minimally from the line of best fit. This indicates that discriminability may not differ by identification position in either experiment.

Figure 4.6

zROC Plot for Wilson et al. (2019) Experiment One

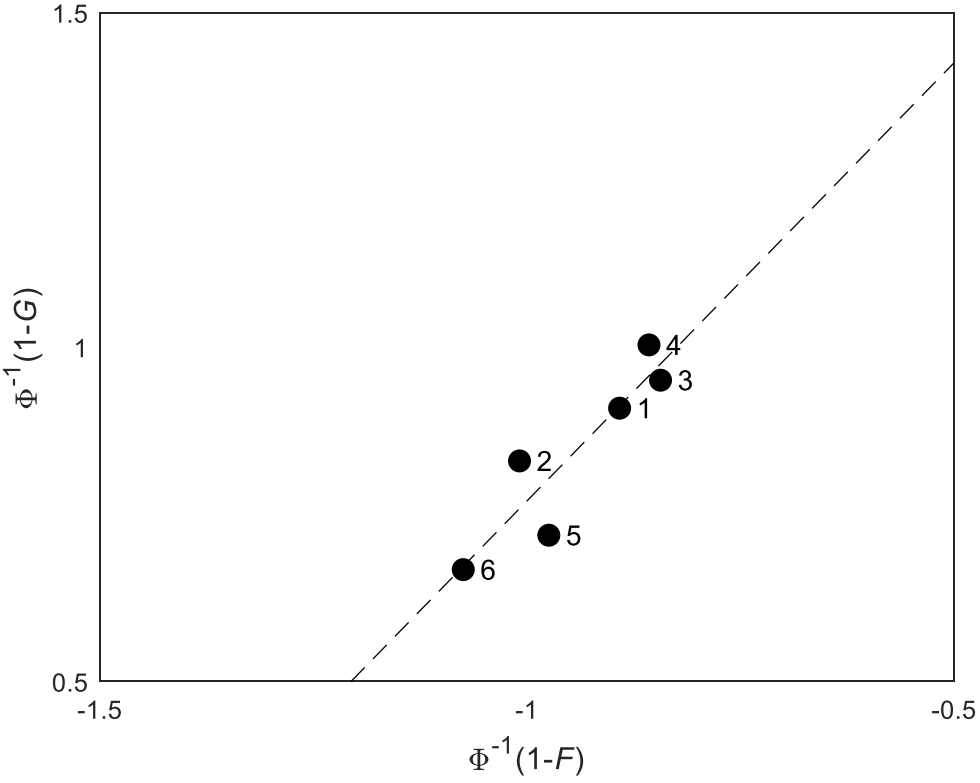
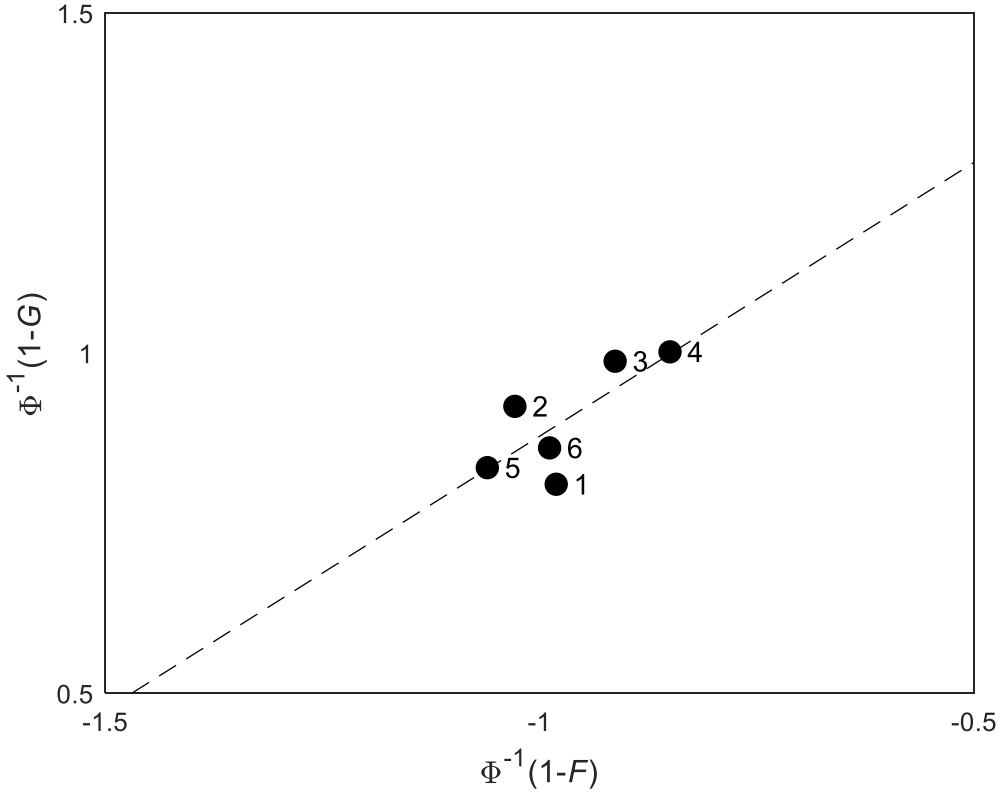


Figure 4.7

zROC Plot for Wilson et al. (2019) Experiment Two



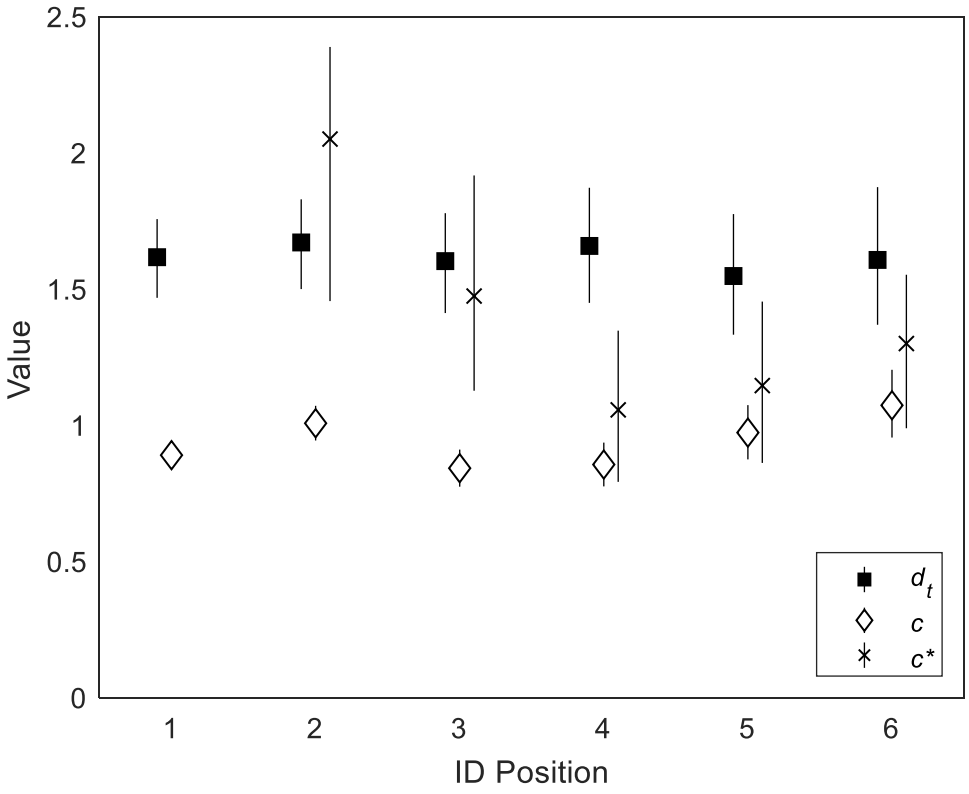
4.6.2.2 SDT-ISL Likelihood Ratio Tests

We fit the SDT implementation of the ISL model to the data from experiment one and experiment two with s_t fixed to .8 at all identification positions to ensure that d_t estimates are comparable between the two experiments and to our data. Figure 4.8 and Figure 4.9 show values of d_i, c_i and c_i^* and bootstrap 95% CIs by identification position for experiment one and experiment two, respectively. As for our previous analysis, we removed outliers in the bootstrap distributions before calculating the 95% CIs. Once again, there were relatively few foil IDs after the presentation of the target, leading to greater uncertainty in the c_i^* estimates than for the other parameters.

Likelihood ratio tests indicate that d_i does not differ by identification for experiment one, $\chi^2(5) = 1.90$, $p = .86$, or experiment two, $\chi^2(5) = 1.19$, $p = .95$. This does not support our hypothesis that discriminability would increase with identification position. It also does not concur with the increase in underlying discriminability from position one to position two reported by Wilson et al. (2019) for experiment one.

Figure 4.8

SDT-ISL Parameters for Wilson et al. (2019) Experiment One

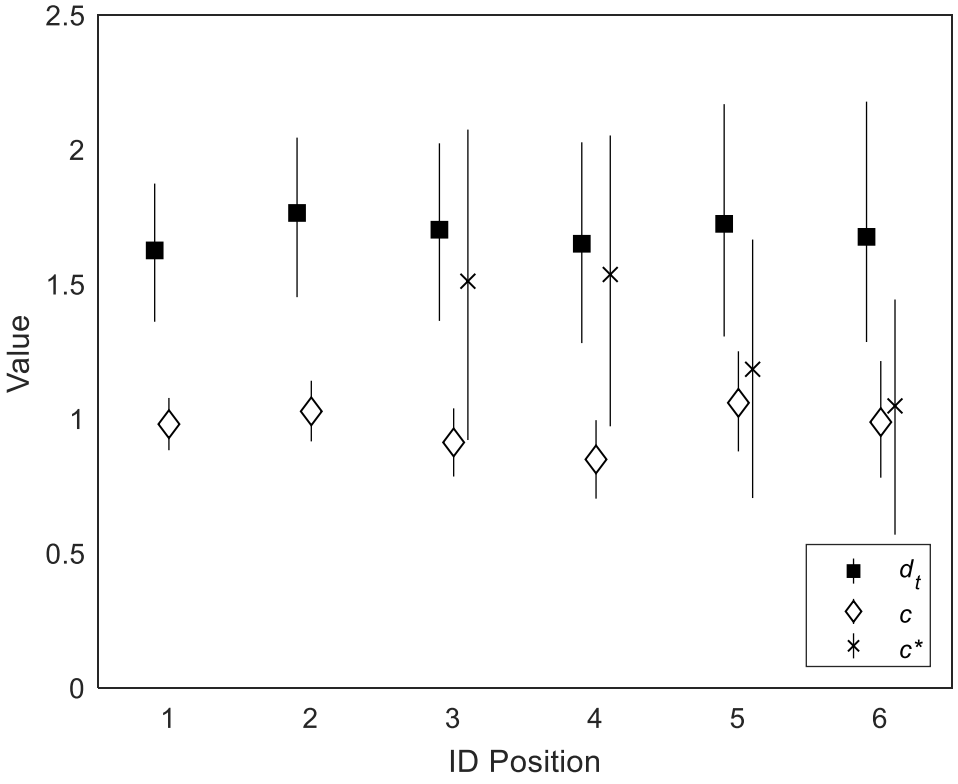


For experiment one, c_i significantly differed by position according to a likelihood ratio test, $\chi^2(5) = 45.40$, $p < .001$, decreasing from positions one to four, then increasing from positions four to six. Parameter c_i^* also significantly differed by position, $\chi^2(4) = 15.61$, $p < .01$, decreasing from position two to four, then increasing from position four to six. We conducted likelihood ratio tests comparing c_i and c_i^* at each identification position, with a Bonferroni corrected alpha level of .01 (.05/5). The difference between c_i and c_i^* was

significant at position two, $\chi^2(1) = 13.59$, $p < .001$ and position three, $\chi^2(1) = 18.67$, $p < .001$, but was not significant at position four, $\chi^2(1) = 3.51$, $p = .06$, five, $\chi^2(1) = 2.42$, $p = .12$, or position six, $\chi^2(1) = 3.82$, $p = .051$. This is similar to the pattern of results in our data, responding after the presentation of a missed target is more conservative than prior to the presentation of the target. In this case, the effect is short lived, only surviving for positions two and three.

Figure 4.9

SDT-ISL Parameters for Wilson et al. (2019) Experiment Two



For experiment two, likelihood ratio tests revealed no significant difference in c_i by position, $\chi^2(5) = 11.15$, $p = .048$, or c_i^* by position, $\chi^2(3) = 4.24$, $p = .24$. Note that Figure 4.9 does not show c_2^* because there were no foil identifications at position two when the target was in position one. As for experiment one, we compared c_i and c_i^* at each identification position using likelihood ratio tests, with a Bonferroni corrected alpha level of .0125 (.05/5).

The difference between c_i and c_i^* was significant at position four, $\chi^2(1) = 8.09$, $p < .01$, but was not significant at position three, $\chi^2(1) = 5.05$, $p < .05$, position five, $\chi^2(1) = .43$, $p = .51$, or position six, $\chi^2(1) = .11$, $p = .74$. As per our data and Wilson et al. (2019) experiment one, there is some evidence here of more conservative responding after missing the target at early positions in the lineup.

The pattern of parameter estimates for Wilson et al. (2019) experiment two shown in Figure 4.9 is similar to that for our data and Wilson et al. (2019) experiment one. It is possible that experiment two may not have had sufficient power to detect any differences in parameters by position, as indicated by the relatively wide 95% CIs.

4.7 General Discussion

The aim of this study was to investigate how underlying discriminability and response bias change over the course of the sequential lineup task (Lindsay & Wells, 1985). In order to achieve this, we developed a modelling framework that accounts for the effects of both the position of the target in the lineup and the position at which an identification decision is made. The Independent Sequential Lineup (ISL) model measures target identification rates and pre- and post-target foil identification at each lineup position, while its signal detection parameterisation, SDT-ISL, measures underlying discriminability and pre- and post-target response bias by position. Distinguishing between the probability of pre- and post-target foil identifications was critical to the models' ability to fit the data.

We fit the models to data from our sequential lineup experiment, which was conducted with a stopping rule, and to data from two experiments reported in Wilson et al. (2019), which were conducted without a stopping rule. We found a small increase in underlying discriminability from identification position one to position two in our experiment, but underlying discriminability did not increase beyond identification position two. For the Wilson et al. experiments, we did not find any differences in underlying

discriminability by identification position. This is contrary to the results reported for experiment one of Wilson et al., based on fits of the SDT-SEQ model to subsets of the data where a target appeared in a particular position. As previously discussed, the results of SDT-SEQ may be misleading in this case because the probability of a foil identification is assumed not to vary by serial position. It is interesting to note that an important result reported by Wilson et al. – an increase in underlying discriminability from position one to position two – was not evident in our reanalysis of their data, but was observed in our own experimental data.

4.7.1 Underlying Discriminability and the Diagnostic Feature Detection Hypothesis

An increase in underlying discriminability from position one to position two is consistent with the prediction of diagnostic feature detection theory (DFDT). The presentation of each lineup item offers an additional opportunity to isolate features uniquely shared by the perpetrator and the lineup items. Underlying discriminability may not have increased beyond position two in these data due to the typically high level of similarity between lineup items. Witnesses may be able to discount the features shared between all lineup items after viewing only the first and second items; subsequent provide little extra information. Alternatively, the failure to observe greater discriminability beyond position two may be the result of output interference (Criss et al., 2011) offsetting the beneficial effect of comparing across lineup items. This is similar to finding in basic recognition memory research employing a multiple trial design that underlying discriminability declines over the course of the test sequence (Criss et al., 2011). This interpretation of our data treats the increase in underlying discriminability from position one to position two as a "within-participant" effect. While this interpretation is intuitive, our experiment was conducted with a stopping rule, which means that the observed effect could also be explained as a "between-participant" effect. When the target is in position one, all participants in that condition make

an identification response at position one. When the target is in position two, some subset of participants have already identified a foil at position one, and therefore cannot identify the target at position two. It is possible that underlying discriminability is greater at position two because participants with poor memory or a bias to lenient responding false alarm at position one, which means they are not included in the analysis as subsequent positions. The results of our experiment, and of the Wilson et al. (2019) reanalysis where we imposed a stopping rule, may reflect some combination of within- and between-participant changes in underlying discriminability over the course of the sequential lineup.

4.7.2 Response Bias

The prevailing view in the literature is that the probability of a false alarm increases over the course of the lineup, indicating that responding becomes more lenient (Carlson et al., 2016; Carlson et al., 2008; Clark & Davey, 2005; Horry et al., 2012b; Lindsay & Wells, 1985). Our results clarify this view; pre-target false alarm rates tended to decrease with lineup position, while post-target false alarm rates tended to increase with position. That is, responding prior to the presentation of the target became more conservative as the lineup progressed while responding after the presentation of the target became more lenient as the lineup progressed.

More conservative responding in positions prior to the presentation of the target may reflect a rational adjustment to the familiarity elicited by pre-target foils. If a foil is rejected, then subsequent foils of greater or equal familiarity should also be rejected. This may require participants to adjust their criterion to be more conservative when relatively familiar foils are presented. Like the underlying discriminability results, the stopping rule constraint means that there is also a between-participant explanation for this pattern of results. If each participant is assumed to hold their decision criterion constant throughout the lineup, lenient participants will tend to identify at early lineup positions, while only more conservative

participants will survive to identify at later positions. The apparent conservative criterion shift with position in pre-target responding may therefore be a selection effect.

It seems that the experience of viewing the target, but failing to identify, causes participants to reset their response criteria to a more conservative level, although this effect weakens as the lineup progresses. This is a within-participant effect; having experienced a target that was not identified, the following foil is rarely chosen because a conservative criterion shift occurs. Given that identifications immediately following the target are rare, they may communicate information about the reliability of witnesses at the group level. Witnesses who identify the foil immediately following the target may have particularly poor memory for the perpetrator because experiencing the target as familiar ordinarily leads to shift to a more conservative response criterion for the next item.

Our results imply that the probability of an innocent suspect in a fair lineup being identified from a sequential lineup differs depending on where they are placed in the lineup. The probability of a guilty suspect being identified appears to be unaffected. If the aim is to reduce false identifications, the optimal position for the suspect in a sequential lineup is in the final lineup position. The issue of position effects on innocent suspect identifications should be considered by jurisdictions that currently employ the sequential lineup procedure.

4.7.3 Future Directions

One aspect not explored in this study is how position effects might operate in target absent lineups when an innocent suspect is selected to resemble the perpetrator to a greater degree than the foils. The target absent column of the ISL model data structure shown in Table 4.1 can be expanded to include an innocent suspect. In this case, the data structure for the target present columns is replicated for the position of the innocent suspect and there would be an addition G parameter for the probability of not identifying an innocent suspect at position j. An experiment could manipulate innocent suspect position and the degree of

similarity between the designated innocent suspect and the perpetrator. Our interpretation of DFDT is that it predicts less of an increase in underlying discriminability by position the more strongly an innocent suspect resembles the perpetrator. This is because the perpetrator and innocent suspect will share a number of distinctive features and the innocent suspect should therefore tend to be identified at a higher rate rates regardless of its position in the lineup.

4.7.4 Conclusions

In this study, we developed and introduced the ISL model for investigating serial position effects in the sequential lineup task. We applied the ISL model to three sequential lineup experiments to test a prediction of the diagnostic feature detection hypothesis, that underlying discriminability should increase as the lineup progresses. We found some support for this prediction in our experiment, but there were no position effects on underlying discriminability in a reanalysis of two experiments in Wilson et al. (2019). However, we cannot rule out the possibility that these divergent results are due to differences between our lineup task and the one used in Wilson et al. (2019). Taken together, our results suggest that the beneficial effect of isolating distinctive features proposed by the diagnostic feature detection hypothesis is small and therefore may be counteracted by other features of the sequential lineup task. Responding prior to the presentation of the target became more conservative as the lineup progressed, consistent with the behaviour of a rational observer. Failing to identify a target led to more conservative responding on the next item, although this effect weakened as the lineup progressed. Further work is required to understand how other aspects of the sequential lineup task, such as lineup size or item similarity, may interact with serial position effects on underlying discriminability and response bias

Chapter 5

5.1 Discussion

Lineup tasks on which the items are sequentially presented have a long history in the eyewitness identification field (Lindsay & Wells, 1985). Sequential presentation was originally proposed as a method for reducing the rate of innocent suspect identification, known as false identification, observed on simultaneous lineup. The original procedure (Lindsay & Wells, 1985) is effective in this regard because it tends to make witnesses less willing to identify from the lineup overall (Palmer & Brewer, 2012; Steblay et al., 2011b). Experiments that manipulate presentation format, among other procedural manipulations, have highlighted a lack of theoretical explanation of the mapping between discriminability as expressed via the decision and criterion as expressed via a confidence rating. This thesis is an exploration of the value in applying formal mathematical models to data so that clear conclusions can be drawn regarding the expression of memory strength, referred to throughout as discriminability, and response bias on sequentially presented lineup tasks.

Discriminability and response bias were estimated from lineup decision outcomes and confidence ratings using measurement models based in signal detection theory, which provide a mathematical mapping between these observed decision outcomes and the latent variables of interest. Due to the stopping rule constraint of the sequential lineup procedure described by Lindsay and Wells (1985), it was necessary to develop a novel signal detection model, SDT-SEQ in order to measure discriminability and response bias on this task. This model captures the stopping rule constraint by specifying a “first above criterion” decision rule. The SDT-SEQ model was used, along with candidate models of the simultaneous lineup task, to compare discriminability and response bias on the simultaneous lineup, the sequential stopping rule lineup and the UK lineup. While SDT-SEQ is an appropriate model for measuring overall discriminability and response bias for the sequential stopping rule lineup, it

is ill-suited to investigating how discriminability and response bias change over the course of this procedure. This is because the model accounts only for the position of the target in the lineup, not the position at which an identification is made. As a result, a modelling framework dubbed the Independent Sequential Lineup (ISL) model was developed, which accounts for both the position of the target in the lineup, if any, and the position at which an identification is made and distinguishes between pre- and post-target foil identifications. This is a new insight into this problem and allows for a more nuanced understanding of position effects.

