Entscheidungen von Augenzeug:innen in Gegenüberstellungen: Validierung und Anwendung eines multinomialen Modells zur Erfassung der zugrundeliegenden Prozesse

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Nicola Marie Menne
aus Geldern

Düsseldorf, Februar 2024

aus dem Institut für Experimentelle Psychologie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Axel Buchner

2. Prof. Dr. Raoul Bell

Tag der mündlichen Prüfung: 16.07.2024

Inhaltsverzeichnis

Zusammenfassung

Die Tatsache, dass viele Justizirrtümer auf Fehlurteile durch Augenzeug:innen zurückzuführen sind, unterstreicht die Notwendigkeit, ein besseres Verständnis über die Prozesse zu erlangen, die den Entscheidungen von Augenzeug:innen zugrunde liegen. In der vorliegenden Dissertation wird ein neues multinomiales Verarbeitungsbaummodell vorgestellt, das es ermöglicht, unter Berücksichtigung aller Entscheidungsausgänge einer Gegenüberstellung zwei detektionsbasierte Prozesse – die Detektion der An- oder Abwesenheit der tatbeteiligten Person – und zwei nicht-detektionsbasierte Prozesse – eine auf Unfairness oder eine auf Raten basierte Auswahl – zu erfassen. Zu Validierungszwecken wurden für jeden der vier Modellparameter zwei bereits publizierte Datensätze mit geeigneten experimentellen Manipulationen reanalysiert, die den betreffenden Parameter in vorhersagbarer Weise beeinflussen sollten. In allen acht modellbasierten Reanalysen waren die Parameter nachweislich sensitiv für die Manipulationen der Prozesse, die sie erfassen sollten. Nach erfolgreicher Validierung wurden die Chancen des Modells für die Gegenüberstellungsforschung durch die Anwendung auf zwei konkrete Forschungsfragen aufgezeigt: (1) In den Experimenten 1a bis 1d wurde überprüft, ob die Verwendung digital kombinierter (gemorphter) Vergleichsbilder zu unfairen Gegenüberstellungen führt. Die Fairness von Gegenüberstellungen mit gemorphten und nicht-gemorphten Vergleichsbildern wurde wie üblich über die Entscheidungen von Pseudozeug:innen – Personen, die die Tat nicht beobachtet hatten – beurteilt. Mithilfe des Modells konnte zudem die Auftretenswahrscheinlichkeit des auf Unfairness basierten Auswahlprozesses einer tatverdächtigen Person direkt anhand der Daten der Augenzeug:innen erfasst werden. Basierend auf den Daten der Pseudozeug:innen hätte man schlussfolgern müssen, dass die Gegenüberstellungen mit gemorphten Vergleichsbildern unfairer waren als die Gegenüberstellungen mit nicht-gemorphten Vergleichsbildern. In den modellbasierten Analysen der Daten von Augenzeug:innen konnte hingegen kein Effekt der Morphing-Manipulation auf die Wahrscheinlichkeit einer auf Unfairness basierten Auswahl in sachgerecht durchgeführten Gegenüberstellungen gefunden werden. Diese divergierenden Befunde verdeutlichen die Unterschiede zwischen Augenzeug:innen und Pseudozeug:innen und demonstrieren damit, wie wichtig es ist, die Fairness einer Gegenüberstellung direkt anhand der Daten der Augenzeug:innen zu erfassen. (2) Weiterhin wurde in den Experimenten 2a und 2b der Einfluss der Gegenüberstellungsgröße auf die zugrundeliegenden Prozesse der Entscheidungen von Augenzeug:innen untersucht. Kleinere im Vergleich zu größeren Gegenüberstellungen verbesserten die Detektion der Anwesenheit der tatbeteiligten Person und reduzierten die Wahrscheinlichkeit einer ratebasierten Auswahl. Wenn jedoch eine ratebasierte Auswahl erfolgt, ist die Wahrscheinlichkeit, dass diese Auswahl zufällig auf die tatverdächtige Person fällt, in größeren Gegenüberstellungen niedriger als in kleineren Gegenüberstellungen. Ausblickend auf zukünftige Forschung wird abschließend eine erweiterte Variante des Modells vorgestellt, die auch eine Berücksichtigung der Konfidenzurteile von Augenzeug:innen erlaubt. Wie die vorliegende Dissertation zeigt, bietet das Modell vielseitige Möglichkeiten, um die Gegenüberstellungsforschung gewinnbringend voranzubringen.

Abstract

The fact that numerous wrongful convictions can be attributed to eyewitness misidentification highlights the need to enhance the current understanding of the processes underlying eyewitness decisions. In this dissertation, a new multinomial processing tree model is introduced. By considering all decisions observed in lineups, the model serves to estimate two detection-based processes—the detection of the presence or absence of the culprit—and two non-detection-based processes—biased suspect selection and guessing-based selection. For validation purposes, two published data sets were reanalyzed for each of the four model parameters involving experimental manipulations intended to predictably influence the respective parameter. In all eight model-based reanalyses, the parameters were sensitive to manipulations of the processes they were designed to measure. After successful validation, the model's potential for lineup research was demonstrated by applying it to two research questions: (1) In Experiments 1a to 1d, it was tested whether the use of digitally combined (morphed) filler photographs leads to unfair lineups. To mirror the approach that is commonly applied, the fairness of lineups with morphed and non-morphed fillers was assessed based on the decisions of mock-witnesses—persons who had not witnessed the crime. The biased selection of the suspects was also directly measured from eyewitness data using the model. The mock-witness data indicated that lineups with morphed fillers were more unfair than lineups with non-morphed fillers. However, the model-based analyses of eyewitness data yielded no effect of the morphing manipulation on biased suspect selection in properly conducted lineups. These divergent findings demonstrate the differences between eyewitnesses and mock witnesses and thus highlight the importance of measuring lineup fairness directly from eyewitness data. (2) Furthermore, in Experiments 2a and 2b, the influence of lineup size on the processes underlying eyewitness decisions was examined. Smaller compared to larger lineups increased culprit-presence detection and decreased guessing-based selection. However, if guessing-based selection occurs, the probability of randomly selecting the suspect is reduced in larger compared to smaller lineups. Finally, an outlook for future research was provided by introducing an extended version of the model which allows for the consideration of eyewitnesses' confidence judgments. The findings derived from this dissertation show that the model is a promising measurement tool for enhancing lineup research.

Einleitung

Eine Gegenüberstellung stellt eine polizeiliche Ermittlungsmaßnahme dar, um zu prüfen, ob eine tatverdächtige Person tatsächlich schuldig oder unschuldig ist (Wells & Luus, 1990). Dabei wird einem:einer Augenzeug:in eine Reihe von Personen vorgeführt, von denen eine die tatverdächtige Person ist, während die anderen als sicher unschuldige Vergleichspersonen dienen. Obwohl die Entscheidungen von Augenzeug: innen in Gegenüberstellungen oftmals wertvolle Hinweise für den Ermittlungsprozess liefern, bergen sie auch ein hohes Risiko für Fehlurteile und Justizirrtümer. Ist die tatverdächtige Person tatsächlich schuldig, kann sie korrekt wiedererkannt werden, jedoch besteht ebenfalls das Risiko, dass fälschlicherweise eine Vergleichsperson ausgewählt wird oder eine falsche Zurückweisung der Gegenüberstellung erfolgt. Ist die tatverdächtige Person hingegen unschuldig, kann die Gegenüberstellung korrekt zurückgewiesen werden, allerdings kann auch hier fälschlicherweise eine Vergleichsperson oder, im schlimmsten Fall, die unschuldige tatverdächtige Person gewählt werden. Besonders alarmierend sind in diesem Zusammenhang die Ergebnisse der gemeinnützigen US-amerikanischen Organisation „Innocence Project", die sich für die Aufklärung von Justizirrtümern einsetzt. Inzwischen konnte die Unschuld von über 350 verurteilten Personen durch nachträgliche DNS-Analysen bewiesen werden. Dabei waren mehr als 60 % der Justizirrtümer auf das fälschliche Wiedererkennen der tatverdächtigen Person durch Augenzeug:innen zurückzuführen (Innocence Project, 2020). Um die Fehlurteile von Augenzeug:innen zu reduzieren, beschäftigt sich die Forschung seit Jahren mit der Optimierung von Gegenüberstellungen (Wells et al., 2020; Wells et al., 1998).

Die Güte von Gegenüberstellungen wurde lange Zeit mithilfe des *Diagnostizitätsverhältnisses* (Wells & Lindsay, 1980) beurteilt. Das Diagnostizitätsverhältnis beschreibt das Verhältnis der Auswahlrate tatbeteiligter Personen zur Auswahlrate unschuldiger tatverdächtiger Personen. Ein hohes Verhältnis impliziert somit eine hohe Wahrscheinlichkeit, dass die ausgewählte tatverdächtige Person tatsächlich schuldig ist (Wells, 2014). Deshalb wurde angenommen, dass Gegenüberstellungsprozeduren mit einem höheren Diagnostizitätsverhältnis gegenüber anderen mit einem geringeren Verhältnis zu bevorzugen seien. Aus signalentdeckungstheoretischer Perspektive konfundiert das Diagnostizitätsverhältnis jedoch zwei wesentliche Größen, nämlich die Fähigkeit der Augenzeug:innen, zwischen tatbeteiligter und unschuldiger tatverdächtiger Person zu differenzieren – die Diskriminationsfähigkeit – und die Bereitschaft, generell eine Person aus der Gegenüberstellung zu wählen – die Antwortneigung (z. B. Meissner et al., 2005; Palmer & Brewer, 2012). Dies hat die unerwünschte Konsequenz, dass das Diagnostizitätsverhältnis nicht nur mit zunehmender Diskriminationsfähigkeit ansteigt, sondern auch dann, wenn lediglich die Antwortneigung konservativer wird (Gronlund et al., 2014; Wixted & Mickes, 2012).