Study one developed the SDT-SEQ model and used it to compare sequential stopping rule presentation and simultaneous presentation in two corpora of previous studies and newly collected experimental data. Overall, simultaneous and sequential stopping rule presentation did not significantly differ in discriminability. While datasets from some recently published studies showed a simultaneous advantage, the effect size for the difference across all datasets was close to zero. Sequential stopping rule presentation was associated with more conservative responding than simultaneous presentation, confirming the findings of previous research. Study two compared simultaneous presentation to UK lineup presentation and sequential stopping rule presentation using a larger sample than the experiment conducted in study one. Results showed that simultaneous presentation was associated with greater discriminability than sequential stopping rule presentation, but not UK lineup presentation, although this result was dependent on which model was fit to the simultaneous lineup data. There was no significant difference in discriminability between UK lineup presentation and sequential stopping rule presentation. Responding was most lenient for UK lineup presentation, followed by simultaneous lineup presentation, followed by sequential stopping rule presentation. Study three developed the ISL modelling framework and applied it to data from a large sample sequential stopping rule lineup experiment and to data from a previous

study that had employed SDT-SEQ to investigate position effects on the same task (Wilson et al., 2019). A large sample is required to investigate position effects because there are six possible target positions, in addition to a lineup with no target, which effectively results in the sample being divided in to seven experimental conditions. Results indicated that discriminability increased from position one to position two in our experiment, but not in the data from Wilson et al. (2019), as originally reported for their first experiment. Responding became more conservative with serial position at the group level, prior to the presentation of the target, and tended to become more lenient after the presentation of the target, although this effect diminished as the lineup progressed.

5.2 Modelling Approach

This thesis increases understanding of how structural aspects of the lineup task affect discriminability and response bias, demonstrating the utility of adopting signal detection measurement models in lineup research. Applying models to the question of lineup presentation format addressed the limitations of previous studies that employed either choosing rate-based analyses that confound discriminability and response bias (e.g. Carlson et al., 2008; Steblay et al., 2011a), or ROC analysis, which cannot disambiguate the relative contribution of discriminability and task structure to observable decision accuracy (e.g. Andersen et al., 2014; Mickes et al., 2012).

Previous studies fit the SDT-INT and SDT-MAX models to data from both simultaneous and sequentially presented lineups on the basis that these models represent the both the detection and identification components of a lineup decision (Carlson et al., 2016; Duncan, 2006; Horry et al., 2015; Horry et al., 2012b; Palmer & Brewer, 2012; Palmer et al., 2010; Smalarz et al., 2019). This demonstrates some understanding in the field of the importance of the match between task and model. This thesis extends this understanding by highlighting the need to consider the match between task structure and the model's decision

rule. As discussed in Section 2.4.5., the decision rules of SDT-MAX and SDT-INT are not compatible with the stopping rule constraint of the sequential lineup. Failure to consider this aspect of the task in the models may lead to inaccurate parameter estimates and incorrect conclusions about the data due to model misspecification. To this end, both SDT-SEQ and the ISL model specify decision rules that can account for the stopping rule constraint of the Lindsay and Wells (1985) sequential lineup. Additionally, it is necessary to consider the match between the research question and the model. SDT-SEQ measures discriminability for the sequential stopping rule lineup overall in order to compare discriminability on this task to other presentation formats, as shown in studies one and two. The ISL model aims to understand how discriminability and response bias change over the course of the sequential stopping rule lineup by using an expanded data structure that accounts for the serial position at which an identification was made in addition to the position of the target. Each model is ill suited for the aim of the other. As discussed in study three, SDT-SEQ is uninformative with respect to position effects because its data structure does not include the position at which an identification was made, a problem shared with other models that were previously employed to investigate position effects (Carlson et al., 2016; Horry et al., 2012b; Wilson et al., 2019). Additionally, subsetting the data by the position of the target when there is no innocent suspect in the target absent condition requires approximations for counting correct rejections that compromise the statistical tests used to assess model fit. The ISL model could be used to estimate overall discriminability and response bias for the sequential stopping rule lineup by assuming that there were no position effects. However, the model is complex relative to SDT-SEQ, which is designed to measure these quantities, and constraining discriminability and response bias to be equal over all identification positions defeats the purpose of employing the ISL model in the first place.

Future lineup research employing measurement models would benefit from careful consideration of the match between task, model and research question. A failure to consider this aspect of model based research may compromise accurate measurement of the latent variables that determine observed decision outcomes. This in turn may lead researchers to draw incorrect inferences about the effects of procedural alternations on latent variables. If policy makers adopt these procedures, they may not perform "as advertised", despite a solid foundation of empirical evidence.

5.2.1 Evaluating the Model Assumptions

The models developed and employed in this thesis estimate discriminability and response bias at the population level with the aim of understanding which lineup procedures are likely to maximise discriminability. Each participant provides a single lineup decision and the data for a given experimental condition is an aggregate of these decisions. This constraint means that the models make a number of assumptions about processes at the individual level of identification. First, it is as if all participants share the same estimate for discriminability and response bias. In reality, discriminability will differ between participants, as will the placement of the decision criteria. It is possible that this participant-level variability in criterion placement, dubbed "criterion noise", decreases parameter estimation accuracy (Benjamin et al., 2009; Smith et al., 2016; Wixted et al., 2018). However, adding model parameters to account for criterion noise increases degrees of freedom and model complexity. Criterion noise is modelled by assuming that each criterion is a random draw from a probability distribution, with the mean and standard deviation estimated as free parameters. Deciding whether to add criterion noise, or any additional parameters, to models involves trading off an undesirable increase in complexity with a desirable increase in predictive ability. In the case of criterion noise, there is evidence that the additional model complexity does not result in a suitably large increase in predictive ability (Kellen et al., 2012). The

models also assume that the shape and form of the underlying signal strength distributions are Gaussian at the group level. This is a standard theoretical assumption in signal detection models, but it may be violated if the shape and form of the signal strength distributions vary between participants. In this case, signal detection models that assume mixture distributions for signal strength may provide a path forward (DeCarlo, 2010). A recently proposed information-theoretic version of signal detection theory that need not assume underlying Gaussian distributions may also help overcome this limitation (Feldman, 2021).

Despite these sources of unmodelled variability, the models accounted for the data reasonably well across three studies. This is gratifying, as it provides some evidence that the models' assumptions are a suitable approximation to reality for the purpose of making comparisons between procedures at the task level. However, unmodelled variability in the parameters at the participant and item level may nonetheless lead to bias in the task level estimates of discriminability and response bias (Rouder & Lu, 2005). Investigating this possibility by estimating discriminability and response bias at the participant (and item) level will require changes to both the experimental paradigm and the modelling approach. For the experimental aspect, multiple lineup decisions must be elicited from each participant (Mansour et al., 2017). It is not possible to model at the individual level when collecting one lineup decision per participant. From a modelling aspect, a hierarchical signal detection modelling approach that estimates parameters at the item, participant and group levels must be adopted (Rouder & Lu, 2005). Implementing this within a Bayesian framework might also provide additional information about the precision of the models' estimates by allowing the specification of posterior distributions around the parameter values rather than the point value estimates generated using a frequentist maximum-likelihood approach (DeCarlo, 2012). The bootstrapping procedures employed in studies two and three are a frequentist method for gaining some sense of the precision of the point estimate parameter values.

5.2.2 Implementation Issues

One issue seldom discussed in the literature is the technical challenge of developing and implementing signal detection models of the lineup task. First, it necessary to derive likelihood functions that give the probability of each decision outcome for a particular decision rule. As shown in Wixted et al. (2018) when deriving likelihood functions for the Ensemble model, this sometimes proves difficult for even experienced mathematical psychologists. The likelihood functions must then be implemented in a programming language that performs constrained optimisation. It is possible that these technical demands may act as barriers to entry for some lineup researchers who wish to employ signal detection models, particularly in the absence of collaborators with the necessary mathematical and programming skills. One aim of this thesis was to provide guidance to researchers seeking to employ modelling. To this end, an example implementation of SDT-MAX in the R language is included in Appendix D. The structure of the code in Appendix D could be used for other lineup models by replacing the likelihood functions for SDT-MAX with functions for a different model. Additionally, Cohen et al. (2020) recently developed an R package called `sdtlu` that estimates parameter values for models mathematically identical to SDT-MAX and SDT-SEQ. The increasing accessibility of lineup models has the potential to benefit the field by encouraging more model-based, theoretically motivated lineup studies. However, developing packages for fitting models may also lead to unintended consequences. In particular, inclusion of a model in an easy to use statistical software package may lead to it being favoured by researchers, despite its relatively poor performance. The widespread use of the poor performing SDT-INT model (Horry et al., 2015; Horry et al., 2012b; Palmer et al., 2010, 2012; Smith et al., 2018) may be one such case, as this seems to be at least partially driven by the fact that an implementation developed by Palmer and Brewer (2010) is accessible on request from the authors. Additionally, accessible software packages may

impede understanding of the inner workings of a model, leading to a failure to consider carefully the critical issue of the match between task, model and research question discussed in Section 5.2.

5.2.3 The Applied Benefit of Models

A primary aim of the modelling approach employed in this thesis is to contribute to the development of theories that advance understanding of eyewitness memory and memory in general. These theories may then be used to develop future lineup procedures and advise on existing procedures. However, models may be useful for procedural development even in the absence of theory, because they measure important underlying psychological variables. If some procedural option for presenting a lineup reliably impairs the expression of discriminability, then this option can be *a priori* eliminated when designing new lineup procedures, even if the theoretical explanation for how and why the effect occurs is unresolved. This is not true of ROC analysis, despite its utility for addressing certain applied questions. If two procedures differ in structure and one produces a reliably greater AUC, it is not clear whether this result is due to a difference in discriminability or a difference in task structure. ROC analysis is always tied to the task that produced the ratings – the shape of the ROC is an expression of discriminability via the constraints of the task. If one procedure produces greater AUC it should be preferred, but this does not provide any insight in to how a procedure achieves greater performance. Thus, the results of ROC analysis cannot provide guidance for the development of new lineup procedures. This also presents a problem when comparing two procedures because the model based discriminability and AUC can be different for the same task. This can be resolved with recourse to a measurement model.

5.2.4 Statistical Power and Model Fit

As is common in model based research, the models in this thesis are partially evaluated on the extent to which they fit the data. The limitations of this approach are well

known; a model may provide a good fit to a dataset despite not capturing the true underlying data generating process and a model may fail to fit data even though its core assumptions are a reasonable description of reality (McClelland, 2009; Navarro et al., 2004; Pitt & Myung, 2002). In some situations, a relatively inflexible model that fails to fit many datasets may be preferred. A lack of flexibility may indicate that a model can only predict a small number of outcomes, i.e. its predictions are constrained, which means it may be easier to devise critical tests that can falsify the model. An additional limitation of frequentist methods for quantifying model fit is that they are necessarily affected by statistical power. The null hypothesis for goodness-of-fit is that the data predicted by the model does not deviate from the observed data. This hypothesis, and by extension the model itself, is rejected if the *p* value for the test that quantifies the deviation between predicted and observed data falls below the chosen alpha level. When cell counts in the observed data are low, there may be insufficient power to reject even an unsuitable model at a given alpha level that does not capture the true data generating process. Given sufficiently high cell counts in the observed data, even a suitable model will always be rejected at the same alpha level, because small absolute deviations between the predicted and observed data will nonetheless result in an extreme goodness of fit value under the null. Two studies differing in statistical power might therefore draw different conclusions about a model based on its fit to data. A low power study may fail to reject a model and conclude that it may be used in further studies, while a high power study may reject the same model and conclude that it should not be used in further studies. This means that some flexibility is required when evaluating models in light of the results of frequentist goodness-of-fit tests and that it may be desirable to break with convention and adjust the alpha level when fitting models to data with high power.

This thesis primarily employed the models as measurement instruments and tested for differences in model parameters between experimental conditions, rather than evaluating the

models themselves. When testing for differences in parameters between conditions, it is desirable to have high power, because this increases the possibility of detecting potentially small effects. It is possible that the high power in study two contributed to the poor fit of SDT-MAX and the Ensemble model to the same simultaneous lineup data. Absolute deviations between predicted and observed data were small; the models captured the general pattern of the data despite a statistically significant misfit (see Section 3.6.2).

5.2.5 Selecting between Lineup Models based on Decision Rules

One issue highlighted by this thesis is the need to consider methods for selecting between competing models for a given lineup task. The models in this thesis, excluding the ISL model, have the same number of parameters when fit to a given dataset, differing only in their decision rule for comparing the familiarity values elicited by the lineup items to the decision criteria. This rules out the use of statistics such Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC), which balance goodness-of-fit and the number of free parameters (Kuha, 2004). Additionally, the limited heuristics of selecting between the models based on goodness-of-fit cannot be applied if models perform similarly to each other. A potential path forward may be conducting cross-fitting simulations, similar to those employed in study one to distinguish SDT-SEQ from models of the simultaneous lineup (see Navarro et al., 2004; Schultheis & Singhaniya, 2015; Wagenmakers et al., 2004). These techniques involve simulating data according to competing models and then bootstrapping and cross-fitting each model to the data from the other model. These procedures can be used to explore the effect of differences in functional form between the models and to quantify to model flexibility, the extent to which a model is able to fit data generated by another model. A highly flexible model is not desirable because it indicates that the model can predict a large set of possible outcomes, which may make it difficult to falsify.

In effect, selecting between unequal variance signal detection models of a given lineup task is selecting between different decision rules for characterising the data generating process at the group level. However, it is important to note that the decision rules of lineup models are not process-level explanations for how individual witnesses complete a lineup task. They are characterisations of the expression of discriminability via a lineup task for a population of eyewitnesses. Much like the issue of participant level variance in discriminability and response bias discussed in Section 5.2.1, it is likely that participants employ a range of decision rules, which can be investigated only with changes to the experimental paradigm and modelling approach. The aim of selecting between competing lineup models is to find a decision rule that allows accurate measurement of the underlying latent variables at the level of a population of eyewitnesses.

5.3 Sequential Item Presentation

5.3.1 Discriminability

The main aim of the modelling work conducted in this thesis was to clarify the effects of sequential item presentation on discriminability. The results of studies one and two imply that sequential item presentation may impair discriminability to a small extent, although the evidence is far from conclusive. Sequential item presentation appears to interact with other procedural factors to facilitate the expression of discriminability for a given lineup task; the main effect is relatively weak. On this basis, the field should not prematurely conclude that sequential item presentation impairs discriminability relative to simultaneous presentation, a view often expressed in the literature (Gronlund et al., 2014; Mickes et al., 2012; Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019; Wixted & Mickes, 2014, 2018). Study three found some evidence of an increase in discriminability from serial position one to position two in the sequential stopping rule lineup, with this increase being driven by a reduction in the pre-target false identification rate. As the ISL model is novel and was able to

be applied only to the results of one other recent study (Wilson et al., 2019), it is difficult to integrate this result with earlier literature on position effects. Overall, it would appear that discriminability changes little over the course of the sequential stopping rule lineup, at least when there are six lineup members. It is possible that results would differ if a much larger number of lineup members was used (Levi, 2007, 2012). The feasibility of applying the ISL model to longer lineups is questionable, as the addition of each item requires a large increase in the number of participants in order to retain statistical power. Few previous studies have investigated how sequential item presentation affects discriminability (Palmer & Brewer, 2012; Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019), as distinct from the many studies that have investigated the effect of sequential item presentation on AUC. There is, at present, too little evidence to provide a strong indication as to the magnitude and direction of the true effects of sequential item presentation on discriminability, particularly given the potential methodological differences, approaches to stimulus selection and, potentially, sample characteristics of studies that have explored this question.

5.3.2 Response Bias

This thesis clarified the nature of the relationship between sequential item presentation and response bias, replicating both of the seemingly contrary results observed in previous studies. The sequential stopping rule lineup was associated with more conservative responding than the simultaneous lineup (Meissner et al., 2005; Palmer & Brewer, 2012), and the UK lineup was associated with more lenient responding than the simultaneous lineup (Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019). Responding was most conservative for the sequential stopping rule procedure in both studies one and two. This further demonstrates the robustness of this effect, which may be among the most reliable in the literature. In study two, responding was most conservative for sequential stopping rule presentation and most lenient for UK lineup presentation. This indicates that the divergent

pattern of results observed for the sequential stopping rule and UK lineup procedures in previous studies was likely attributable to a genuine difference in response bias between the two procedures, rather than differences in methodology or the characteristics of the participant sample. As for discriminability, these results imply that sequential item presentation alone exerts minimal influence on witness response bias. Other factors, particularly the presence or absence of a stopping rule, seem to have a much greater effect on response strategies.

In study three, the ISL model revealed an increase in the false alarm rate, i.e. response bias, over the course of the lineup prior to the presentation of the target and a decrease following the presentation of the target. That such a pattern of results is apparent in the data is a testament to the advantages of the ISL model for answering this question but, from an applied perspective, it is not a positive result for this lineup procedure. Differences in the probability of a false alarm at different serial positions imply that there are differences in the innocent suspect identification rate at different serial positions in real police lineups. This is counter to the aim of procedural fairness; innocent suspects should be equally likely to be selected (or rejected) no matter where they are placed in a sequential lineup.

In general, it seems that the response bias effects reported in the literature for sequentially presented lineups are both larger and more reliable than effects on discriminability. This might be because the aspects of the procedure that are likely responsible for response bias, such as the use of a stopping rule, are implemented in a relatively consistent way across studies. In contrast, aspects that are likely to affect discriminability tend to vary across studies, particularly in terms of the general difficulty of recognising the target stimuli, the type of distractor task and the duration between encoding and test.

5.3.3 Summary

This thesis advances the field's understanding of sequential item presentation. From a theoretical perspective, results indicate that the effect of sequential item presentation on decision behaviour may be weaker than previously assumed and that the conclusions of previous studies with respect to discriminability and response bias may be due to the interaction of sequential item presentation with other procedural factors. The end goal of measuring the effect of such procedural factors on latent variables is to guide the development of lineup procedures. If the results of this thesis reflect true underlying effects, there is minimal theoretical justification for designing lineup procedures that employ sequential item presentation. Across three studies, no sequentially presented lineup was associated with greater discriminability than the simultaneous lineup. There was also a complex pattern of response bias shifts over the course of the sequential stopping rule lineup, which cannot occur in the simultaneous lineup due to the structure of the task. Although, there is some evidence that the position of the target in the array affects witness responding in simultaneous lineups (Palmer et al., 2017).

One issue yet to be explored in the literature is whether the position of the target on the UK lineup affects discriminability and response bias. It is possible that the procedure may induce position effects similar to the primacy and recency effects found in list learning designs (e.g. Oberauer, 2003) because viewing all items twice before making an identification decision could be considered a kind of memory test where the aim is to remember which item, if any, is the target. Like the position effect investigation of the sequential stopping rule lineup in study three, one difficult aspect of conducting a study to address this question would be the requirement to collect a large sample, particularly if the lineup was conducted with the usual nine members.

5.4 Diagnostic Feature Detection Theory

As the studies in this thesis investigated the effects of sequential item presentation, the results are germane to the predictions of diagnostic feature detection theory (Wixted & Mickes, 2014). The diagnostic feature detection theory (DFDT) prediction that simultaneous item presentation is associated with greater discriminability than sequential item presentation was not supported in study one, but was supported by a small discriminability advantage for simultaneous presentation in study two. However, the effect in study two was restricted to the comparison between the simultaneous lineup and the sequential stopping rule lineup and was not evident when comparing the simultaneous lineup and the sequentially presented UK lineup. The prediction that discriminability increases over the course of the sequential lineup was partially supported in study three, although the increase only occurred from position one to position two. These results are mixed in their support for the predictions of theory. However, the design of these studies are limited in their ability to offer a strong test of DFDT. The aims of this thesis were best achieved by employing lineup tasks that were close analogues to those used in real police investigations. These tasks include procedural factors that may have counteracted the effects on discriminability predicted by DFDT, as discussed throughout (see sections 2.11.1, 3.7.1, 4.7.1 and 5.3.1). That is, the effects proposed by DFDT may have occurred in spite of no significant effect being observed in the predicted direction. In this case, it is unlikely that the results of these studies, and many others that employ close analogues of "real world" lineup tasks (e.g. Seale-Carlisle et al., 2019), are sufficient to falsify the theory.

As discussed in section 2.11.1, one aspect of providing strong tests of DFDT is the design of experiments that aim to isolate the specific effects proposed by the theory while controlling for confounding factors. One possibility would be to extend the "simultaneous showup" procedure from Colloff et al. (2019) to a sequentially presented lineup. Colloff and

Wixted (2019) presented a number of unidentifiable foils surrounding a single suspect who could be identified and compared this procedure to a single suspect showup. Discriminability was greater for the "simultaneous showup" condition, which provides support for the DFDT notion that providing the opportunity to compare across items improves discriminability. An analogous "sequential showup" experiment could be conducted where a varying number of unidentifiable foils are presented one at a time prior to a suspect who can be identified. This would be a more rigorous test of the prediction that discriminability increases with the opportunity to compare across increasing numbers of sequentially presented items. Perhaps discriminability improves for a certain number of items until interference takes hold, beyond which it decreases. This would be a stronger test of DFDT than comparing sequentially and simultaneously presented lineups. At a more basic level, it might be necessary to conduct face matching experiments using computer generated stimuli (see Carlson et al., 2019; Flowe & Ebbesen, 2007), manipulating the number of common features and the degree of distinctiveness of unique features. This would require a review of face perception literature to determine what constitutes a feature for the purposes of recognition and which features are important for face identification (Abudarham et al., 2019).

5.5 Limitations of Diagnostic Feature Detection Theory

Research drawing on DFDT often advocates for its use on the basis that it is formalised, contrasting it with verbal theories that offer less precision in their predictions (e.g. Carlson et al., 2019; Wooten et al., 2020). DFDT was originally instantiated in a signal detection framework, described in Section 1.8.2 (Wixted & Mickes, 2014) and researchers have adopted this framework to generate testable predictions about the effect of procedural changes on discriminability (Colloff et al., 2021; Wooten et al., 2020). However, this framework allows researcher degrees of freedom that affect the predictions of the model.

Take the DFDT prediction tested in previous studies (Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019) and in study two that the simultaneous lineup is associated with greater discriminability than the UK lineup. This prediction is shown in Table 5.1 using the formalisation from Wixted and Mickes.

Table 5.1

Study Two Prediction generated by DFDT Formalisation for the UK vs Simultaneous Lineup

Procedure	Suspect	Parameter	f1	f2	f3	f4	Σ	d_a
UK Lineup	Innocent	μ_{innocent}		1	0	0	1	1
		$\sigma^2_{\text{innocent}}$		1.5	1	1	3.5	
	Guilty	μ_{guilty}		1	1	1	3	
		σ^2_{guilty}		1.5	1.5	1.5	4.5	
Simultaneous Lineup	Innocent	μ_{innocent}			0	0	0	1.26
		$\sigma^2_{\text{innocent}}$			1	1	2	
	Guilty	μ_{guilty}			1	1	2	
		σ^2_{guilty}			1.5	1.5	3	

Note: f1 and f2 are features shared by the perpetrator and innocent suspect, f3 and f4 are features unique to the perpetrator

This example uses five features (f1 – f4), two uniquely shared by the perpetrator and two shared between the perpetrator and innocent suspect, which are likely to be general factors of witness's description such as race and age. Features shared by the guilty perpetrator and the lineup items are assumed to provoke stronger familiarity and therefore have a mean of 1 and a standard deviation of 1.22, while features not shared with the perpetrator have mean familiarity of 0 and a standard deviation of 1. Values are aggregated by summing the means and standard deviations of all features. The d_a measure can then be calculated for a given procedure:

$$d_a = \frac{\mu_{guilty} - \mu_{innocent}}{\sqrt{(\sigma^2_{guilty} + \sigma^2_{innocent})/2}}$$

This statistic quantifies the difference between the means of the perpetrator and innocent

suspect feature distributions while accounting for differences in their standard deviation. Table 5.1 shows the DFDT prediction under the assumption that not all common features are discounted on the UK lineup because items are sequentially presented, leading to lower d_a compared to the simultaneous lineup.

However, DFDT can also predict equivalent discriminability for the two procedures. Wixted and Mickes (2014) argue that the same amount of diagnostic information is available at the end of the sequential stopping rule lineup as for a simultaneous lineup of equivalent length, because all items have been seen by the end of a sequential stopping rule lineup. They surmise that this should lead to a discriminability increase over the course of the sequential stopping rule lineup and equivalent discriminability between the two procedures by the end of the sequential lineup. In this case, a discriminability decrement for sequential presentation is explained by the stopping rule. Some witnesses identify early and do not see all lineup items, thereby failing to discount common features, which reduces discriminability at the group level. However, this does not occur on the UK lineup; all items are seen prior to an identification being made. Thus, on simultaneous and UK lineups of the same length, witnesses should be able to discount all common features on either procedure, resulting in equivalent discriminability. This prediction is formalised in Table 5.2.

Table 5.2

Alternative Prediction generated by DFDT Formalisation for the UK vs Simultaneous Lineup

Procedure	Suspect	Parameter	f1	f2	f3	f4	Σ	d_a
UK Lineup	Innocent	μ_{innocent}			0	0	1	1.26
		$\sigma^2_{\text{innocent}}$			1	1	2	
	Guilty	μ_{guilty}			1	1	2	
		σ^2_{guilty}			1.5	1.5	3	
Simultaneous Lineup	Innocent	μ_{innocent}			0	0	0	1.26
		$\sigma^2_{\text{innocent}}$			1	1	2	
	Guilty	μ_{guilty}			1	1	2	
		σ^2_{guilty}			1.5	1.5	3	

There is nothing within the DFDT framework that indicates which of these two contrary predictions is consistent with the theory. Rather, justifying whether the UK lineup will impair discriminability relies on invoking empirical results external to the theory in relation to a feature detection mechanism that is unspecified.

A similar issue affects the prediction of Wooten et al. (2020) that the number of foils on the simultaneous lineup does not affect discriminability if they are sufficiently well matched to the suspect. In Wooten et al. (2020), this prediction is generated using six features, two unique to the perpetrator and four shared by the perpetrator and innocent suspect. The argument made by Wooten et al. (2020) is that relatively few well matched foils will allow the discounting of all four common features and thus additional foils provide no benefit, i.e. $d_a = 1.26$ for either a 3- 6- or 9-item lineup, similar to the prediction shown in Table 5.2. It is unclear whether their results conform to this prediction because they employed ROC analysis, which cannot separate differences in discriminability and structural effects caused by changes to lineup size (see Section 2.4.4). It is possible to mount a plausible alternative argument; that two foils is insufficient to discount all common features, no matter how well matched they are to the suspect. The DFDT formalisation could be used to generate this alternative prediction that a 3-item lineup will lead to lower d_a than a 6- or 9-item lineup by simply adjusting how many of the four common features are discounted in the 3-item condition.

In summary, DFDT's predictions about the effects of procedural manipulations on discriminability are not well constrained, which makes the theory difficult to falsify. The set of possible predictions are limited only by the set of explanations that researchers and peer reviewers are willing to entertain for why a given manipulation should or should not impair the ability to discount common features. Constraining DFDT's predictions and improving falsifiability may require the development and integration of a formal process account of

feature detection. A recent paper Colloff et al. (2021) provides an extended examination of feature detection within DFDT, which may foreshadow further developments in this area. It may be necessary to revisit global memory matching models such as WITNESS (Clark, 2003), which more explicitly represent the matching of features stored in memory and features of a test item. This may allow for falsification of the model.

5.6 The Limitations of Receiver Operating Characteristic (ROC) Analysis

This thesis emphasised a major limitation of ROC analysis as a measure of discriminability; the shape of an ROC for a given lineup procedure may be affected by aspects of the tasks unrelated to discriminability. In this case, Area Under the Curve (AUC) for a procedure reflects some unknown combination of task constraints and discriminability, which means that discriminability and AUC can be dissociated (Wixted & Mickes, 2018). As discussed in Section 2.4.4, the stopping rule constraint of the Lindsay and Wells (1985) sequential lineup means that its ROC curve is non-monotonic, falling back toward the major diagonal representing chance performance in the region of ROC space where responding is very lenient (Cohen et al., 2020; Rotello & Chen, 2016; Wilson et al., 2019). This means that a difference in AUC between the sequential stopping rule lineup and other lineup tasks may not reflect a difference in discriminability. This occurred in study three, where there was no significant difference in AUC between the simultaneous and sequential stopping rule lineups, but discriminability was greater for the simultaneous lineup when estimated by the Ensemble model. This demonstrates that AUC is a poor measure of discriminability for certain lineup tasks, and should not be used to compare lineup tasks when they differ in structure. While many authors acknowledge the possibility of dissociation between AUC and discriminability, they also tend to present the analyses as equivalent for applied purposes (e.g. Colloff & Wixted, 2019), adopting the position that AUC and discriminability as estimated by a signal detection model are generally in agreeance about which procedure is superior (Wixted &

Mickes, 2018). This view should be revised. AUC and discriminability will tend to agree about which of two lineup procedures is superior *except* when the procedures differ in structure. This is consistent with the results of previous studies that have employed both ROC analysis and models to compare lineup procedures. Studies finding no dissociation between discriminability and AUC have manipulated aspects of the lineup procedure that did not alter structural aspects of the lineup task (Colloff et al., 2016; Colloff et al., 2017; Colloff & Wixted, 2019). Some studies have manipulated task structure and would therefore be expected to show a dissociation, but conducted only ROC analysis (Gronlund et al., 2012; Neuschatz et al., 2016; Wooten et al., 2020). Finally, one study on which a dissociation would be expected because task structure was manipulated – the position of the target in the sequential stopping rule lineup – reported results consistent with such a dissociation (Wilson et al., 2019). Researchers should carefully consider whether their experiments manipulate task structure before citing Wixted and Mickes (2018) as a justification for interpreting differences in AUC between conditions as differences in discriminability.

5.7 Methodological Limitations

One possible limitation of the studies in this thesis is that participants seemed to find the memory test relatively easily, as indicated by discriminability scores of close to two in all studies. This is substantially stronger memory than reported by Seale-Carlisle et al. (2019), the most recent study other than those conducted here to find greater discriminability for a simultaneously presented lineup. The high overall discriminability across all tasks may have led to a kind of ceiling effect, where differences in discriminability between lineup tasks could not be expressed. There are two aspects of the experimental procedure might be responsible for the apparent ease of the memory test.

First, the distractor task employed between the encoding and test phases was relatively short at around 90 seconds. This does not leave much time between encoding and

test and is certainly much less than in real criminal cases. It would have been ideal to use a longer interval, but this is difficult to implement in Mechanical Turk studies. Participants must be paid for all time spent completing the task, so a longer interval increases the cost per participant for running experiments or requires participants to complete a two-part task. Other than offering financial incentives, there is little to prevent participants from simply collecting their payment for an encoding phase and failing to return for a test phase. The cost issue is particularly problematic for lineup studies as large samples are required because each witness only provides one lineup decision.

Second, it may have been relatively easy to encode the relevant information from the stimulus videos. Each video lasts approximately 10 to 15 seconds with the target as the only visible person. Thus, there is almost no information for the participant to encode other than the identity of the target and no other aspects of the video that might divert their attention from encoding the target. Resource constraints prevented the production of more elaborate, ecologically valid videos for each target. Target videos were filmed as part of a project to create a large corpus of target videos and accompanying photographs, totalling 194 targets in all. The intention was to use these stimuli in face-matching and lineup studies. Another aspect of the stimuli that may have reduced task difficulty is the use of female faces. There is some evidence that female faces are more easily recognised compared to male faces, particularly by female participants (Cross et al., 1971; Lewin & Herlitz, 2002). Female appearance tends to be more variable than male appearance due to a tendency for greater differences in, for example, hair style (Wright & Sladden, 2003). These additional cues may increase the distinctiveness of female faces, facilitating recognition (Valentine, 1991).

5.8 Stimulus Selection

An important aspect of lineup research is controlling for the presence of stimulus effects. The approach adopted in this thesis was to select a pool of sixteen similar stimuli that

such that, for any pool item selected as the target, all remaining pool members are appropriately matched foils. This increases ecological validity because it bears a closer resemblance to the real police lineup task, in which all witnesses see different lineups. This approach differs from early lineup studies in that it aims to increase variability in the stimuli in order to "wash out" stimulus effects, rather than aiming to control variability by selecting one target and accompanying set of "unbiased" foils (Clark et al., 2008; Wells & Windschitl, 1999). Other studies have employed a similar approach by randomly selecting foils for a single target from a large pool on each trial (e.g. Mickes et al., 2012; Wilson et al., 2019) or using multiple targets, each with a set of foils (e.g. Colloff et al., 2016; Colloff & Wixted, 2019). One way in which this pool approach might be extended is to integrate a face generation algorithm, such as a Generative Adversarial Network (Karras et al., 2019), into the delivery of a face matching experiment. Target faces could be generated at random for each participant in the study phase. In the test phase, foil faces for a series of lineup trials could be generated for the target based on some kind of similarity parameters that govern the behaviour of the algorithm. In this way, it would be possible to approximate the real-world situation of all witnesses seeing different lineup. The similarity tuning parameters could be set to either hold constant or vary the similarity of the lineups seen by each participant. One limitation of this approach for the traditional "one shot" lineup paradigm is the difficulty of generating simulated crime videos for artificially generated target faces. However, multiple trial face matching experiments may provide a path to theoretical explanations at the level of the individual witness.

5.9 Future Directions

The results of this thesis provide many avenues for further exploration. In general, relatively few published studies have fit signal detection models to data in order to compare discriminability and response bias between simultaneous and sequentially presented lineups

(Palmer & Brewer, 2012; Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019). This thesis adds to this limited body of empirical evidence but it is not yet possible to draw any strong conclusions about the effect of sequential item presentation on discriminability and response bias. Additional studies, conducted with appropriate sample sizes, are required to gain a better understanding of the direction and size of the population-level effects.

5.9.1 The Absence of a Stopping Rule

This thesis investigated the sequential stopping rule lineup proposed by Lindsay and Wells (1985) because it is one of the most heavily studied procedural manipulations in lineup research (Steblay et al., 2011b). This body of research led to its adoption in many jurisdictions. However, surveys of police jurisdictions in the US reveal that none currently enforce the stopping rule and that generally witnesses may request a second lap through the lineup items (Wells, 2014). While the stopping rule is of theoretical interest due to the constraints it places on the task, investigating the task in this form is of minimal applied relevance if the stopping rule is not enforced in real lineups. When the sequential lineup is conducted without a stopping rule, multiple identifications are generally allowed, as there is little utility in allowing witnesses to continue viewing items after an identification if subsequent items cannot be identified. Future model based research is required to assess the effect of allowing witnesses to make multiple identifications on a sequential lineup on which a yes/no response is elicited for each item (Wilson et al., 2019), in addition to revisiting the effect of allowing multiple laps (Horry et al., 2015). Such an experiment could employ a similar procedure to (Wells et al., 2015a), allowing participants to make multiple identifications but requiring them to adjudicate between them to produce a final decision at the end of the procedure.

With yes/no responses for each item, it would then be possible to use different rules for counting identifications, perhaps comparing a dataset in which the first identification is

counted to dataset where some rule is developed for counting subsequent identifications. The difficulty with any such rule is that counting the *i*-th of multiple identifications rules out identifications at certain serial positions. For example, always taking the second of multiple identification decisions will result in counting no identifications decisions made at position one. Based on the results of study three, this may affect the aggregated data due to differences in response bias at different serial positions. Datasets using different rules for counting multiple identifications could be compared to data where participants' final adjudicated identification is counted in order to establish whether allowing witnesses to adjudicate between identification decisions is of detriment or benefit to memory. This would give a better understanding of how the sequential lineup procedure as actually employed in many jurisdictions compares to the simultaneous lineup (Wells et al., 2015a).

5.9.2 The UK Lineup Procedure

Study two is the first study to compare the UK lineup to the simultaneous *and* sequential stopping rule lineup procedures. The UK lineup procedure merits further study given that it is used in real police investigations and only two published studies have investigated it to this point (Seale-Carlisle & Mickes, 2016; Seale-Carlisle et al., 2019). As the discriminability decrement for UK lineup presentation compared to simultaneous lineup presentation reported in these studies was not replicated in study three, the true direction and effect size of the difference in discriminability between these procedures is unclear. As for sequential item presentation more generally, additional evidence in the form of replication studies may clarify the size and direction of the true effect in the population. If the UK lineup does impair discriminability, then it will be necessary to clarify which aspects of the procedure are responsible. Seale-Carlisle et al. (2019) concluded that sequential item presentation was the primary factor, but the findings of both studies one and two in this thesis are not consistent with this view. The analogue of the UK lineup procedure used in study

three differed in many ways from that used in Seale-Carlisle and Mickes (2016) and Seale-Carlisle et al. (2019). Study three used six rather than nine lineup items, which were still photos shown for five seconds rather than 15-second long videos of a rotating head-and-shoulders profile. Additionally, study three allowed participants to revisit items before identifying. In a series of experiments, Seale-Carlisle et al. (2019) concluded that discriminability was minimally affected by item type (photo vs. video), number of items (six vs. nine) and task duration. However, it could be that these factors interact in some way that is not evident when they are isolated and that lesser discriminability would have been observed for the UK lineup if the analogue in study two more closely matched the task used in Seale-Carlisle et al. (2016) and Seale-Carlisle et al. (2019). Further research is needed to understand the conditions under which the UK lineup will lead to lesser discriminability than the simultaneous lineup.

The lenient responding for the UK lineup compared to the simultaneous lineup reported in Seale-Carlisle and Mickes (2016) and Seale-Carlisle et al. (2019) was replicated in study three; whatever aspect of the procedure is responsible for this appears to have been present in both studies. This finding merits further investigation, as lenient choosing may not be desirable from a policy perspective and is the opposite effect to that which was intended by Lindsay and Wells (1985) when notion of a sequentially presented lineup was first introduced. It is not immediately clear why responding on the UK lineup should be more lenient than the simultaneous lineup. The tasks share a similar format; both involve the presentation of a set of lineup items, after which a single lineup decision is expressed. There is no theoretical explanation for why the UK lineup encourages participants to choose more readily. None of the current models predict this result and so it remains an observation that must be explained by a process model.

5.9.3 Item Similarity

The stimulus pool approach employed in this thesis attempted to mitigate effects caused by differences in item similarity both within the lineup and between the target and the foils. The overall similarity of lineups, and the similarity of the foils to the target, randomly varied between participants, which should have minimised the possibility that results were due to the use of a particular stimulus set. This thesis did not explore the relationship between item similarity, discriminability and response bias, partly because subsetting the data by the sets of items contained in each lineup shown to participants reduces power to such an extent that analysis is unviable. Recent research has suggested discriminability is poor when the similarity between items in a lineup is both very low and very high, i.e. that there is an inverse U-shaped function between within-lineup similarity and discriminability (Carlson et al., 2019). Item similarity varies across studies and the effect of item similarity on discriminability may interact with effects caused by other procedural manipulation. Characterising the relationship between within-lineup similarity and discriminability may therefore provide a path to integrating the results of lineup studies that have reported diverging patterns in discriminability when comparing the same procedural manipulations,. Some changes to methodology would be required if the aim is to have a "similarity profile" for all stimuli in published eyewitness memory studies. A starting point is to collect pairwise similarity ratings for all lineup items in a separate experiment.

Item similarity likely has a greater effect on sequentially presented lineups compared to the simultaneous lineup. This is because the relative similarity of all items can be determined more or less immediately on the simultaneous lineup, but such information may not be available immediately, or at all, on sequentially presented lineups. The position effects observed in study three for the sequential stopping rule lineup may be particularly susceptible to differences in item similarity. One possibility is that witnesses may not identify the most

familiar item to the perpetrator in the set of items because an item presented prior to the most familiar item exceeds their decision criterion (Clark & Davey, 2005). Additionally, there may be large changes in criterion between serial positions when an item that is relatively similar to the target follows an item that is relatively dissimilar. This is a generalisation of the increase in response bias observed after the presentation of the target in study three, which the ISL model would need to be reparameterised to accommodate. The proposed effect rests on the assumption that witnesses behave rationally in adjusting their criterion over the course of the sequential stopping rule lineup. That is, any identified item must be more familiar than the items rejected at previous serial positions. It would be irrational to reject a highly familiar item at position one and then identify a relatively unfamiliar item at position two. It is currently unclear whether participant responses conform to this expectation.

5.9.4 Shifting the Research Paradigm

Much lineup research works from the standpoint of improving on a pre-existing framework for collecting eyewitness evidence. It is generally assumed that legal system requires a binary decision to either identify from a lineup or reject from witnesses due to the way in which identification evidence is typically presented at trial. A major research-based reform to this framework is the policy of collecting of post-decision confidence immediately following identification decisions (Brewer, 2006; Brewer & Wells, 2006). Simplified, the prevailing view is that confidence and accuracy of identifications (but not rejections) are positively associated at the group level, particularly when confidence is high (Wixted & Wells, 2017). On this basis, triers of fact may make inferences about the reliability of witness decisions on the basis of post-decision confidence estimates. Within this binary decision paradigm, lineup research has aimed to isolate procedural factors that affect the accuracy of lineup decision outcomes and the latent variables that determine them. However, the small effects found within these studies (and others) in relation to discriminability raise the

possibility that the field is expending significant effort testing procedural changes that are of minimal applied benefit (Brewer & Doyle, 2021). On this basis, researchers may need to free themselves from the constraints of the lineup task as currently studied in order to more effectively address the aims of lineup research. In this vein, some researchers have set out to develop alternative procedures that eschew the binary identification/rejection decision. Brewer et al. (2020) proposed a lineup on which confidence ratings are collected for every member, while Mu et al. (2017) proposed a lineup on which two-alternative forced choice decisions are made for each possible pairwise item comparison. A procedure on which all items are ranked as matches to a witness's memory of the perpetrator might also be adapted from tasks in the basic recognition memory literature (Kellen & Klauer, 2014). Effectively, these procedures all provide a ranking for witnesses' belief that each lineup item is the perpetrator (McCormick et al., 2019).

There are two general arguments advanced in favour of these procedures. First, they offer additional information about witness beliefs compared to a binary decision and confidence estimate. A profile of confidence for a lineup provides information about witness beliefs that each item is the perpetrator, in addition to their beliefs relative to the other items. The pairwise procedure provides a ranking of preference for which item is the perpetrator in addition to a measure of reliability by determining the extent to which witnesses violate transitivity when making pairwise choices. That is, if item A is preferred to B, and B is preferred to C, then C should not be preferred to A. Theoretical research in recognition memory has shown that there is memorial information contained in the second and third ranked choices of a forced choice ranking task (Kellen & Klauer, 2014). Second, these alternative procedures provide a more fine grained scale of evidence for suspect guilt in the form of the confidence rating or the preference/ranking assigned to the suspect as perpetrator relative to the foils. The higher the ranking for the suspect relative to the foils, the stronger

the evidence for suspect guilt. It is possible that this evidence could be presented at trial in place of binary judgements, which may lead to an improvement in criminal justice outcomes.

From a theoretical perspective, the absence of a binary judgement complicates but does not preclude the measurement of discriminability. It is possible to calculate a measure of discriminability from the forced-choice ranking task (Kellen & Klauer, 2014). Interestingly, this task eliminates response bias entirely as there is no decision threshold. It may be possible to adapt a signal detection model of a classification task to estimate discriminability for the confidence-based procedure and the pairwise procedure. In order to fit these models, it may be necessary to develop a rule for converting observed rankings to binary decisions, which somewhat defeats the purpose of the approach. Additionally, as these procedures focus on gradated evidence of suspect guilt, it is necessary to designate innocent suspects for target absent lineups in experimental studies (e.g. Brewer et al., 2020).

This shift away from the binary decision outcomes proposed procedures provides a path forward for eyewitness science that is grounded in theoretical approaches to memory and decision making and has the potential to improve on the existing binary decision framework that constrains research activity. As indicated by Brewer and Doyle (2021), if these procedures are to be adopted by the stakeholders in the legal system, it is critical that these stakeholders have input in to the development of these procedures.

5.10 The Promise of Signal Detection Theory

The introduction of signal detection theory (SDT) to eyewitness memory research provided a framework for increasing understanding by addressing the limitations of previous approaches to understanding witness decision behaviour. This might have started the building of a renewed focus on building a theoretical base for guiding lineup research (e.g. Clark, 2008; Turtle et al., 2008; Wells, 2008). This flowering of theoretically motivated lineup research is yet to occur. Rather, researchers have retained their applied focus but adopted

AUC as the preferred measure in place of the diagnosticity ratio. A program of research comparing AUC between procedures that differ more or less arbitrarily may establish which manipulations maximise discrimination accuracy, but is unlikely to increase theoretical understanding, particularly if researchers are not sensitive to the conditions under which AUC and discriminability are dissociated.

SDT has been applied to many domains and tasks for both theoretical and applied purposes (Goodenough et al., 1972; Hutchinson, 1981). However, some lineup researchers have interpreted and applied SDT in rather idiosyncratic ways that seem disconnected from the framework's history and underlying principles. For example, Lee and Penrod (2019) recently proposed a "multi-d" model that calculates pairwise *d'* from the yes/no task between each item type; target, innocent suspect and foils. This was partly motivated by the incorrect assumption that models previously employed in the literature, i.e. SDT-MAX and SDT-INT, collapsed the 2 x 3 decision outcomes of the lineup in to 2 x 2 structure of the yes/no task by treating foil identifications as rejections. As the multi-d model effectively treats the lineup as a series of pairwise yes/no detection tasks between item types, it does not capture the structure of the lineup task. Using it to compare lineup tasks that differ in structure may therefore lead to inaccurate estimates of the difference in discriminability between procedures.

In another example, Smith et al. (2020) recently proposed a method for constructing ROC curves that extend through the entirety of ROC space in order to avoid some of the limitations inherent in comparing procedures using partial AUC (Lampinen, 2016; Smith et al., 2019; Wixted et al., 2017). Smith et al. (2020) first propose that the field should focus primarily on the detection task completed by the investigator to determine whether the suspect should be charged or released, not the memory test conducted by the witness. They assert that investigators use witness identification decisions to set their response criteria for

charging or releasing a suspect. Based on previous research stating that the diagnosticity ratio largely indexes response bias (Wixted & Mickes, 2012), they propose calculating a diagnosticity ratio for each possible decision outcome at each level of confidence and ordering the points in the ROC from highest diagnosticity ratio (most conservative) to lowest (most lenient). This involves plotting target, foil and innocent suspect identifications, as well as misses and correct rejections, on the same two-dimensional axis. They also proposed an alternative method to sorting by diagnosticity for ordering points in ROC space and indicated that did not lead to significant differences in the resulting ROC curves. There are a number of shortcomings of this approach.

Smith et al. (2020) are correct that the witness and investigator complete separate signal detection tasks, but it is not theoretically sound to specify an explicit mapping that links witness decisions to investigator decisions. Investigators integrate information from witnesses as part of their decision to charge or release a suspect. The assumption of many SDT models, including those used here, is that evidence from different sources is integrated into a unidimensional signal strength axis. SDT does not specify the exact process by which investigators integrate this evidence, as Smith et al. (2020) attempt. Building on this wayward assumption, Smith et al. (2020) coerce the multi-class data from a lineup into two dimensions by assuming that each lineup decision outcome in a given confidence category is associated with a particular level of evidence for suspect guilt/innocence. That is, that the multi-class points from witness decisions can be ordered in the investigator's ROC space along a single dimension of guilt. As per standard practice (Gronlund et al., 2014), the target identification rates and innocent suspect identification rates at different levels of confidence are treated as hit rates and false alarm rates pair for the purposes of plotting the ROC curve. Treatment of other response classes is somewhat unusual. The target present foil identification rates and target absent foil identification rates are treated as hit and false alarm rates, respectively, and

plotted against each other, as are the miss rates and correct rejection rates. All points are then ordered from lenient to conservative according to either the diagnosticity ratio or some other rule – not according to confidence ratings – and plotted to produce a ROC curve that extends from zero to one in ROC space. Allowing multiple methods for deciding on the order of the points in ROC space is open to researcher degrees of freedom. A recent study employing the Smith et al. ROC curves ordered the points according to a different *a priori* scheme rather than by diagnosticity ratio (Lucas et al., 2020). There are likely to be conditions under which different methods of ordering the points lead to different conclusions about the data, particularly if other researchers propose additional methods for ordering the points that are acceptable to reviewers. If the aim is to address the shortcomings of partial AUC by including all lineup decision outcomes in an ROC-type analysis, a more principled option may be to explore the volume under the ROC surface, which generalises two-dimensional ROC analysis to multi-class classification problems (Kang & Tian, 2013).

These examples demonstrate the potential for a robust and longstanding theoretical approach like SDT to be misapplied when research lacks engagement with foundational principles. The theoretical promise of SDT for lineup research lies with its ability to measure underlying variables of interest. Improper application of SDT compromises this measurement ability, which negates the benefit of applying the theory in the first place. If the field adopts a version of SDT disconnected from fundamental principles, it may hamper theoretical development and once again lead to the recommendation procedures that have unintended effects on witness decision behaviour, as in the case of the sequential stopping-rule lineup (Clark, 2012a).

5.11 Conclusion

The aim of this thesis was to investigate the effect of sequential item presentation on discriminability and response bias using signal detection measurement models. These models

improve on existing measures of lineup task performance that confound discriminability and response bias (Wixted & Mickes, 2012) and discriminability and the effect of task structure. To ensure measurement accuracy and to allow comparison between tasks, it is critical to ensure a suitable degree of match between the model, the lineup task and the research question under investigation. To this end, novel models of the sequential stopping rule lineup (Lindsay & Wells, 1985) were developed, SDT-SEQ and the ISL model. Both models account for the stopping rule constraint, while the ISL model also accounts for the serial position in the lineup at which an identification is made in order to investigate serial position effects. Across three studies, these models of the sequential lineup, along with models of the simultaneous lineup, were fit to both newly collected and archival experimental lineup data. In each experiment, participants encoded a target video and, after completing a distractor task, were presented with a police lineup that either did or did not contain the target. Participants could either identify a lineup item or reject the lineup, and provided a post-decision confidence judgement. Studies one and two manipulated the presentation format of lineup items, while study three manipulated the position of the target in the sequential stopping rule lineup.

In studies one and two, there was some evidence that sequential item presentation impairs discriminability to a small extent compared to simultaneous item presentation, although this effect was not found in all datasets. Conducting a sequential lineup with a stopping rule caused participants to respond more conservatively compared to a simultaneous lineup or the sequentially presented UK lineup, on which responding was most lenient. These results imply that factors other than sequential item presentation may be primarily responsible for differences in response bias between lineup procedures. In study three, there was a small increase in discriminability from serial position one to position two in a new experiment, but not in data from a previous experiment that had reported the same effect based on parameter

estimates from SDT-SEQ (Wilson et al., 2019). A unique aspect of the ISL model fit to these data is that it distinguishes between response bias prior to and following the appearance of the target in the lineup. Responding prior to the presentation of the target became more conservative over the course of the lineup and there was a shift to conservative responding immediately following the presentation of the target, which weakened as the lineup progressed.

Each study in this thesis also provided a test of some high-level predictions of diagnostic feature detection theory. As a whole, results were generally inconclusive with respect to the validity of the theory's somewhat loosely constrained predictions. Testing this theory will require the design of more focused experiments that bear less relation to real lineup tasks. From an applied perspective, this thesis implies that the designers of future lineup procedures should be wary of adopting sequential item presentation. There appears to be no benefit to discriminability and the effect on response bias is inconsistent.

This thesis demonstrates that adopting a measurement model approach can advance understanding in eyewitness memory research by quantifying the effects of changes to lineup procedure on latent variables. It is hoped that the work conducted here inspires further model-based lineup research.

Appendix A – Model Equations

We suppose that each witness is presented with a lineup consisting of *n* items. Let $[n] = \{1, \ldots, n\}$ be the set of item positions in the lineup and, for a given witness, let $X = \{x_i : i \in [n]\}$ be the set of familiarity values of the corresponding items. In a target present (TP) lineup, there is one *target* and $n-1$ *foils*. Similarly, in a target absent (TA) lineup, there may be one designated *suspect* and $n-1$ foils. Alternatively, if there is no designated suspect then the lineup consists only of foils.

We assume that each x_i is a value of a continuous random variable that defines a probability distribution for that item. We also assume that the values of each random variable are independently and identically distributed. Let $T(\cdot)$ and $t(\cdot)$ be the cdf and pdf of the target distribution and let $S(\cdot)$ and $s(\cdot)$ be the cdf and pdf of the designated innocent suspect distribution. We assume that the random variables corresponding to foils all have the same distribution. Accordingly, let $F(\cdot)$ and $f(\cdot)$ be the cdf and pdf, respectively, of the common foil distribution.

The witness identification task can be notionally partitioned into a *detection task* and an *identification task* (Duncan, 2006). The aim of the detection task is to determine if the lineup contains a target. The aim of the identification task is to determine which lineup item is the target. Accordingly, we define three events of interest. In a TP lineup, a *target detection* (TD) occurs when a witness identifies any item as the target and a *target identification* (TID) occurs when the item so identified is in fact the target. In a TA lineup, a *false alarm* (FA) occurs when a witness identifies any item as the target and a *suspect identification* (SID) occurs when the item so identified is the designated innocent suspect (if present). From these events, additional events can be defined. A *miss* occurs when there is no target detection to a TP lineup; a *correct rejection* occurs if there is no false alarm to a TA lineup; a *foil*

identification (FID) occurs if there is target detection but no correct identification in a TP lineup or if there is a false alarm but no suspect identification in a TA lineup.

When a witness identifies a lineup item, they assign to it a confidence level according to a set of *k* decision criteria, $C = \{c_i : i \in [k]\}$. For a given $c \in C$, we define $P_{TID}(c)$, $P_{TD}(c)$, $P_{SID}(c)$ and $P_{FA}(c)$ as the proportions of target identifications, target detections, suspect identifications, and false alarms, respectively, assigned a level of confidence at least as great as *c*. We note that $P_{FID}(c) = P_{FA}(c) - P_{SID}(c)$ for TA lineups and $P_{FID}(c) = P_{TD}(c) - P_{TID}(c)$ for TP lineups. If there is no designated suspect then $P_{SID}(c) = P_{FA}(c)/n$.

We assume that the elements of *C* are ordered $c_1 < c_2 < \ldots < c_k$. The expected proportion of target identifications in confidence band *i* is then equal to $P_{TD}(c_i) - P_D(c_{i+1})$ for $i < k$ and equal to $P_{TD}(c_i)$ for $i = k$. The proportion of non-identifications (i.e., misses) is equal to $1 - P_{TD}(c_1)$. Similar considerations apply to false alarms and correct rejections as well as to target and suspect identifications although these cases, these are undefined if no identification is made.

SDT-MAX

SDT-MAX is a simultaneous lineup model that implements the following decision rule. Let $m \in [n]$ such that $x_m = \max(X)$. Then if $x_m \geq c_1$, identify item *m* as the target at confidence level, $c = \max(\{c \in C : c \leq x_m\})$, otherwise reject the lineup. The general equations for this model are as follows.

For a TP lineup,

$$P_{TD}(c) = 1 - T(c)F(c)^{n-1}.$$

Let $t \in [n]$ be the position of the target item in a TP lineup. Then $P_{TID}(c)$ is the joint probability that $m = t$ and $x_m \geq c$. The probability that $m = t$ is

$$\Pr\left(m = t\right) = \int_{-\infty}^{\infty} t\left(x\right) F\left(x\right)^{n-1} dx \ . \tag{1}$$

The additional requirement that $x_m \geq c$ gives the equation,

$$P_{TID}\left(c\right) = \int_{c}^{\infty} t\left(x\right) F\left(x\right)^{n-1} dx \, .$$

Similarly, for a TA lineup with a designated innocent suspect,

$$P_{FA}\left(c\right) = 1 - S(c) F\left(c\right)^{n-1}$$

and

$$P_{SID}\left(c\right) = \int_{c}^{\infty} s\left(x\right) F\left(x\right)^{n-1} dx \, .$$

If there is no designated innocent suspect then

$$P_{FA}(c) = 1 - F(c)^{n}$$

and

$$P_{SID}\left(c\right) = P_{FA}\left(c\right) / n.$$

Normal distribution implementation. In order to fit this and other models we present, it is necessary to specify the forms of the different probability distributions. Following standard practice, we assume that they are normal. In this and following sections, let $\phi(x; \mu, \sigma)$ be the normal pdf with mean μ and standard deviation σ, evaluated at $x \in \mathbb{R}$, and let $\Phi\left(x; \mu, \sigma\right) = \int_{-\infty}^{x} \phi\left(x; \mu, \sigma\right)$ be the corresponding normal cdf. Let μ_t and σ_t be the mean and standard deviation of the target distribution and let μ_s and σ_s be the mean and standard deviation of the innocent suspect distribution. Without loss of generality, the mean and standard deviation of the foil distribution is zero and one, respectively. In this case, we call $\phi(x; 0, 1)$ and $\Phi(x; 0, 1)$ the standard normal cdf and pdf respectively, and write $\phi(x; 0, 1) = \phi(x)$ and $\Phi(x; 0, 1) = \Phi(x)$.

Substituting $\phi(x; \mu_t, \sigma_t)$ for $t(x)$, $\phi(x; \mu_s, \sigma_s)$ for $s(x)$ and $\phi(x)$ for $f(x)$, we derive the following equations for SDT-MAX:

$$P_{TID}(c) = \int_c^{\infty} \phi(x;\mu_t,\sigma_t)\Phi(x)^{n-1}\,dx$$
$$P_{TD}(c) = 1-\Phi(c;\mu_t,\sigma_t)\Phi(c)^{n-1}$$
$$P_{SID}(c) = \int_c^{\infty} \phi(x;\mu_s,\sigma_s)\Phi(x)^{n-1}\,dx$$
$$P_{FA}(c) = 1-\Phi(c;\mu_s,\sigma_s)\Phi(c)^{n-1}.$$

If there is no designated innocent suspect then

$$P_{FA}(c) = 1-\Phi(c)^n$$
$$P_{SID}(c) = \frac{1}{n}\left(1-\Phi(c)^n\right).$$

SDT-INT

The SDT-INT model is a simultaneous lineup model that implements the decision rule: if $\text{sum}(X) \geq c_1$ then choose lineup item *m* at confidence level $c = \max\left(\{c \in C : c \leq \text{sum}(X)\}\right)$, otherwise reject the lineup. The general equations for this model are as follows:

For a TP lineup,

$$P_{TD}(c) = \int_{-\infty}^{\infty} t * f^{*(n-1)}(x)dx\,,$$

where $t * f$ is the convolution of the density functions *t* and *f*, and f^{*n} is the convolution of *f* with itself *n* times. That is, $f * f = f^{*2}$. Then $P_{TID}(c)$ is the joint probability that $\text{sum}(X) \geq c$ and $m = t$. This can be expressed as the following product,

$$P_{TID}(c) = \Pr(\text{sum}(X) \geq c \mid m = t) \cdot \Pr(m = t)$$

where $\Pr(m = t)$ is given by Equation 1 above. We now require an equation for the conditional probability $\Pr(\text{sum}(X) \geq c \mid m = t)$.

Let $X \setminus \{x_m\} = \{x_i \in X : i \neq m\}$ be the set of memory strengths of the $n-1$ items, excluding the maximum and let $Y = \text{sum}(X \setminus \{x_m\})$. Then

$$\Pr(\text{sum}(X) \geq c) = \Pr(Y \geq c - x_m).$$

Because Y is the sum of $n-1$ random variables, its pdf is equal to the convolution of the component pdfs, truncated at the upper limit of x_m. Let f_x be the distribution f truncated at x. That is,

$$f_x(u) := \begin{cases} \dfrac{f(u)}{F(x)} & u \le x, \\ 0 & u > x \end{cases}$$

Suppose $m = t$. Then,

$$\Pr(Y \ge c - x) = \int_{c-x}^{\infty} f_x^{*(n-1)}(u)\, du$$

and hence

$$P_{TID}(c) = \int_{-\infty}^{\infty} t(x) F(x)^{n-1} \int_{c-x}^{\infty} f_x^{*(n-1)}(u)\, du\, dx$$

Similarly, for a TA lineup with a designated innocent suspect,

$$P_{FA}(c) = \int_{-\infty}^{\infty} s * f^{*(n-1)}(x) dx,$$

$$P_{SID}(c) = \int_{-\infty}^{\infty} s(x) F(x)^{n-1} \int_{c-x}^{\infty} f_x^{*(n-1)}(u)\, du\, dx.$$

If there is no designated innocent suspect,

$$P_{FA}(c) = \int_{c}^{\infty} f^{*n} dx.$$

Gaussian Implementation. Because the convolution of two more Gaussian distributions is Gaussian, it follows that:

$$P_{TD}(c) = 1 - \Phi\left(c; \mu_t, \sqrt{n - 1 + \sigma_t^2}\right).$$

Let ϕ_x be the standard normal pdf truncated at the upper limit x and let μ_x and σ_x^2 be its mean and variance, respectively. It is known that

$$\mu_x = \frac{-\phi(x)}{\Phi(x)}, \quad \sigma_x^2 = 1 - x\frac{\phi(x)}{\Phi(x)} - \left(\frac{\phi(x)}{\Phi(x)}\right)^2.$$

Because the memory strengths of the foils are independent, the mean and variance of the convolution of $n-1$ truncated distributions is μ_x and σ_x^2 each multiplied by $n-1$. By the central limit theorem, this convolution is approximately normal. Therefore,

$$\Pr\left(Y \geq c - x\right) \approx 1 - \Phi\left(c - x; (n-1)\mu_x, \sqrt{n-1}\sigma_x\right)$$

and

$$P_{TID}(c) \approx \int_{-\infty}^{\infty} \left(1 - \Phi\left(c - x; (n-1)\mu_x, \sqrt{(n-1)}\sigma_x\right)\right) \phi\left(x; \mu_t, \sigma_t\right) \Phi\left(x\right)^{n-1} dx.$$

Similarly, if there is a designated innocent suspect

$$P_{FA}\left(c\right) = 1 - \Phi\left(c; \mu_s, \sqrt{n-1+\sigma_s^2}\right)$$

$$P_{SID}\left(c\right) \approx \int_{-\infty}^{\infty} \left(1 - \Phi\left(c - x; (n-1)\mu_x, \sqrt{(n-1)}\sigma_x\right)\right) \phi\left(x; \mu_s, \sigma_s\right) \Phi\left(x\right)^{n-1} dx.$$

If there is no designated innocent suspect then

$$P_{FA}\left(c\right) = 1 - \Phi\left(c; 0, \sqrt{n}\right)$$

$$P_{SID}\left(c\right) = \frac{1}{n}\left(1 - \Phi\left(c; 0, \sqrt{n}\right)\right).$$

SDT-SEQ

The SDT-SEQ model is a sequential lineup model that implements the following decision rule. Let $K = \left\{i \in \left[n\right] : x_i \geq c_1\right\}$. That is, *K* is the set of positions of those items with memory strengths large enough to be identified. Then, if *K* is not empty choose lineup item $m = \min\left(K\right)$ at confidence level, $c = \max\left(\left\{c \in C : c \leq x_m\right\}\right)$, otherwise reject the lineup.

We start by deriving an equation for $P_{TID}\left(c\right)$ and *mutatis mutandis* for $P_{SID}\left(c\right)$. If the first lineup item is the target then $P_{TID}\left(c\right) = 1 - T\left(c\right)$. If the second item is the target then $P_{TID}\left(c\right) = \left(1 - T\left(c\right)\right)F\left(c_1\right)$, and so on. In general, if the target is at position *t* then the probability of target identification is equal to the joint probability of identifying the target and

not identifying an item at any preceding position. Let p_i be the probability that item *i* is the target. Then

$$P_{TID}(c) = (1 - T(c)) \sum_{i=1}^{n} p_i F(c_1)^{i-1} \quad .$$

The equation for $P_{SID}(c)$ is directly analogous. In this case, let q_i be the probability that item *i* is the designated innocent suspect. Then

$$P_{SID}(c) = (1 - S(c)) \sum_{i=1}^{n} q_i F(c_1)^{i-1} \quad .$$

The equations for $P_{TD}(c)$ and $P_{FA}(c)$ are more complex. In a sequential lineup, target detection occurs whenever the witness identifies any item in a TP lineup, whether it is the target or not. Suppose item *t* is the target and let $P_{TD}(c)_t$ be the probability of target detection in the corresponding lineup. This is the sum of three probabilities; $P_{TD}(c)_{i<t}$, the probability of identifying a preceding item as the target, $P_{TD}(c)_{i=t}$ the probability of correctly identifying the target at position *t*, and $P_{TD}(c)_{i>t}$, the probability of identifying a following item as the target. Then

$$P_{TD}(c)_{i<t} = (1 - F(c)) \sum_{i=1}^{t-1} F(c_1)^{i-1}$$

$$P_{TD}(c)_{i=t} = (1 - T(c)) F(c_1)^{t-1}$$

$$P_{TD}(c)_{i>t} = T(c)(1 - F(c)) \sum_{i=t+1}^{n} F(c_1)^{i-2}$$

with $P_{TD}(c)_{i<1} = 0$ and $P_{TD}(c)_{i>n} = 0$. It follows that $P_{TD}(c)_t = P_{TD}(c)_{i<t} + P_{TD}(c)_{i=t} + P_{TD}(c)_{i>t}$ and

$$P_{TD}(c) = \sum_{i=1}^{n} p_i P_{TD}(c)_i \quad .$$

By analogy, for a TA lineup with the designated innocent suspect in position *s*,

$$P_{FA}(c)_{i<s} = (1 - F(c)) \sum_{i=1}^{s-1} F(c_1)^{i-1}$$
$$P_{FA}(c)_{i=s} = (1 - S(c)) F(c_1)^{s-1}$$
$$P_{FA}(c)_{i>s} = S(c)(1 - F(c)) \sum_{i=s+1}^{n} F(c_1)^{i-2}.$$

Hence, $P_{FA}(c)_s = P_{FA}(c)_{i<s} + P_{FA}(c)_{i=s} + P_{FA}(c)_{i>s}$ and, if there is a designated suspect,

$$P_{FA}(c) = \sum_{i=1}^{n} q_i P_{FA}(c)_i \ ,$$

otherwise,

$$P_{FA}(c) = (1 - F(c)) \sum_{i=1}^{n} F(c_1)^{i-1}.$$

Normal distribution implementation. To implement the above as normal distributions, the following substitutions are made:

$$T(c) = \Phi(c; \mu_t, \sigma_t)$$
$$S(c) = \Phi(c; \mu_s, \sigma_s)$$
$$F(c) = \Phi(c).$$

Appendix B – Model Simulations and Cross Fits

We first generated random parameter values for d_t, s, and five criteria (c), matching the seven-parameter model used in subsequent fits to our experimental data. Parameters were sampled from uniform distributions with ranges taken from fits to the Palmer and Brewer (2012) database for d_t [0.3 – 2.6] and c [0.4 – 2.2] and from Wixted et al. (2018) and Wilson et al. (2019) and the subsequent fits to our experimental data for s [0.6 – 1.2]. We generated 100 sets of parameter values, which we used to simulate 100 datasets according to each model. The likelihood functions for generating predicted data according to each model are available in Appendix A. The models generate predicted proportions, not frequencies, but frequencies are required to fit the models to the data. In order to avoid any issues with low cell counts, we assumed 10000 TP and 10000 TA lineups for each dataset, multiplying the predicted TP/TA decision rates from the model by these amounts. Each model was then fit to the 100 datasets it generated, in addition to the 100 datasets generated by the other models.

Results

Table 1 and Table 2 below show the correlations between the parameter values of d_t and s for all cross fits. In general, correlations are good when the models are fit to their own data, although SDT-SEQ, and to some extent SDT-INT, have problems with outliers when fit to their own data, likely caused by the model falling in to local minima in the parameter space. When these outliers are excluded, recovery improves substantially, as evident on the main diagonal of Figures 2 and 3. Figure 1 shows mean χ^2 for each model fit to its own data and that of the other models. It is evident that the models tend to fit their own data better than the data generated by the other models. Figures 2 - 8 show scatterplots of the generating and recovered parameter values for d_t, s and $c_1 - c_5$ respectively. Criterion recovery is generally good when the models are fit to their own data, with the exception of some outliers for SDT-SEQ and SDT-INT when fit to their own data.

Table 1

Correlation between generating and recovered d_t for each cross fit

		Fitting Model		
		SDT-SEQ	SDT-MAX	SDT-INT
Generating Model	SDT-SEQ	.53	.31	.56
	SDT-MAX	.47	1.00	.98
	SDT-INT	.97	.99	.99

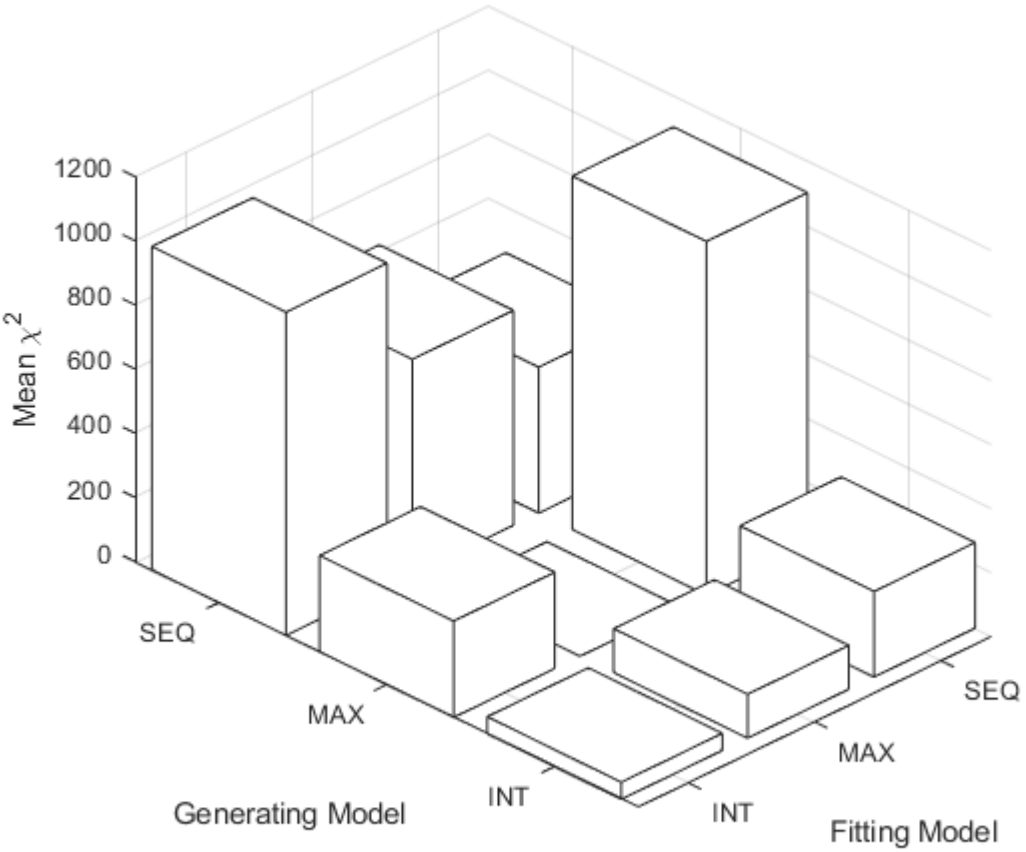
Table 2

Correlation between generating and recovered s for each cross fit

		Fitting Model		
		SDT-SEQ	SDT-MAX	SDT-INT
Generating Model	SDT-SEQ	.53	.23	.21
	SDT-MAX	.60	1.00	.38
	SDT-INT	.29	.59	.67

Figure 1

Mean χ^2 values for each model fit to its own simulated data and cross fit to the data generated by the other models.



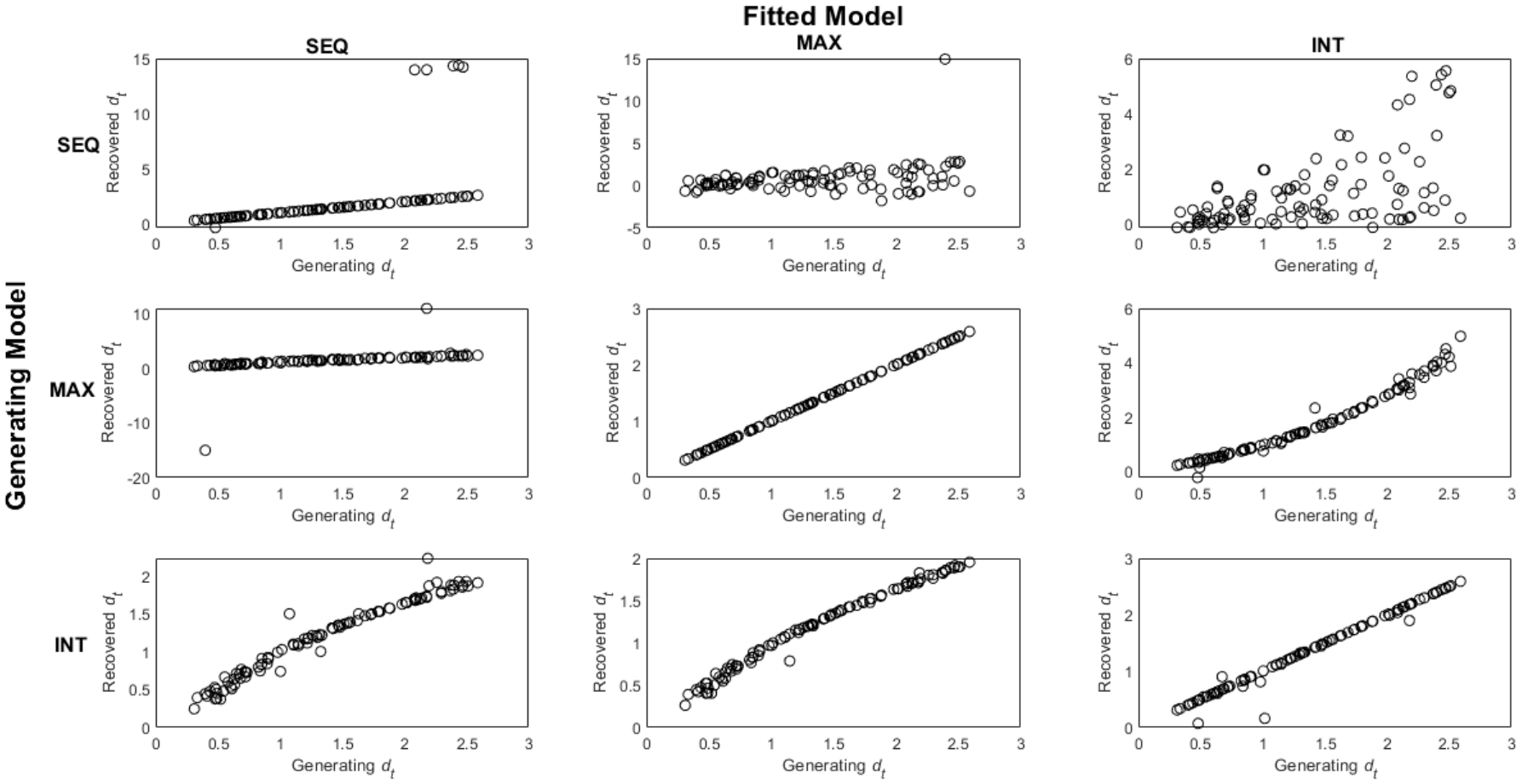


Figure 2. Scatter plots for d_t cross fits

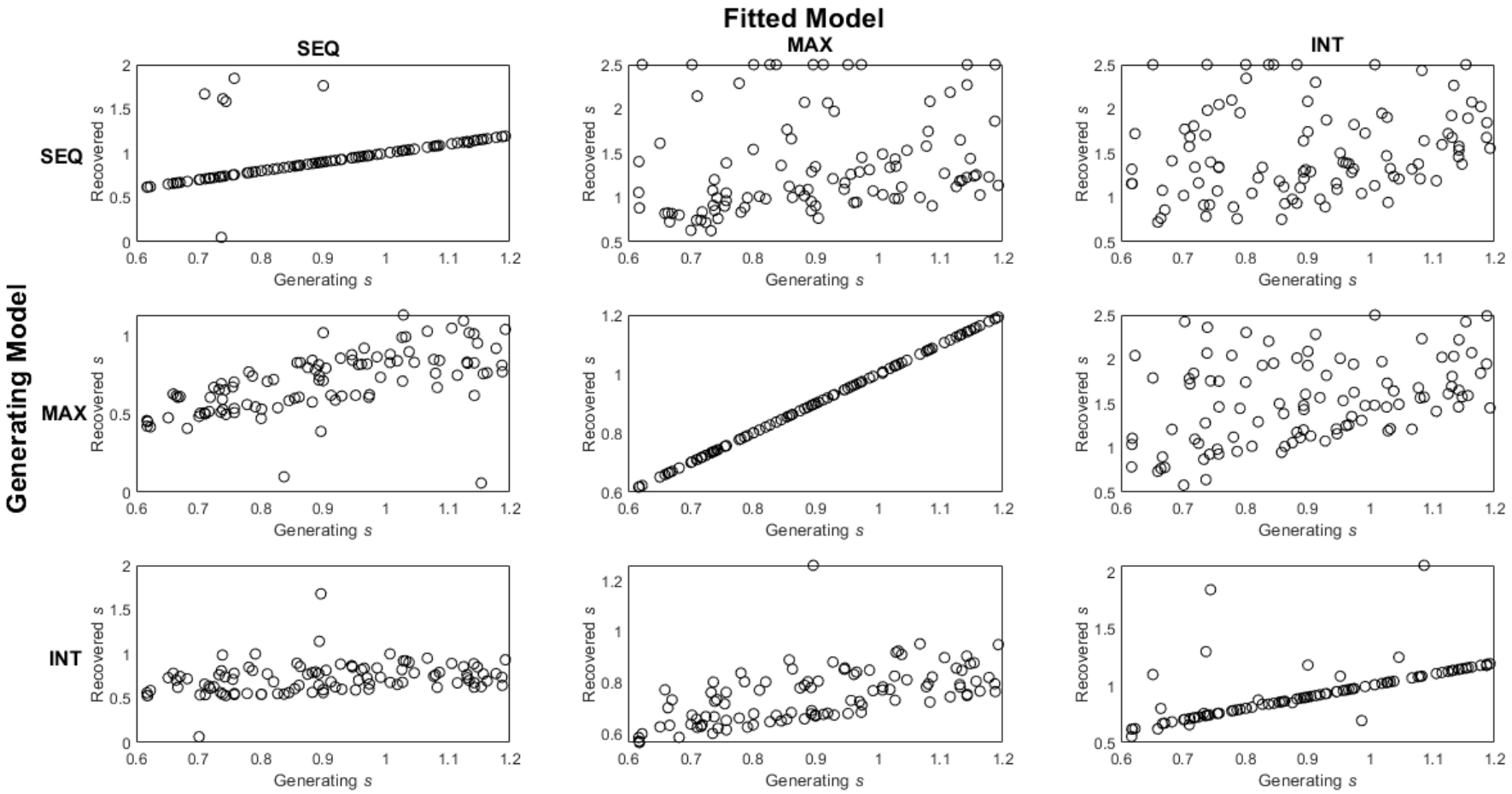


Figure 3. Scatter plots for *s* cross fit

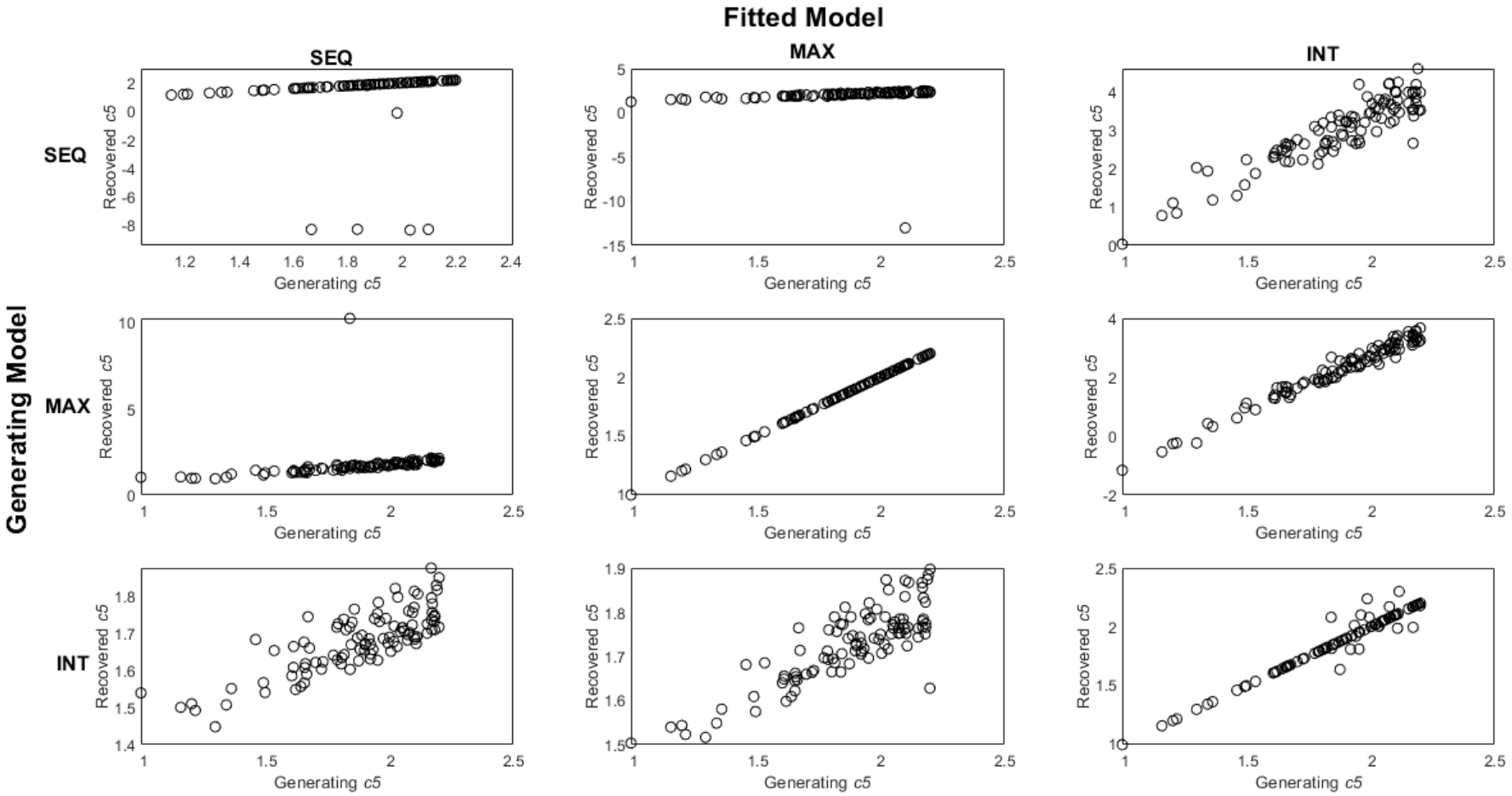


Figure 4. Scatter plots for *c5* cross fit

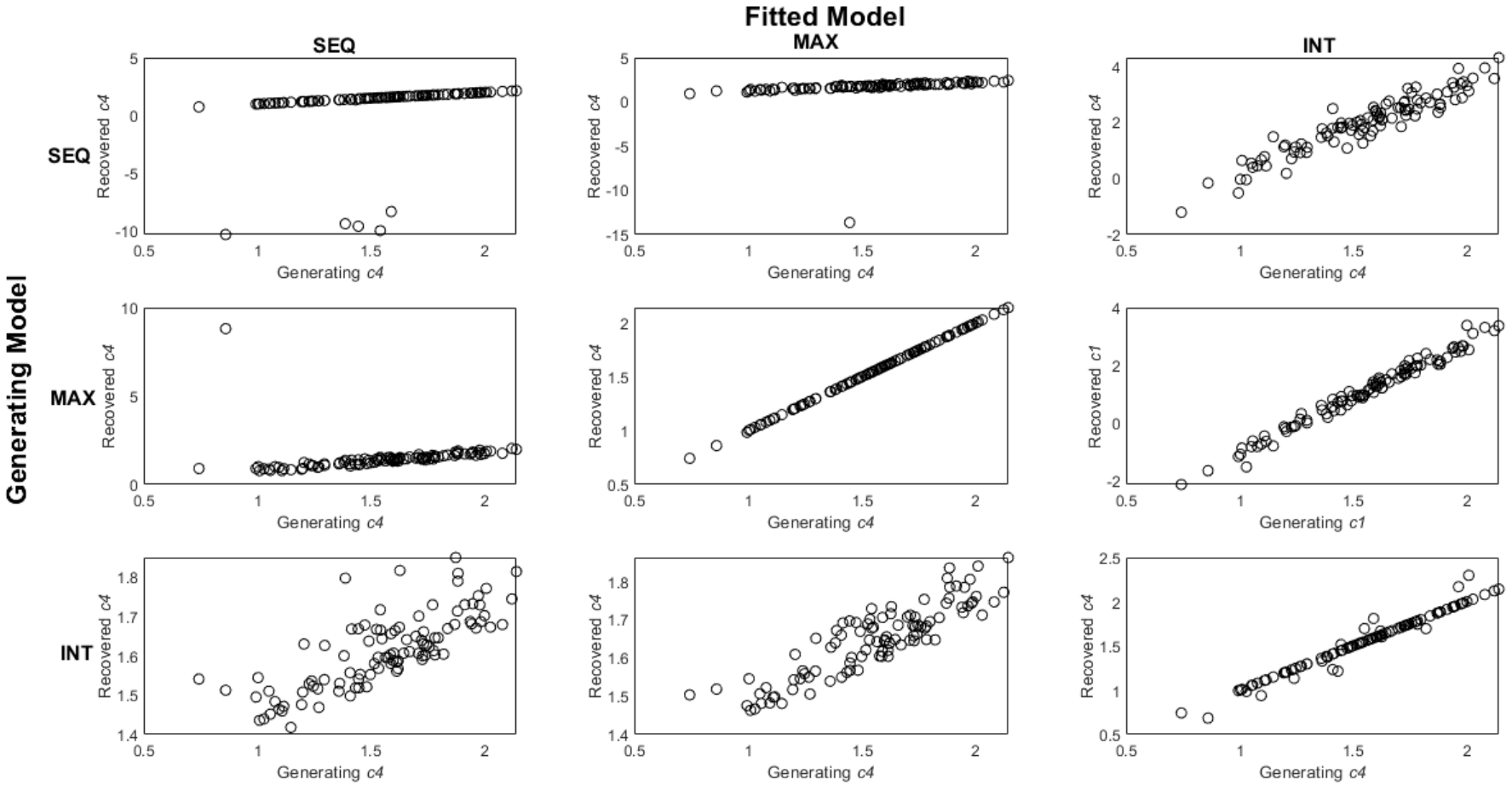


Figure 5. Scatterplots for *c4* cross fit

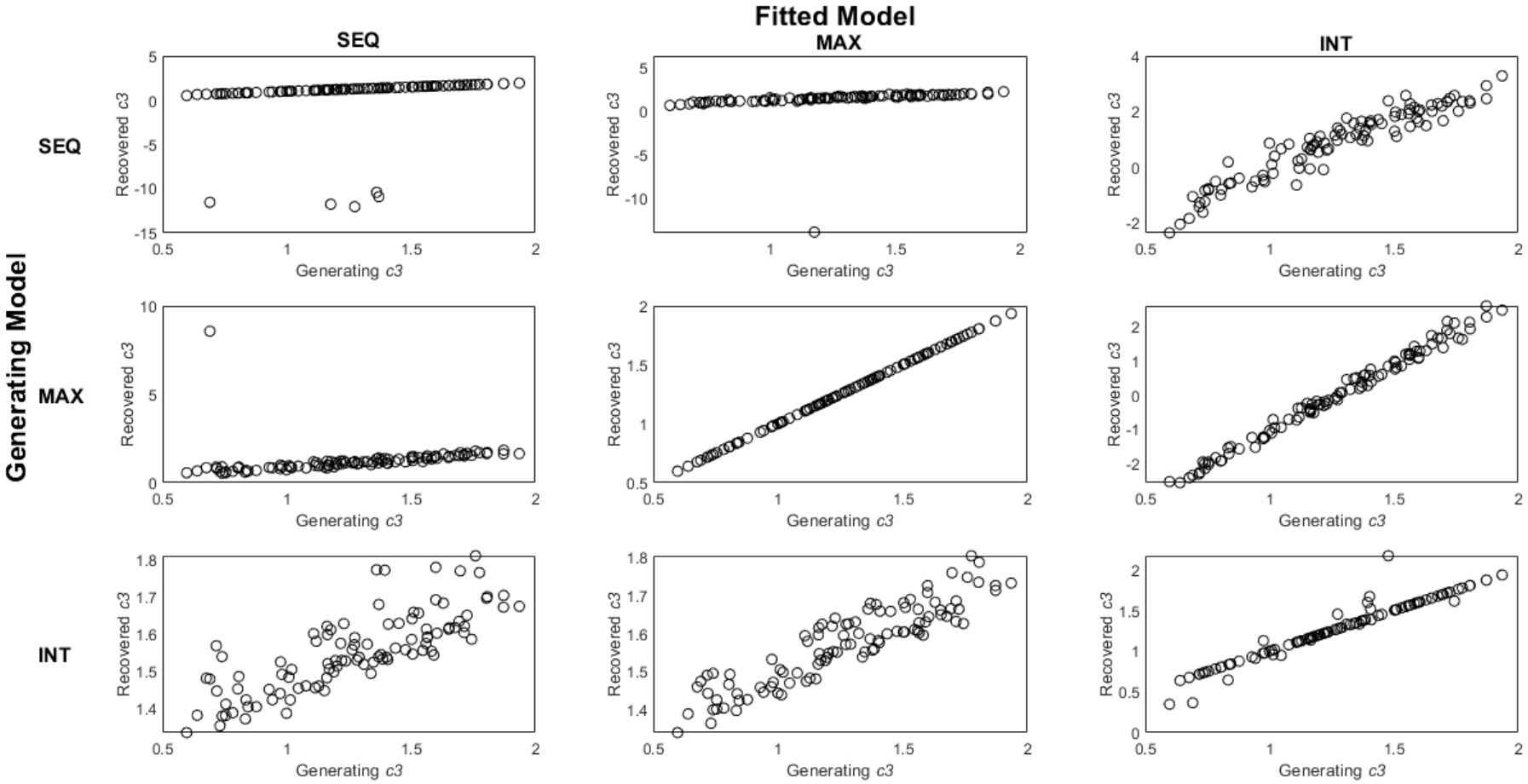


Figure 6. Scatterplots for *c3* cross fit

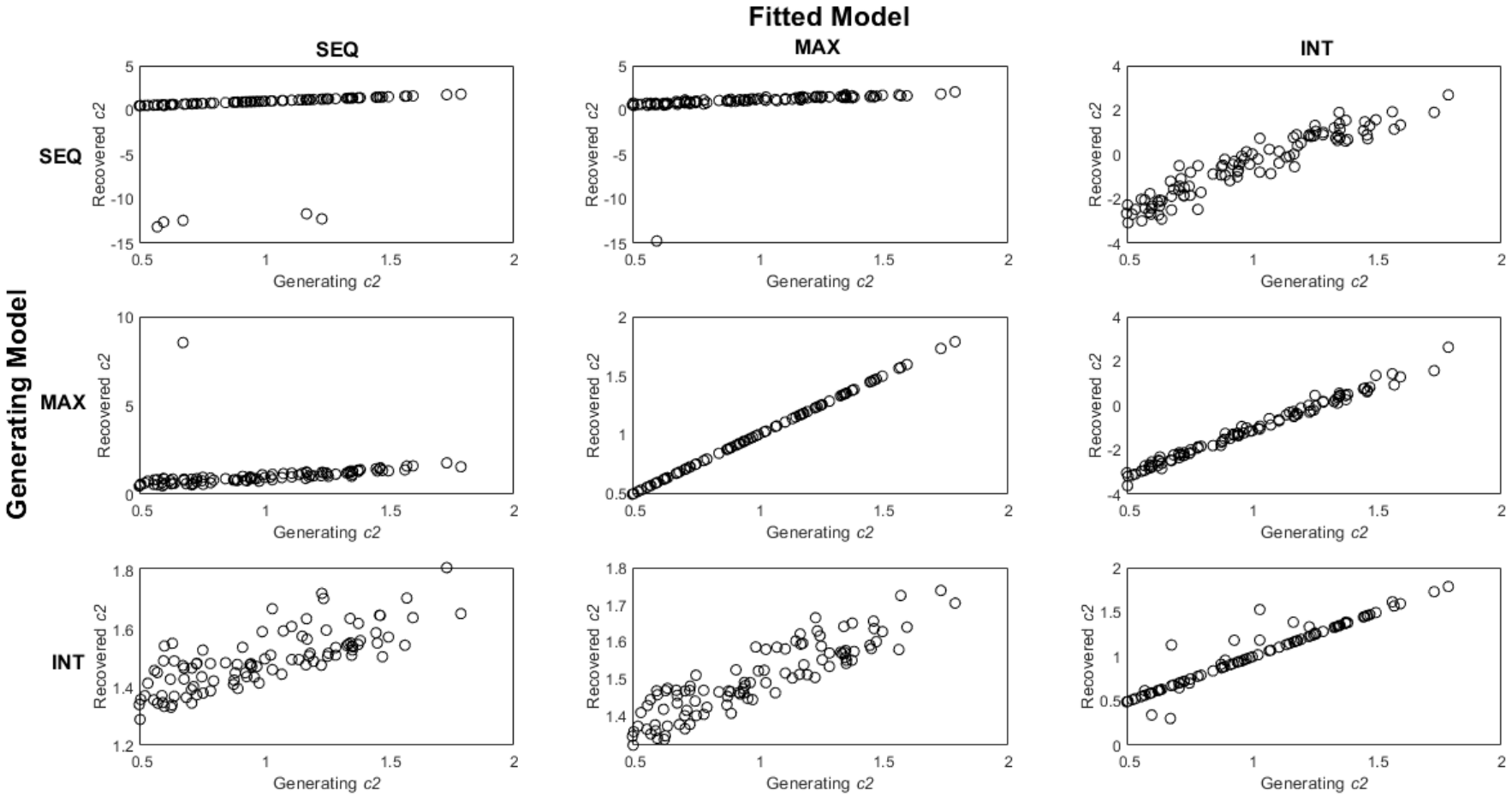


Figure 7. Scatterplots for *c2* cross fit

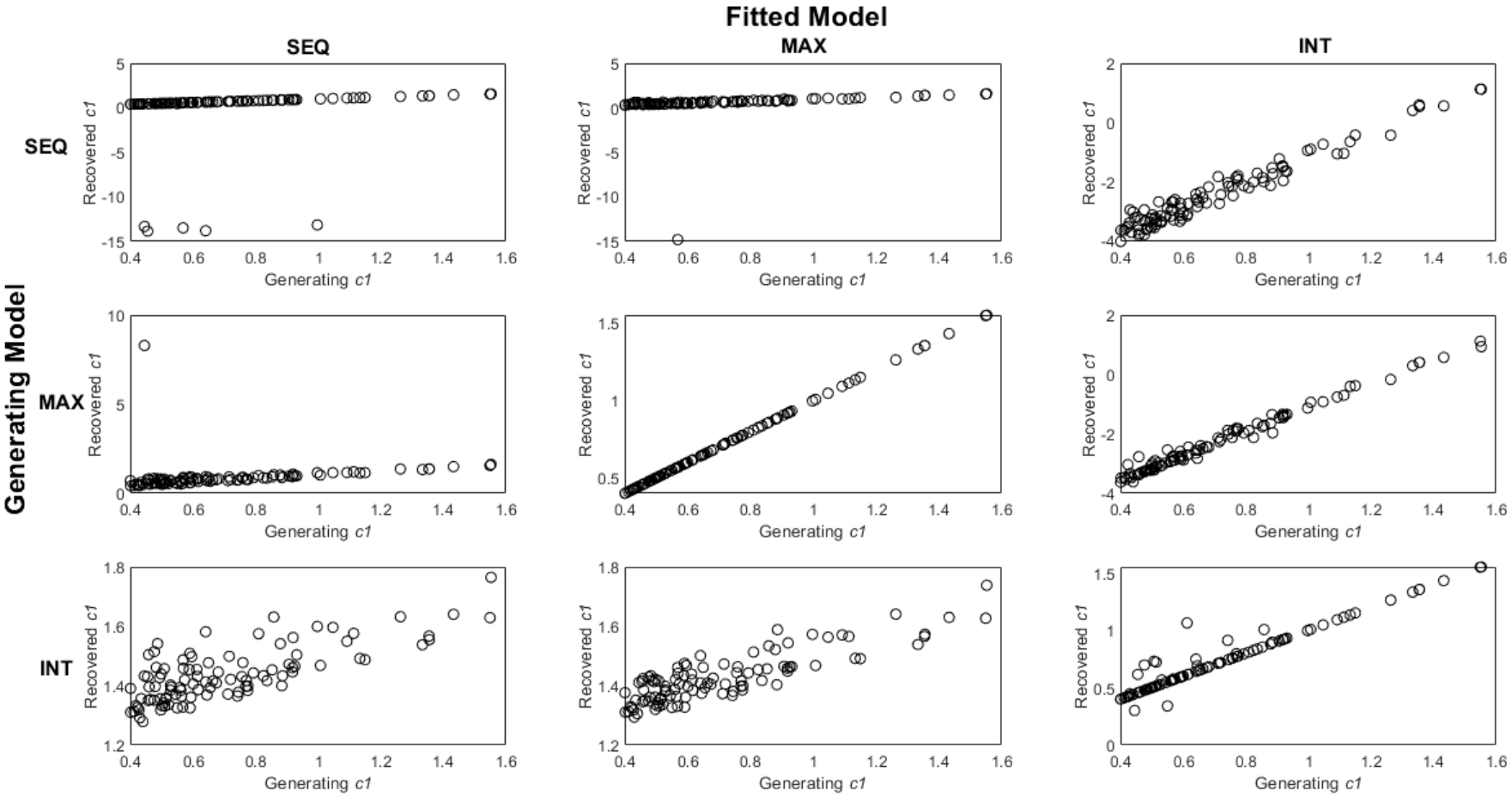


Figure 8. Scatterplots for *c1* cross fit

Appendix C – Full Tables of Parameter Estimates for Study One

Table 1

d_t, d_s, c and C for each dataset in the Palmer and Brewer (2012) corpus, estimated by the relevant models for each presentation format

	d_t				d_s				c				C			
	Simultaneous		Sequential		Simultaneous		Sequential		Simultaneous		Sequential		Simultaneous		Sequential	
	MAX	INT	SEQ	INT	MAX	INT	SEQ	INT	MAX	INT	SEQ	INT	MAX	INT	SEQ	INT
Carlson et. al (2008, Exp 1)	2.39	4.27	2.28	3.90	-.27	0.55	.43	1.89	1.73	1.86	2.08	2.54	.53	-0.28	.95	0.59
Carlson et. al (2008, Exp 2)	1.50	1.80	1.54	1.99	1.28	1.88	1.32	2.70	1.40	-0.22	1.77	0.95	.65	-1.12	1.00	-0.04
Clark & Davey (2005, Exp 1)	.49	0.80	1.48	1.50	.90	1.46	1.13	1.18	.79	-1.30	1.05	-1.18	.55	-1.70	.31	-1.93
Clark & Davey (2005, Exp 2)	1.08	1.40	1.64	1.70	.80	1.33	1.18	1.74	.93	-1.10	1.29	-0.74	.39	-1.80	.46	-1.59
Greathouse & Kovera (2009)	.94	1.18	2.08	2.00	.02	0.21	-.05	0.02	.75	-1.87	.93	-1.75	.28	-2.45	-.11	-2.75
Kneller et. al (2001)	1.57	1.61	1.83	1.40	-	-	-	-	.94	-1.20	1.29	0.20	.16	-2.00	.37	-0.50
Levi (2006)	1.80	1.99	1.34	1.64	-	-	-	-	1.23	-0.17	1.61	1.39	.33	-1.17	.94	0.57
Lindsay, Lea, & Fulford (1991)	1.69	2.43	2.10	3.89	.64	1.33	.51	2.13	1.33	-0.27	2.13	2.82	.49	-1.49	1.08	0.87
Lindsay & Wells (1985)	1.71	2.30	1.86	2.64	1.38	2.58	.97	2.11	1.40	0.10	1.86	1.43	.55	-1.05	.93	0.11
MacLin & Phelan (2007)	1.38	1.61	1.40	1.97	-	-	-	-	1.24	-0.02	2.12	3.22	.55	-0.82	1.41	2.23
MacLin et. al (2005, Exp 1)	1.06	1.14	1.17	1.44	-	-	-	-	.98	-1.00	1.50	1.03	.45	-1.57	.92	0.30
MacLin et. al (2005, Exp 2)	1.34	1.58	1.18	1.49	-	-	-	-	1.25	0.04	1.72	1.85	.58	-0.76	1.13	1.10
Melara et. al (1989)	.34	0.34	.56	0.70	-	-	-	-	.47	-3.09	1.62	1.56	.30	-3.26	1.34	1.21
Memon & Gabbert (2003)	1.32	1.46	1.19	1.70	-	-	-	-	1.14	-0.37	2.13	3.34	.48	-1.10	1.54	2.49
Parker & Ryan (1993)	1.14	1.77	.90	1.37	.29	0.96	-.21	0.25	1.25	0.20	1.16	-0.01	.68	-0.69	.71	-0.69
Pozzulo et. al (2008)	1.37	1.56	1.56	1.90	-	-	-	-	1.19	-0.19	1.70	1.62	.51	-0.97	.91	0.67
Pozzulo & Marciniak (2006)	1.35	1.59	1.21	0.99	-	-	-	-	1.30	0.25	1.05	-0.69	.63	-0.54	.45	-1.18
Rose et. al (2005)	1.78	1.94	1.28	1.39	-	-	-	-	1.19	-0.32	1.36	0.43	.30	-1.29	.72	-0.27
Sporer (1993)	1.11	1.25	1.34	1.64	-	-	-	-	.97	-1.03	1.50	0.99	.41	-1.66	.83	0.17

Steblay et. al (2011)	.30	0.32	.19	0.22	-	-	-	-	1.53	1.17	1.54	1.21	1.38	1.01	1.45	1.10
Wells & Pozzulo (2006)	.74	0.85	.19	0.16	-	-	-	-	1.26	0.18	1.22	-0.05	.89	-0.25	1.12	-0.13
Wilcock et. al (2005)	1.07	1.32	1.78	2.05	-	-	-	-	1.19	-0.09	1.61	1.16	.66	-0.75	.72	0.14

Table 2

Fit statistics and d_t, d_s and c for each dataset in the post-2011 corpus, estimated by SDT-MAX and SDT-SEQ for simultaneous and sequential presentation respectively

		$\chi^2(1)$		*p*		d_t		d_s		*c*	
Study	Condition	sim	seq	sim	seq	sim	seq	sim	seq	sim	seq
Pica & Pozzulo (2017)	-	.18	.90	.68	.34	2.37	2.22	-	-	1.31	1.80
Flowe et al. (2016)	Upright	2.60	1.14	.11	.29	1.64	1.38	-	-	1.33	1.45
	Inverted	.56	1.03	.46	.31	.77	.58	-	-	1.00	1.10
Carlson et al. (2016)	Backloading	10.29	.35	<.01	.56	1.05	1.32	-	-	1.16	1.28
	No backloading	.26	.92	.61	.17	1.07	1.17	-	-	.83	1.09
Pozzulo et al. (2016)	-	3.23	.94	.07	.33	1.95	1.28	-	-	1.28	1.47
Sucic et al. (2015)[a]	-	2.45	8.67	.12	<.01	1.24	1.43	1.09	1.16	1.02	1.39
Carlson & Carlson (2014)	No weapon, no feature	.84	.18	.36	.68	.89	.77	-.67	-.48	.82	.89
	No weapon, feature	.02	.01	.90	.91	.48	.20	-.15	-.49	.89	.73
	Weapon, no feature	1.11	3.19	.29	.07	1.21	.52	-.19	-.11	1.03	.83
	Weapon, feature	1.76	10.09	.18	<.01	.56	.76	-.11	-.51	.97	.80
Pozzulo et al. (2013)	-	.18	4.46	.67	.03	1.78	1.30	-	-	1.15	1.48
Mickes et al. (2012)	Experiment 1a	.23	.10	.63	.75	2.02	1.29	-	-	1.51	1.60

[a] Sucic et al. (2015) included an additional "don't know" response option. We excluded "don't know" responses when analysing the data, adjusting the total number of target present and target absent trials accordingly.

Appendix D – SDT-MAX R Code Walkthrough

First, download the R script "Supplement_SDT-MAX.R", which can be found at https://osf.io/jw64c/. This file contains functions for fitting an unequal variance version of SDT-MAX, also known as the Independent Observations model, to simultaneous lineup data. This document will explain what each piece of the code does and how it works together.

Data Structure

A critical aspect of working with this code is formatting observed data in a way that is appropriate for the likelihood functions that predict data according to the model. The experimental data from Kaesler et al. (2021) is formatted below:

```
#               Confidence Bin   100-91 90-81 80-66 65-51 50-0 Rejec
t
#                    Criterion     c5    c4    c3    c2    c1
# Target IDs (TID)                 24    25    30    9     11    NaN
# TP Foil IDs + TID (TD)           24    26    35    13    22    19
# Target Absent Foil IDs (FA)      4     11    25    16    24    61
```

The criteria are organised from most conservative (c_5) to least conservative (c_1); c_1 is the choose/no-choose threshold. The last column contains all rejections. There are no rejections associated with target identification decisions, so NaN appears in that cell. Incorrect target present (TP) rejections appear in the last cell of the TD row, which contains all detections on TP trials. Correct rejections appear in the last cell of the FA row, which contains all detections on TA trials. Summing the TD row should give the total number of TP trials, summing the TA row should give the total number of TA trials.

The code below sets up the observed data as a matrix and also specifies the lineup size, which is used by the likelihood functions.

```
obsData <- matrix(data = c(24,25,30,9,11,NaN,
                           24,26,35,13,22,19,
                           4,11,25,16,24,61),

                  nrow = 3,
                  ncol = 6,
                  byrow = TRUE)
```

```
n <- 6 #lineup size
```

Likelihood Functions

The next section of code contains the three functions that predict each outcome in our data, TID, TD and FA, according to the MAX decision rule. These functions take the model parameters and lineup size as input, and return probabilities as output. Derivations for the relevant equations can be found in Appendix A.

```
# Probability of Target ID (TID) on TP trials according to MAX model
QT <- function(c,d,s,n){
  m <- function(x) dnorm(x,mean = d, sd = s)*pnorm(x)^(n-1)
  p <- rep(0,length(c))
  for (i in 1:length(c)){
    a <- integrate(m,c[i],15)
    p[i] <- a$value
  }
  return(p)
}

# Probability of detection (TD) on TP trials MAX model
# This is any detection, both TID and TP foil ID
TP <- function(c,d,s,n){
  p <- rep(0,length(c))
  for (i in 1:length(c)){
    p[i] <- pnorm(((c[i])-d)/s)*pnorm(c[i])^(n-1)
  }
  p <- 1 - p
  return(p)
}

# Probability of TA foil ID (FA) MAX model
TA <- function(c,n){
  p <- rep(0,length(c))
  for (i in 1:length(c)){
    p[i] = pnorm(c[i])^n
  }
  p = 1 - p
  return(p)
}
```

These functions are called by a wrapper function **`genPred`**. It takes a vector of parameters, the observed data and the lineup size as input and returns predicted data as output in the same format as the observed data. The likelihood functions return probabilities, so

genPred converts these to counts by multiplying by the total number of observed TP/TA lineups. This is only reason for passing the observed data to **genPred**.

```
genPred <- function(pars, obsData, n){

  # Unpack vector of parameters for use by likelihood functions
  c <- pars[1:(length(pars)-2)]
  d <- pars[length(pars)-1] # second-to-last parameter is always d
  s <- tail(pars,1) # last parameter is always s

  # Calculate total number of TP and TA lineups
  totalTP <- sum(obsData[2,])
  totalTA <- sum(obsData[3,])

  TID <- QT(c(c, -15),d,s,n)
  TID <- c(TID[1],diff(TID))

  TD <- c(TP(c,d,s,n),1)
  TD <- c(TD[1],diff(TD))

  FA <- c(TA(c,n),1)
  FA <- c(FA[1],diff(FA))

  # Convert proportions to counts
  TID <- TID*totalTP
  TD <- TD*totalTP
  FA <- FA*totalTA

  predData <- rbind(TID,TD,FA)
  rownames(predData) <- c()

  return(predData)
}
```

Chi-Square Goodness-of-Fit Test

We use a χ^2 test to assess model fit. The **chiSq** function takes a vector of parameters, the observed data and the lineup size as input. It passes these to **genPred** to generate predicted data, which is compared to the observed data, returning a χ^2 value as output.

```
chiSq <- function(pars,obsData,n){

  predData <- genPred(pars,obsData,n)
  lastcell <- ncol(obsData)
  ncrit <- ncol(obsData)-1 #number of criteria
  f <- rep(0,nrow(obsData)+2) #for storing and summing chi-sq fit va
lue
```

```
  for (i in 1:ncrit){

    # TID
    a <- predData[1,i]
    b <- obsData[1,i]
    f[1] <- f[1] + (b-a)^2/a

    # Foil ID on TP lineup
    a <- predData[2,i]-predData[1,i]
    b <- obsData[2,i]-obsData[1,i]
    f[2] = f[2] + (b-a)^2/a

    # FA
    a <- predData[3,i]
    b <- obsData[3,i]
    f[3] <- f[3] + (b-a)^2/a
  }

  # Incorrect Rejection TP
  a <- predData[2,lastcell]
  b <- obsData[2,lastcell]
  f[4] <- (b-a)^2/a

  # Correct Rejection TA
  a <- predData[3,lastcell]
  b <- obsData[3,lastcell]
  f[5] <- (b-a)^2/a

  f <- sum(f)

  return(f)
}
```

We have now defined the functions necessary for fitting SDT-MAX to the data; **`QT`**, **`TP`** and **`TA`** are the likelihood functions that give the probability of TID, TD and FA respectively according to the MAX model, **`genPred`** is a wrapper function that calls these functions and returns predicted data in the same format as the observed data, and **`chiSq`** is the function that calls **`genPred`** and returns a goodness-of-fit value that reflects how close the predicted data is to the observed data.

Setting Starting Values for the Optimisation

We now turn to some setup for the optimisation step. One aspect of model fitting that can be difficult is setting appropriate starting values. If a parameter space is particular "lumpy", the best fitting parameter values recovered by optimisation may change depending on which starting values are used. This is because the optimisation routine may fall in to local minima near the starting values when searching the parameter space rather than finding the global minimum. Additionally, arbitrarily chosen starting values may sometimes lie outside the feasible region of the function being optimised, in which case the optimisation routine will not run.

We have attempted to mitigate these issues by setting random-like starting values for the criteria and plausible starting values for d_t and s_t, i.e. values that are close to those typically recovered in recognition memory experimenets. Note that this method for setting starting values may not work for every possible dataset, particularly those where participants behaved unusually. The function **`startVals`** takes the observed data and lineup size as input and returns a vector of starting values as output. The order of the parameters in this vector is important, as **`genpred`** unpacks this vector in the order $c_{max}, \ldots, c_1, d_t, s_t$ for passing to the likelihood functions.

```
startVals <- function(obsData,n) {

  ncrit <- ncol(obsData)-1 #number of criteria

  # Random-like criteria starting points
  c0 <- obsData[nrow(obsData),]
  c0[length(c0)] <- max(c(c0,0.5))
  a0 <- cumsum(c0)/sum(c0)
  c0 <-   qnorm((1-a0)^(1/n))
  c0 <- c0[1:length(c0)-1]

  # Concatenate to form starting parameter vector in order accepted
by genpred
  x0 = c(c0,1.5,1) #cMax, ..., c1, dt, s

  return(x0)
}
```

```
# Call function to generate starting values
x0 <- startVals(obsData,n)
x0

## [1] 2.590990 2.084184 1.606439 1.399098 1.124807 1.500000 1.00000
0
```

Inequality Constraints

We now set up the inequality constraints to ensure that the criteria do not "cross over" during optimsation. For example, unconstrained optimisation might converge on a solution where c_4 is greater than c_5, which is not interpretable.

In our case, we have a vector of parameters ordered $(c_5, c_4, c_3, c_2, c_1, d_t, s_t)$. We can consider this a vector labeled $(x_1, \ldots, x_7)$. We want to satisfy the inequality constraint that $x_1 \geq x_2 \geq x_3 \geq x_4 \geq x_5$. We have no inequality constrains on x_6 (d_t) or x_7 (s). Expressed as a series of linear equations, our inequality constraints are:

$$x_1 - x_2 \geq 0; x_2 - x_3 \geq 0; x_3 - x_4 \geq 0; x_4 - x_5 \geq 0;$$

The `constrOptim` function we use for the optimisation takes inequality constraints in the form $Ax \geq b$, where Ax is a $k * p$ matrix of k inequality constrains and p parameters, and b is a vector of the constraints to be satisified. We have four inequality constrains ($k = 4$) and seven parameters ($p = 7$), so Ax is a 4 x 7 matrix, and b is a vector of length four.

To represent our linear equations in matrix form, conforming to the requirement that $Ax \geq b$, we need to generate the following matrix for Ax:

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]    1   -1    0    0    0    0    0
## [2,]    0    1   -1    0    0    0    0
## [3,]    0    0    1   -1    0    0    0
## [4,]    0    0    0    1   -1    0    0
```

and the following vector for b:

```
## [1] 0 0 0 0
```

This can achieved for any number of criteria with the following code:

```
# Set inequality constraint matrix for criteria, stops them crossing
over
```

```
ncrit <- ncol(obsData)-1 # number of criteria
nparam <- length(x0) # number of parameters
A <- matrix(0,ncrit-1,nparam); # k x p matrix
b <- rep(0,ncrit-1) # vector of >= constraints to satisfy
for (i in 1:ncrit-1) {
  A[i,i] <- 1; A[i,i+1] <- -1
}
```

Optimisation

We now use the R function **constrOptim** to minimise the function **chiSq**. **constrOptim** takes the vector of starting values **theta = x0** and passes it as the first argument (**pars**) of the function to be minimised, **f = chiSq**. It takes the constraint matrix **ui = A** and the constraint vector **ci = b**. The last two input arguments to the function are the extra inputs required by **chiSq**; **obsData** and the lineup size **n**. **constrOptim** returns a list including the value of **chiSq** at the optimal solution and the best fitting parameters.

Consult the **optim** documentation for information about the remaining input arguments, which are options for the optimisation routine.

```
# Minimise chiSq function using the starting values and constraints
defined above
out <- constrOptim(theta = x0, f = chiSq, grad = NULL, ui = A, ci =
b, mu = 1e-04, method = "Nelder-Mead", outer.iterations = 100, obsDa
ta = obsData, n = n)
```

When we run the optimisation it starts with the parameters in **x0** and, at each iteration, it searches for a new set of parameter values that minimise χ^2. It usually stops when it converges on an optimal solution, but it may stop for other reasons without converging, such as reaching the maximum number of iterations specified in the options.

Output

Once the optimisation has finished we can examine the list stored in **out**. **out$par** contains the best fitting parameter values in the same order as specified in **x0**. **out$value** is the smallest χ^2 value at the optimal solution, associated with the parameters in **out$par**. **out$counts** shows how many times **chiSq** was called, i.e. how many iterations were

required to find the optimal solution. **`out$convergence`** shows whether the optimisation converged. Zero indicates that it did converge in this case. Consult the documentation for **`optim`** for information on other convergence messages that may indicate unsuccessful convergence.

```
out
```

```
## $par
## [1] 2.7214683 2.2022708 1.6910396 1.4885855 1.1593367 1.8250110 0
.9384431
##
## $value
## [1] 13.43876
##
## $counts
## function gradient
##     1004       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
##
## $outer.iterations
## [1] 3
##
## $barrier.value
## [1] 0.0002934577
```

Our degrees of freedom for calculating a *p*-value for our χ^2 test are based on the format of our data and our number of parameters. The degrees of freedom are calculated as the number rows in the dataset multiplied by the number of criteria, minus the number of parameters estimated. At least one degree of freedom is required to fit the model; the output of a saturated model with zero degrees of freedom is not interpretable. The code below uses **`pchisq`**, the χ^2 distribution function, to find *p* for our minimum χ^2 value with the appropriate degrees of freedom. The *p*-value is non-significant, so we can conclude that the model fit the data.

```
# Degrees of freedom and p-value
df <- (nrow(obsData)*ncrit) - nparam
pval <- pchisq(out$value, df = df, lower.tail = FALSE)
pval

## [1] 0.09761866
```

We can pass the best-fitting parameters in **`out$par`** to **`genpred`** to get the predicted data associated with the optimal solution. This can be used for plotting model ROCs. Additionally, if the model did not fit, it can be useful to compare the observed and predicted data to see where it failed.

Finally, we can extract the parameters from **`out$par`** and collate all information relevant to our results in to a list to make it easy to read at a glance.

```
# Predicted Data - Useful for plotting model ROCs
predData <- genPred(out$par, obsData, n)
predData[1,ncol(predData)] <- NaN #replace meaningless number in rej
ect cell of pred Data with NaN

# Collate in to list for easy viewing
results <- list(
  c = out$par[1:(length(out$par)-2)], # same as extracting pars in g
enpred
  d = out$par[length(out$par)-1],
  s = tail(out$par,1),
  x0 = x0,
  n = n,
  nparam = nparam,
  fitVal = out$value,
  df = df,
  pval = pval,
  obs = obsData,
  pred = predData,
  converge = out$convergence)

results

## $c
## [1] 2.721468 2.202271 1.691040 1.488586 1.159337
##
## $d
## [1] 1.825011
##
## $s
## [1] 0.9384431
```

```
##
## $x0
## [1] 2.590990 2.084184 1.606439 1.399098 1.124807 1.500000 1.00000
0
##
## $n
## [1] 6
##
## $nparam
## [1] 7
##
## $fitVal
## [1] 13.43876
##
## $df
## [1] 8
##
## $pval
## [1] 0.09761866
##
## $obs
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]   24   25   30    9   11  NaN
## [2,]   24   26   35   13   22   19
## [3,]    4   11   25   16   24   61
##
## $pred
##          [,1]      [,2]     [,3]      [,4]     [,5]     [,6]
## [1,] 23.45591 23.279170 25.80194  8.672849 10.35475      NaN
## [2,] 25.45453 28.471077 36.23977 13.703964 17.90724 17.22342
## [3,]  2.72695  8.570629 23.01633 14.454322 28.14878 64.08299
##
## $converge
## [1] 0
```

References

Abudarham, N., Shkiller, L., & Yovel, G. (2019). Critical features for face recognition. *Cognition, 182*, 73-83. https://doi.org/10.1016/j.cognition.2018.09.002

Andersen, S. M., Carlson, C. A., Carlson, M. A., & Gronlund, S. D. (2014). Individual differences predict eyewitness identification performance. *Personality and Individual Differences, 60*, 36-40. https://doi.org/10.1016/j.paid.2013.12.011

Behrman, B. W., & Davey, S. L. (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

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116*(1), 84-115. https://doi.org/10.1037/a0014351

Bentley, D. (2003). *English criminal justice in the 19th century* (1st ed.). Bloomsbury Publishing Plc.

Bornstein, B. H., & Penrod, S. D. (2008). Hugo who? G. F. Arnold's alternative early approach to psychology and law. *Applied Cognitive Psychology, 22*(6), 759-768. https://doi.org/10.1002/acp.1480

Brewer, N. (2006). Uses and abuses of eyewitness identification confidence. *Legal and Criminological Psychology, 11*(1), 3-23. https://doi.org/10.1348/135532505X79672

Brewer, N., & Doyle, J. (2021). Changing the face of police lineups: Delivering more information from witnesses. *Journal of Applied Research in Memory and Cognition*. https://doi.org/10.1016/j.jarmac.2020.12.004

Brewer, N., Weber, N., & Guerin, N. (2020). Police lineups of the future? *American Psychologist, 75*(1), 76-91. https://doi.org/10.1037/amp0000465

Brewer, N., Weber, N., & Semmler, C. (2007). A role for theory in eyewitness identification research. In R. Lindsay, D. F. Ross, J. Read, & M. P. Toglia (Eds.), *The handook of eyewitness psychology volume ii: Memory for people* (pp. 201-218). Erlbaum.

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *J Exp Psychol Appl*, *12*(1), 11-30. https://doi.org/10.1037/1076-898X.12.1.11

British Broadcasting Corporation. (2003). *Police offer virtual id parades*. Retrieved May 19 from http://news.bbc.co.uk/2/hi/technology/2850803.stm

Brown, E., Deffenbacher, K., & Sturgill, W. (1977). Memory for faces and the circumstances of encounter. *Journal of Applied Psychology*, *62*(3), 311-318. https://doi.org/10.1037/0021-9010.62.3.311

Carlson, C. A., & Carlson, M. A. (2014). An evaluation of lineup presentation, weapon presence, and a distinctive feature using roc analysis. *Journal of Applied Research in Memory and Cognition*, *3*(2), 45-53. https://doi.org/10.1016/j.jarmac.2014.03.004

Carlson, C. A., Carlson, M. A., Weatherford, D. R., Tucker, A., & Bednarz, J. (2016). The effect of backloading instructions on eyewitness identification from simultaneous and sequential lineups. *Applied Cognitive Psychology*, *30*(6), 1005-1013. https://doi.org/10.1002/acp.3292

Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied*, *14*(2), 118-128. https://doi.org/10.1037/1076-898X.14.2.118

Carlson, C. A., Jones, A. R., Whittington, J. E., Lockamyeir, R. F., Carlson, M. A., & Wooten, A. R. (2019). Lineup fairness: Propitious heterogeneity and the diagnostic feature-detection hypothesis. *Cognitive Research: Principles and Implications*, *4*(1), 20. https://doi.org/10.1186/s41235-019-0172-5

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press. https://doi.org/10.1093/0195171276.001.0001

Clark, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology*, *17*. 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. (2012a). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science*, *7*(3), 238-259. https://doi.org/10.1177/1745691612439584

Clark, S. E. (2012b). 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., Benjamin, A. S., Wixted, J. T., Mickes, L., & Gronlund, S. D. (2015a). Eyewitness identification and the accuracy of the criminal justice system. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 175-186. https://doi.org/10.1177/2372732215602267

Clark, S. E., & Davey, S. L. (2005). The target-to-foils shift in simultaneous and sequential lineups. *Law and Human Behavior*, *29*(2), 151-172. https://doi.org/10.1007/s10979-005-2418-7

Clark, S. E., Erickson, M. A., & 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

Clark, S. E., Howell, R. T., & Davey, S. L. (2008). Regularities in eyewitness identification. *Law and Human Behavior, 32*(3), 187-218. https://doi.org/10.1007/s10979-006-9082-4

Clark, S. E., Moreland, M. B., & Gronlund, S. D. (2014). Evolution of the empirical and theoretical foundations of eyewitness identification reform. *Psychonomic Bulletin and Review, 21*(2), 251-267. https://doi.org/10.3758/s13423-013-0516-y

Clark, S. E., Moreland, M. B., & Rush, R. A. (2015b). Lineup composition and lineup fairness. In T. Valentine & J. P. Davis (Eds.), *Forensic facial identification: Theory and practice of identification from eyewitnesses, composites and cctv* (pp. 129-157). Wiley-Blackwell. https://doi.org/10.1002/9781118469538.ch6

Cohen, A. L., Starns, J. J., & Rotello, C. M. (2020). Sdtlu: An r package for the signal detection analysis of eyewitness lineup data. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01402-7

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed. ed.). L. Erlbaum Associates.

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science, 27*(9), 1227-1239. https://doi.org/10.1177/0956797616655789

Colloff, M. F., Wade, K. A., Strange, D., & Wixted, J. T. (2018). Filler-siphoning theory does not predict the effect of lineup fairness on the ability to discriminate innocent from guilty suspects: Reply to smith, wells, smalarz, and lampinen (2018). *Psychological Science, 29*(9), 1552-1557. https://doi.org/10.1177/0956797618786459

Colloff, M. F., Wade, K. A., Wixted, J. T., & Maylor, E. A. (2017). A signal-detection analysis of eyewitness identification across the adult lifespan. *Psychology and Aging, 32*(3), 243-258. https://doi.org/10.1037/pag0000168

Colloff, M. F., Wilson, B. M., Seale-Carlisle, T. M., & Wixted, J. T. (2021). Optimizing the selection of fillers in police lineups. *Proceedings of the National Academy of Sciences, 118*(8), e2017292118. https://doi.org/10.1073/pnas.2017292118

Colloff, M. F., & Wixted, J. T. (2019). Why are lineups better than showups? A test of the filler siphoning and enhanced discriminability accounts. *Journal of Experimental Psychology: Applied*, *26*(1). https://doi.org/10.1037/xap0000218

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of memory and language*, *64*(4), 316-326. https://doi.org/10.1016/j.jml.2011.02.003

Cross, J. F., Cross, J., & Daly, J. (1971). Sex, race, age, and beauty as factors in recognition of faces. *Perception & Psychophysics*, *10*(6), 393-396. https://doi.org/10.3758/BF03210319

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology*, *54*(3), 304. https://doi.org/10.1016/j.jmp.2010.01.001

DeCarlo, L. T. (2012). On a signal detection approach to m-alternative forced choice with bias, with maximum likelihood and bayesian approaches to estimation. *Journal of Mathematical Psychology*, *56*(3), 196-207. https://doi.org/10.1016/j.jmp.2012.02.004

Dewar, M. T., Cowan, N., & Sala, S. D. (2007). Forgetting due to retroactive interference: A fusion of müller and pilzecker's (1900) early insights into everyday forgetting and recent research on anterograde amnesia. *Cortex*, *43*(5), 616-634. https://doi.org/10.1016/S0010-9452(08)70492-1

Dobolyi, D. G., & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification

overconfidence. *Journal of Experimental Psychology: Applied*, *19*(4), 345-357. https://doi.org/10.1037/a0034596

Duncan, M. (2006). *A signal detection model of compound decision tasks (techincal note drdc tr 2006-256)*. https://apps.dtic.mil/sti/pdfs/ADA473015.pdf

Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review*, *111*(2), 524-542. https://doi.org/10.1037/0033-295X.111.2.524

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology*, *67*(5), 818-835. https://doi.org/10.1037/0022-3514.67.5.818

Ebbesen, E. B., & Flowe, H. D. (2002). *Simultaneous v. Sequential lineups: What do we really know?* https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/20167/1/Simultaneous%20v.%20Sequential%20Lineups%20What%20Do%20We%20Really%20Know.pdf

Ebbinghaus, H. (1885/1965). *Memory : A contribution to experimental psychology*. Dover Publications.

Efron, B. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Egan, J. P. (1958). *Recognition memory and the operating characteristic* (USAF Operational Applications Laboratory Technical Note, Issue.

Epps, D. (2015). The consequences of error in criminal justice. *Harvard Law Review*, *128*(4), 1065-1151.

Feldman, J. (2021). Information-theoretic signal detection theory. *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000300

Fife, D., Perry, C., & Gronlund, S. D. (2014). Revisiting absolute and relative judgments in the witness model. *Psychonomic Bulletin & Review*, *21*(2), 479-487. https://doi.org/10.3758/s13423-013-0493-1

Fitzgerald, R. J., Rubinova, E., & Juncu, S. (2021). Eyewitness identification around the world. In A. M. Smith, M. P. Toglia, & J. M. Lampinen (Eds.), *Methods, measures, and theories in eyewitness identification tasks*. Routledge. https://doi.org/10.4324/9781003138105

Flowe, H. D., & Ebbesen, E. B. (2007). The effect of lineup member similarity on recognition accuracy in simultaneous and sequential lineups. *Law and Human Behavior*, *31*(1), 33-52. https://doi.org/10.1007/s10979-006-9045-9

Flowe, H. D., Smith, H. M., Karoglu, N., Onwuegbusi, T. O., & Rai, L. (2016). Configural and component processing in simultaneous and sequential lineup procedures. *Memory*, *24*(3), 306-314. https://doi.org/10.1080/09658211.2015.1004350

Friedman, R. D. (1986). A close look at probative value. *Boston University law review*, *66*(3-4), 733.

Garrett, B. (2012). *Convicting the innocent: Where criminal prosecutions go wrong*. Harvard University Press. https://doi.org/10.4159/harvard.9780674060982

Gibson, E. (1969). *Principles of perceptual learning and development*. Appleton-Century-Crofts.

Goodenough, D. J., Rossmann, K., & Lusted, L. B. (1972). Radiographic applications of signal detection theory. *Radiology*, *105*(1), 199-200. https://doi.org/10.1148/105.1.199

Greathouse, S. M., & Kovera, M. B. (2009). Instruction bias and lineup presentation moderate the effects of administrator knowledge on eyewitness identification. *Law and Human Behavior*, *33*(1), 70-82. https://doi.org/10.1007/s10979-008-9136-x

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

Gronlund, S. D., & Benjamin, A. S. (2018). The new science of eyewitness memory. In K. D. Federmeier (Ed.), *Psychology of learning and motivation* (Vol. 69, pp. 241-284). Academic Press. https://doi.org/10.1016/bs.plm.2018.09.006

Gronlund, S. D., Carlson, C. A., Dailey, S. B., & Goodsell, C. A. (2009). Robustness of the sequential lineup advantage. *Journal of Experimental Psychology: Applied*, *15*(2), 140-152. https://doi.org/10.1037/a0015082

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S. A., Wooten, A., & Graham, M. (2012). Showups versus lineups: An evaluation using roc analysis. *Journal of Applied Research in Memory and Cognition*, *1*(4), 221-228. https://doi.org/10.1016/j.jarmac.2012.09.003

Gronlund, S. D., Mickes, L., Wixted, J. T., & Clark, S. E. (2015). Conducting an eyewitness lineup: How the research got it wrong. In *Psychology of learning and motivation - advances in research and theory* (Vol. 63, pp. 1-43). https://doi.org/10.1016/bs.plm.2015.03.003

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

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science*, *23*(1), 3-10. https://doi.org/10.1177/0963721413498891

Haeffel, G. J., & Howard, G. S. (2010). Self-report: Psychology's four-letter word. *The American journal of psychology*, *123*(2), 181-188. https://doi.org/10.5406/amerjpsyc.123.2.0181

Horry, R., Brewer, N., Weber, N., & Palmer, M. A. (2015). The effects of allowing a second sequential lineup lap on choosing and probative value. *Psychology, Public Policy, and Law, 21*(2), 121-133. https://doi.org/10.1037/law0000041

Horry, R., Memon, A., Wright, D. B., & Milne, R. (2012a). Predictors of eyewitness identification decisions from video lineups in england: A field study. *Law and Human Behavior, 36*(4), 257-265. https://doi.org/10.1037/h0093959

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

Hutchinson, T. P. (1981). A review of some unusual applications of signal detection theory. *Quality & quantity*, *15*(1), 71-98. https://doi.org/10.1007/BF00144302

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*(2), 291-306. https://doi.org/10.1037/a0015525

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

Kaesler, M. P., Semmler, C., & Dunn, J. C. (2017). Using measurement models to understand eyewitness identification. 39th Annual Meeting of the Cognitive Science Society, London, UK. https://cogsci.mindmodeling.org/2017/papers/0126/paper0126.pdf

Kang, L., & Tian, L. (2013). Estimation of the volume under the roc surface with three ordinal diagnostic categories. *Computational Statistics & Data Analysis*, *62*, 39-51. https://doi.org/10.1016/j.csda.2013.01.004

Karras, T., Laine, S., & Aila, T. (2019, 15-20 June 2019). A style-based generator architecture for generative adversarial networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),

Kellen, D., Christoph Klauer, K., & Singmann, H. (2012). On the measurement of criterion noise in signal detection theory: The case of recognition memory. *Psychological Review, 119*(3), 457-479. https://doi.org/10.1037/a0027727

Kellen, D., Davis-Stober, C. P., Dunn, J. C., & Kalish, M. L. (2021a). The problem of coordination and the pursuit of structural constraints in psychology. *Perspectives on Psychological Science*, 1745691620974771-1745691620974771. https://doi.org/10.1177/1745691620974771

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. (2021b). Testing the foundations of signal detection theory in recognition memory. *Psychological Review*. https://doi.org/10.1037/rev0000288

Key, K. N., Cash, D. K., Neuschatz, J. S., Price, J., Wetmore, S. A., & Gronlund, S. D. (2015). Age differences (or lack thereof) in discriminability for lineups and showups. *Psychology, Crime and Law*, *21*(9), 871-889. https://doi.org/10.1080/1068316X.2015.1054387

Kneller, W., Memon, A., & Stevenage, S. (2001). Simultaneous and sequential lineups: Decision processes of accurate and inaccurate eyewitnesses. *Applied Cognitive Psychology*, *15*(6), 659-671. https://doi.org/10.1002/acp.739

Kuha, J. (2004). Aic and bic: Comparisons of assumptions and performance. *Sociological Methods & Research*, *33*(2), 188-229. https://doi.org/10.1177/0049124103262065

Lampinen, J. M. (2016). Roc analyses in eyewitness identification research. *Journal of Applied Research in Memory and Cognition*, *5*(1), 21-33. https://doi.org/10.1016/j.jarmac.2015.08.006

Lampinen, J. M., Smith, A. M., & Wells, G. L. (2019). Four utilities in eyewitness identification practice: Dissociations between receiver operating characteristic (roc) analysis and expected utility analysis. *Law and Human Behavior*, *43*(1), 26-44. https://doi.org/10.1037/lhb0000309

Lane, S. M., & Meissner, C. A. (2008). A 'middle road' approach to bridging the basic–applied divide in eyewitness identification research. *Applied Cognitive Psychology*, *22*(6), 779-787. https://doi.org/10.1002/acp.1482

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

Leippe, M. R., Wells, G. L., & Ostrom, T. M. (1978). Crime seriousness as a determinant of accuracy in eyewitness identification. *Journal of Applied Psychology*, *63*(3), 345-351. https://doi.org/10.1037/0021-9010.63.3.345

Levi, A. M. (2006). An analysis of multiple choices in msl lineups, and a comparison with simultaneous and sequential ones. *Psychology, Crime & Law*, *12*(3), 273-285. https://doi.org/10.1080/10683160500238782

Levi, A. M. (2007). Research note: Evidence for moving to an 84-person photo lineup. *Journal of Experimental Criminology*, *3*(4), 377-391. https://doi.org/10.1007/s11292-007-9042-0

Levi, A. M. (2012). Much better than the sequential lineup: A 120-person lineup. *Psychology, Crime & Law*, *18*(7), 631-640. https://doi.org/10.1080/1068316X.2010.526120

Lewin, C., & Herlitz, A. (2002). Sex differences in face recognition—women's faces make the difference. *Brain and cognition*, *50*(1), 121-128. https://doi.org/10.1016/S0278-2626(02)00016-7

Lindner, M. A., Eitel, A., Thoma, G. B., Dalehefte, I. M., Ihme, J. M., & Köller, O. (2014). Tracking the decision-making process in multiple-choice assessment: Evidence from eye movements [Article]. *Applied Cognitive Psychology*, *28*(5), 738-752. https://doi.org/10.1002/acp.3060

Lindsay, R., Lea, J. A., Nosworthy, G. J., Fulford, J. A., Hector, J., LeVan, V., & Seabrook, C. (1991a). Biased lineups: Sequential presentation reduces the problem. *Journal of Applied Psychology*, *76*(6), 796-802. https://doi.org/10.1037/0021-9010.76.6.796

Lindsay, R., & 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

Lindsay, R. C., Lea, J. A., & Fulford, J. A. (1991b). Sequential lineup presentation: Technique matters. *Journal of Applied Psychology*, *76*(5), 741-745. https://doi.org/10.1037/0021-9010.76.5.741

Lindsay, R. C. L., & Pozzulo, J. D. (1999). Sources of eyewitness identification error. *International Journal of Law and Psychiatry*, *22*(3), 347-360. https://doi.org/10.1016/S0160-2527(99)00014-X

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

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of verbal learning and verbal behavior*, *13*(5), 585-589. https://doi.org/10.1016/s0022-5371(74)80011-3

Lucas, C. A., Brewer, N., & Palmer, M. A. (2020). Eyewitness identification: The complex issue of suspect-filler similarity. *Psychology, Public Policy, and Law*. https://doi.org/10.1037/law0000243

Luus, C. A. E., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behavior*, *15*(1), 43-57. https://doi.org/10.1007/BF01044829

MacLin, O. H., & Phelan, C. M. (2007). Pc_eyewitness: Evaluating the new jersey method. *Behavior Research Methods*, *39*(2), 242-247. https://doi.org/10.3758/BF03193154

Maclin, O. H., Zimmerman, L. A., & Malpass, R. S. (2005). Pc_eyewitness and the sequential superiority effect: Computer-based lineup administration. *Law and Human Behavior*, *29*(3), 303-321. https://doi.org/10.1007/s10979-005-3319-5

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum Associates Publishers; US. https://doi.org/10.4324/9781410611147

Malmberg, K. J., Raaijmakers, J. G. W., & Shiffrin, R. M. (2019). 50 years of research sparked by atkinson and shiffrin (1968). *Memory & Cognition*, *47*(4), 561-574. https://doi.org/10.3758/s13421-019-00896-7

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

Mansour, J., Beaudry, J., & Lindsay, R. (2017). Are multiple-trial experiments appropriate for eyewitness identification studies? Accuracy, choosing, and confidence across

trials. *Behavior Research Methods*, *49*(6), 2235-2254. https://doi.org/10.3758/s13428-017-0855-0

McAdoo, R. M., & Gronlund, S. D. (2016). Relative judgment theory and the mediation of facial recognition: Implications for theories of eyewitness identification. *Cognitive Research: Principles and Implications*, *1*(1), Article 11. https://doi.org/10.1186/s41235-016-0014-7

McCarty, D. (1929). *Psychology for the lawyer*. Prentice-Hall.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in cognitive science*, *1*(1), 11-38. https://doi.org/10.1111/j.1756-8765.2008.01003.x

McCormick, K., Semmler, C., & Dunn, J. C. (2019, June 5 - 9). *Is eyewitness memory continuous or all-or-none?* [Poster presentation]. Meeting of the Society for Applied Research in Memory and Cognition, Cape Cod, MA. http://hdl.handle.net/2440/122735

McQuiston-Surrett, D., Malpass, R. S., & Tredoux, C. G. (2006). Sequential vs. Simultaneous lineups: A review of methods, data, and theory. *Psychology, Public Policy, and Law*, *12*(2), 137-169. https://doi.org/10.1037/1076-8971.12.2.137

Mecklenburg, S. H., Bailey, P. J., & Larson, M. R. (2008). The illinois field study: A significant contribution to understanding real world eyewitness identification issues. *Law and Human Behavior*, *32*(1), 22-27. https://doi.org/10.1007/s10979-007-9108-6

Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition*, *33*(5), 783-792. https://doi.org/10.3758/bf03193074

Meisters, J., Diedenhofen, B., & Musch, J. (2018). Eyewitness identification in simultaneous and sequential lineups: An investigation of position effects using receiver operating characteristics. *Memory*, 1-13. https://doi.org/10.1080/09658211.2018.1464581

Melara, R. D., & De Witt-Rickards, T. S. (1989). Enhancing lineup identification accuracy: Two codes are better than one. *Journal of Applied Psychology, 74*(5), 706. https://doi.org/10.1037/0021-9010.74.5.706

Memon, A., & Gabbert, F. (2003). Unravelling the effects of sequential presentation in culprit-present lineups. *Applied Cognitive Psychology*, *17*(6), 703-714. https://doi.org/10.1002/acp.909

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., Moreland, M. B., Clark, S. E., & Wixted, J. T. (2014). Missing the information needed to perform roc analysis? Then compute d', not the diagnosticity ratio. *Journal of Applied Research in Memory and Cognition*, *3*(2), 58-62. https://doi.org/10.1016/j.jarmac.2014.04.007

Mickes, L., Seale-Carlisle, T. M., Wetmore, S. A., Gronlund, S. D., Clark, S. E., Carlson, C. A., Goodsell, C. A., Weatherford, D., & Wixted, J. T. (2017). Rocs in eyewitness identification: Instructions versus confidence ratings. *Applied Cognitive Psychology*, *31*(5), 467-477. https://doi.org/10.1002/acp.3344

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review*, *14*(5), 858-865. https://doi.org/10.3758/BF03194112

Mitchell, D. J., Tal, E., & Chang, H. (2017). The making of measurement: Editors' introduction. *Studies in History and Philosophy of Science Part A, 65-66,* 1-7. https://doi.org/10.1016/j.shpsa.2017.10.001

Moreland, M. B., & Clark, S. E. (2016). Eyewitness identification: Research, reform, and reversal. *Journal of Applied Research in Memory and Cognition*, *5*(3), 277-283. https://doi.org/10.1016/j.jarmac.2016.07.011

Mu, E., Chung, T. R., & Reed, L. I. (2017). Paradigm shift in criminal police lineups: Eyewitness identification as multicriteria decision making. *International Journal of Production Economics*, *184*, 95-106. https://doi.org/10.1016/j.ijpe.2016.11.019

Munsterberg, H. (1908). *On the witness stand*. McClure.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90-100. https://doi.org/10.1016/s0022-2496(02)00028-7

National Research Council. (2014). *Identifying the culprit: Assessing eyewitness identification*. The National Academies Press. https://doi.org/10.17226/18891

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*(1), 47-84. https://doi.org/10.1016/j.cogpsych.2003.11.001

Navon, D. (1992). Selection of lineup foils by similarity to the suspect is likely to misfire. *Law and Human Behavior*, *16*(5), 575-593. https://doi.org/10.1007/BF01044624

Neuschatz, J. S., Wetmore, S. A., Key, K. N., Cash, D. K., Gronlund, S. D., & Goodsell, C. A. (2016). A comprehensive evaluation of showups. In M. K. Miller & B. H. Bornstein (Eds.), *Advances in psychology and law: Volume 1* (pp. 43-69). Springer International Publishing. https://doi.org/10.1007/978-3-319-29406-3_2

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231-259. https://doi.org/10.1037/0033-295X.84.3.231

Oberauer, K. (2003). Understanding serial position curves in short-term recognition and recall. *Journal of memory and language*, *49*(4), 469-483. https://doi.org/10.1016/S0749-596X(03)00080-9

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

Palmer, M. A., Brewer, N., & Weber, N. (2010). Postidentification feedback affects subsequent eyewitness identification performance. *Journal of Experimental Psychology: Applied*, *16*(4), 387-398. https://doi.org/10.1037/a0021034

Palmer, M. A., Brewer, N., & Weber, N. (2012). The information gained from witnesses' responses to an initial "blank" lineup. *Law and Human Behavior*, *36*(5), 439-447. https://doi.org/10.1037/h0093939

Palmer, M. A., Sauer, J. D., & Holt, G. A. (2017). Undermining position effects in choices from arrays, with implications for police lineups. *J Exp Psychol Appl*, *23*(1), 71-84. https://doi.org/10.1037/xap0000109

Parker, J. F., & Ryan, V. (1993). An attempt to reduce guessing behavior in children's and adults' eyewitness identifications. *Law and Human Behavior*, *17*, 11-26. https://doi.org/10.1007/BF01044534

Pica, E., & Pozzulo, J. (2017). The elimination-plus lineup: Testing a modified lineup procedure with confidence. *Journal of Investigative Psychology and Offender Profiling*, *14*(3), 294-306. https://doi.org/10.1002/jip.1477

Pike, G., Brace, N., & Kynan, S. (2002). *The visual identification of suspects: Procedures and practice. Briefing note 2/02*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.608.8076&rep=rep1&type=pdf

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in cognitive sciences*, *6*(10), 421-425. https://doi.org/10.1016/S1364-6613(02)01964-2

Police and Criminal Evidence Act. (1984). *Code of practice for the identification of persons by police officers*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/903812/pace-code-d-2017.pdf

Police Executive Research Forum. (2013). *A national survey of eyewitness identification procedures in law enforcement agencies*. https://www.ncjrs.gov/pdffiles1/nij/grants/242617.pdf

Pollack, I., & Hsieh, R. (1969). Sampling variability of the area under the roc-curve and of d′e. *Psychological Bulletin, 71*. https://doi.org/10.1037/h0026862

Pozzulo, J. D., Dempsey, J., Corey, S., Girardi, A., Lawandi, A., & Aston, C. (2008). Can a lineup procedure designed for child witnesses work for adults? Comparing simultaneous, sequential, and elimination lineup procedures. *Journal of Applied Social Psychology*, *38*(9), 2195-2209. https://doi.org/10.1111/j.1559-1816.2008.00387.x

Pozzulo, J. D., Dempsey, J., & Pettalia, J. (2013). The z generation: Examining perpetrator descriptions and lineup identification procedures. *Journal of Police and Criminal Psychology*, *28*(1), 63-74. https://doi.org/10.1007/s11896-012-9107-5

Pozzulo, J. D., & Marciniak, S. (2006). Comparing identification procedures when the perpetrator has changed appearance. *Psychology, Crime & Law*, *12*(4), 429-438. https://doi.org/10.1080/10683160500050690

Pozzulo, J. D., Reed, J., Pettalia, J., & Dempsey, J. (2016). Simultaneous, sequential, elimination, and wildcard: A comparison of lineup procedures. *Journal of Police and Criminal Psychology*, *31*(1), 71-80. https://doi.org/10.1007/s11896-015-9168-3

Quinlivan, D. S., Wells, G. L., Neuschatz, J. S., Luecht, K. M., Cash, D. K., & Key, K. N. (2017). The effects of pre-admonition suggestions on eyewitnesses' choosing rates and retrospective identification judgments. *Journal of Police and Criminal Psychology*, *32*(3), 236-246. https://doi.org/10.1007/s11896-016-9216-7

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). Proc: An open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Robinson, E. (1935). *Law and the lawyers*. Macmillan.

Rose, R. A., Bull, R., & Vrij, A. (2005). Non-biased lineup instructions do matter--a problem for older witnesses. *Psychology, Crime & Law*, *11*(2), 147-159. https://doi.org/10.1080/10683160512331316307

Rotello, C. M., & Chen, T. (2016). Roc curve analyses of eyewitness identification decisions: An analysis of the recent debate. *Cognitive Research: Principles and Implications*, *1*(1), 10. https://doi.org/10.1186/s41235-016-0006-7

Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, *22*. https://doi.org/10.3758/s13423-014-0759-2

Rouder, J. N., & Lu, J. (2005). An introduction to bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573-604. https://doi.org/10.3758/BF03196750

Schultheis, H., & Singhaniya, A. (2015). Decision criteria for model comparison using the parametric bootstrap cross-fitting method. *Cognitive Systems Research*, *33*(Supplement C), 100-121. https://doi.org/10.1016/j.cogsys.2014.09.003

Seale-Carlisle, T. M., & Mickes, L. (2016). Us line-ups outperform uk line-ups. *Royal Society Open Science*, *3*(9), 160300. https://doi.org/10.1098/rsos.160300

Seale-Carlisle, T. M., Wetmore, S. A., Flowe, H. D., & Mickes, L. (2019). Designing police lineups to maximize memory performance. *Journal of Experimental Psychology: Applied*. https://doi.org/10.1037/xap0000222

Semmler, C., Brewer, N., & Douglass, A. B. (2012). Jurors believe eyewitnesses. In B. L. Cutler (Ed.), *Conviction of the innocent: Lessons from psychological research* (pp. 185-209). American Psychological Association. https://doi.org/10.1037/13085-009

Smalarz, L., Kornell, N., Vaughn, K. E., & Palmer, M. A. (2019). Identification performance from multiple lineups: Should eyewitnesses who pick fillers be burned? *Journal of Applied Research in Memory and Cognition*, *8*(2), 221-232. https://doi.org/10.1016/j.jarmac.2019.03.001

Smith, A. M., Lampinen, J. M., Wells, G. L., Smalarz, L., & Mackovichova, S. (2019). Deviation from perfect performance measures the diagnostic utility of eyewitness lineups but partial area under the roc curve does not. *Journal of Applied Research in Memory and Cognition*, *8*(1), 50-59. https://doi.org/10.1016/j.jarmac.2018.09.003

Smith, A. M., Wells, G. L., Lindsay, R., & Penrod, S. D. (2016). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior*, No Pagination Specified. https://doi.org/10.1037/lhb0000219

Smith, A. M., Wells, G. L., Smalarz, L., & Lampinen, J. M. (2018). Increasing the similarity of lineup fillers to the suspect improves the applied value of lineups without improving memory performance: Commentary on colloff, wade, and strange (2016). *Psychological Science*, *29*(9), 1548-1551. https://doi.org/10.1177/0956797617698528

Smith, A. M., Yang, Y., & Wells, G. L. (2020). Distinguishing between investigator discriminability and eyewitness discriminability: A method for creating full receiver

operating characteristic curves of lineup identification performance. *Perspectives on Psychological Science*, *15*(3), 589-607. https://doi.org/10.1177/1745691620902426

Sosic-Vasic, Z., Hille, K., Kröner, J., Spitzer, M., & Kornmeier, J. (2018). When learning disturbs memory – temporal profile of retroactive interference of learning on memory formation. *Frontiers in Psychology*, *9*(82). https://doi.org/10.3389/fpsyg.2018.00082

Spanton, R. W., & Berry, C. J. (2020). The unequal variance signal-detection model of recognition memory: Investigating the encoding variability hypothesis. *Quarterly Journal of Experimental Psychology*, *73*(8), 1242-1260. https://doi.org/10.1177/1747021820906117

Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology*, *78*(1), 22-33. https://doi.org/10.1037/0021-9010.78.1.22

Steblay, N. K., Dietrich, H. L., Ryan, S. L., Raczynski, J. L., & James, K. A. (2011a). Sequential lineup laps and eyewitness accuracy. *Law and Human Behavior*, *35*(4), 262-274. https://doi.org/10.1007/s10979-010-9236-2

Steblay, N. K., Dysart, J., Fulero, S., & Lindsay, R. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, *25*(5), 459-473. https://doi.org/10.1023/A:1012888715007

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011b). 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

Steblay, N. K., & Phillips, J. D. (2011). The not-sure response option in sequential lineup practice. *Applied Cognitive Psychology*, *25*(5), 768-774. https://doi.org/10.1002/acp.1755

Stephens, R. G., Dunn, J. C., & Hayes, B. K. (2019). Belief bias is response bias: Evidence from a two-step signal detection model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(2), 320-332. https://doi.org/10.1037/xlm0000587

Stern, W. (1903). *Beiträge zur psychologie der aussage*. J. A. Barth.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of experimental psychology. Learning, memory, and cognition*, *24*(6), 1397-1410. https://doi.org/10.1037/0278-7393.24.6.1397

Sučić, I., Tokić, D., & Ivešić, M. (2015). Field study of response accuracy and decision confidence with regard to lineup composition and lineup presentation. *Psychology, Crime and Law*, *21*(8), 798-819. https://doi.org/10.1080/1068316X.2015.1054383

Swets, J. A. (1959). Indices of signal detectability obtained with various psychophysical procedures. *Journal of the Acoustical Society of America*, *31*, 511-513. https://doi.org/10.1121/1.1907744

Turtle, J., Read, J. D., Lindsay, D. S., & Brimacombe, C. A. E. (2008). Toward a more informative psychological science of eyewitness evidence. *Applied Cognitive Psychology*, *22*(6), 769-778. https://doi.org/10.1002/acp.1481

U.S. National Department of Justice. (1999). *Eyewitness evidence: A guide for law enforcement*. https://www.ncjrs.gov/pdffiles1/nij/178240.pdf

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q J Exp Psychol A*, *43*(2), 161-204. https://doi.org/10.1080/14640749108400966

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

Wais, P. E., Mickes, L., & Wixted, J. T. (2008). Remember/know judgments probe degrees of recollection. *Journal of cognitive neuroscience*, *20*(3), 400-405. https://doi.org/10.1162/jocn.2008.20041

Weber, N., & Varga, M. (2012). Can a modified lineup procedure improve the usefulness of confidence? *Journal of Applied Research in Memory and Cognition*, *1*(3), 152-157. https://doi.org/10.1016/j.jarmac.2012.06.007

Wells, E. C., & Pozzulo, J. D. (2006). Accuracy of eyewitnesses with a two-culprit crime: Testing a new identification procedure. *Psychology, Crime & Law*, *12*(4), 417-427. https://doi.org/10.1080/10683160500050666

Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology*, *36*(12), 1546. https://doi.org/10.1037/0022-3514.36.12.1546

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? *The American psychologist*, *48*(5), 553-571. https://doi.org/10.1037/0003-066X.48.5.553

Wells, G. L. (2008). Theory, logic and data: Paths to a more coherent eyewitness science. *Applied Cognitive Psychology*, *22*(6), 853-859. https://doi.org/10.1002/acp.1488

Wells, G. L. (2014). Eyewitness identification: Probative value, criterion shifts, and policy regarding the sequential lineup. *Current Directions in Psychological Science*, *23*(1), 11-16. https://doi.org/10.1177/0963721413504781

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior*, *44*(1), 3-36. https://doi.org/10.1037/lhb0000359

Wells, G. L., & Lindsay, R. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88*. https://doi.org/10.1037/0033-2909.88.3.776

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest, 7*(2), 45-75. https://doi.org/10.1111/j.1529-1006.2006.00027.x

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22*(6), 603-647. https://doi.org/10.1023/A:1025750605807

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2011). *A test of the simultaneous vs . Sequential lineup methods an initial report of the ajs national eyewitness identification field studies*. http://www.popcenter.org/library/reading/PDFs/lineupmethods.pdf

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

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2015a). Double-blind photo lineups using actual eyewitnesses: An experimental test of a sequential versus simultaneous lineup procedure. *Law and Human Behavior, 39*(1), 1-14. https://doi.org/10.1037/lhb0000096

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25*(9), 1115-1125. https://doi.org/10.1177/01461672992512005

Wells, G. L., Yang, Y., & Smalarz, L. (2015b). Eyewitness identification: Bayesian information gain, base-rate effect equivalency curves, and reasonable suspicion. *Law and Human Behavior*, *39*(2), 99-122. https://doi.org/10.1037/lhb0000125

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

Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition*, *4*. https://doi.org/10.1016/j.jarmac.2014.07.003

Whipple, G. M. (1917). Psychology of testimony. *Psychological Bulletin*, *14*(7), 234-236. https://doi.org/10.1037/h0074099

Wickelgren, W. A. (1966). Consolidation and retroactive interference in short-term recognition memory for pitch. *Journal of Experimental Psychology*, *72*(2), 250-259. https://doi.org/10.1037/h0023438

Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press. http://ebookcentral.proquest.com/lib/adelaide/detail.action?docID=271367

Wilcock, R. A., Bull, R., & Vrij, A. (2005). Aiding the performance of older eyewitnesses: Enhanced non-biased line-up instructions and line-up presentation. *Psychiatry, Psychology and Law*, *12*(1), 129-140. https://doi.org/10.1375/pplt.2005.12.1.129

Wills, W. (1838/1850). *An essay on the principles of circumstancial evidence* (3rd ed.). Henry Butterworth.

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*, 108-125. https://doi.org/10.1016/j.jml.2018.10.002

Wilson, B. M., Seale-Carlisle, T. M., & Mickes, L. (2018). The effects of verbal descriptions on performance in lineups and showups. *Journal of experimental psychology. General*, *147*(1), 113-124. https://doi.org/10.1037/xge0000354

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152-176. https://doi.org/10.1037/0033-295X.114.1.152

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. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science*, *7*(3), 275-278. https://doi.org/10.1177/1745691612442906

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. (2015a). Evaluating eyewitness identification procedures: Roc analysis and its misconceptions. *Journal of Applied Research in Memory and Cognition*, *4*(4), 318-323. https://doi.org/10.1016/j.jarmac.2015.08.009

Wixted, J. T., & Mickes, L. (2015b). Roc analysis measures objective discriminability for any eyewitness identification procedure. *Journal of Applied Research in Memory and Cognition*, *4*(4), 329-334. https://doi.org/10.1016/j.jarmac.2015.08.007

Wixted, J. T., & Mickes, L. (2018). Theoretical vs. Empirical discriminability: The application of roc methods to eyewitness identification. *Cognitive Research: Principles and Implications*, *3*(1), 9. https://doi.org/10.1186/s41235-018-0093-8

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *113*(2), 304-309. https://doi.org/10.1073/pnas.1516814112

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

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

Wright, D. B., & Sladden, B. (2003). An own gender bias and the importance of hair in face recognition. *Acta psychologica*, *114*(1), 101-114. https://doi.org/10.1016/S0001-6918(03)00052-0

Wundt, W. M., & Judd, C. H. (1897). *Outlines of psychology* (2nd rev. English ed., from the 4th rev. German ed. ed.). W. Engelman.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of experimental psychology. Learning, memory, and cognition*, *20*(6), 1341-1354. https://doi.org/10.1037/0278-7393.20.6.1341