Um die Diskriminationsfähigkeit der Augenzeug:innen unabhängig von deren Antwortneigung messen zu können, führten Mickes et al. (2012) die sogenannte *Receiver-Operating-Characteristic-(ROC-)Analyse* in die Gegenüberstellungsforschung ein. Bei einer ROC-Analyse wird die Auswahl-

rate tatbeteiligter Personen gegen die Auswahlrate unschuldiger tatverdächtiger Personen für verschiedene Stufen der Antwortneigung abgetragen. Die Antwortneigung wird dabei typischerweise über die Abfrage der subjektiven Konfidenz der Augenzeug:innen nach einem Urteil erfasst. Die resultierenden Wertepaare können durch eine Linie, die sogenannte ROC-Kurve, verbunden werden. Beim Vergleich verschiedener Gegenüberstellungsprozeduren wird die Prozedur mit der größeren Fläche unter der ROC-Kurve als überlegen angesehen (Mickes et al., 2012).

ROC-Analysen basieren jedoch auf dem Standardmodell der Signalentdeckungstheorie und wurden ursprünglich für einfache Detektionsaufgaben mit vier möglichen Ausgängen (korrekte und falsche Auswahl, korrekte und falsche Zurückweisung) entwickelt. Bei Aufgaben mit einer solchen 2 × 2-Datenstruktur ist es zulässig, nur die korrekte und falsche Auswahlrate zu berücksichtigen, da die Rate falscher beziehungsweise korrekter Zurückweisungen daraus abgeleitet werden kann und somit keine weiteren Informationen liefert (Macmillan & Creelman, 2005). Gegenüberstellungen beinhalten jedoch nicht nur eine schuldige oder unschuldige tatverdächtige Person, sondern auch mehrere Vergleichspersonen. Wie anfangs beschrieben, ergeben sich somit sechs mögliche Ausgänge einer Gegenüberstellung: In Gegenüberstellungen mit und ohne tatbeteiligte Person besteht die Möglichkeit, die tatverdächtige Person auszuwählen, eine Vergleichsperson auszuwählen oder die Gegenüberstellung zurückzuweisen (siehe Tabelle 1 in Einzelarbeit 1). Um ROC-Analysen im Kontext von Gegenüberstellungen trotzdem nutzen zu können, wird diese 2 × 3-Datenstruktur auf eine 2 × 2-Datenstruktur reduziert, indem die Auswahl einer Vergleichsperson als Zurückweisung klassifiziert wird. Dieses Vorgehen wird damit gerechtfertigt, dass die Auswahl einer Vergleichsperson keinerlei strafrechtliche Konsequenzen für die tatverdächtige Person habe und somit funktional äquivalent zur Zurückweisung einer Gegenüberstellung sei (z. B. Mickes et al., 2012; Wixted & Mickes, 2012). Werden jedoch nur zwei der sechs Datenkategorien berücksichtigt (Auswahl schuldiger und unschuldiger tatverdächtiger Personen) und die anderen ignoriert, gehen wertvolle Informationen verloren, die für das Verständnis der den Entscheidungen von Augenzeug:innen zugrundeliegenden latenten Prozesse entscheidend sind (Smith et al., 2017; Smith et al., 2018; Smith et al., 2020; Wells, Smalarz, & Smith, 2015; Wells, Smith, & Smalarz, 2015). Es ist beispielsweise ein gravierender Unterschied, ob in einer Gegenüberstellung ohne tatbeteiligte Person eine Vergleichsperson ausgewählt oder die Gegenüberstellung zurückgewiesen wird. Während es sich im ersten Fall um eine falsche Entscheidung handelt, stellt der zweite Fall eine korrekte Entscheidung dar.

Im Folgenden wird ein neues Messmodell für Gegenüberstellungen vorgestellt (fortan als 2-HT EI-Modell für *two-high threshold eyewitness identification model* bezeichnet), das unter Berücksichtigung der *gesamten* 2 × 3-Datenstruktur von Gegenüberstellungen die Erfassung der latenten Prozesse ermöglicht, die den Entscheidungen von Augenzeug:innen zugrunde liegen. Das 2-HT EI-Modell gehört zur Klasse der multinomialen Verarbeitungsbaummodelle (Batchelder & Riefer, 1999; Erdfelder et al., 2009), die in der Sozial-und Kognitionspsychologie ein nützliches Werkzeug darstellen, um die Wahrscheinlichkeiten des Auftretens latenter Prozesse aus beobachtbaren, katego-

rialen Daten zu schätzen (z. B. Bröder & Meiser, 2007; Erdfelder et al., 2007; Kroneisen & Heck, 2020; Meiser & Bröder, 2002; Smith & Bayen, 2004). Die Modellstruktur des 2-HT EI-Modells wird in Abbildung 1 visualisiert. Der obere Modellbaum repräsentiert die postulierten Prozesse, die den Entscheidungen von Augenzeug:innen in Gegenüberstellungen *mit* tatbeteiligter Person zugrunde liegen. Die Anwesenheit der tatbeteiligten Person wird mit der Wahrscheinlichkeit *dP* detektiert. Wird die Anwesenheit der tatbeteiligten Person nicht detektiert ($1 - dP$), können zwei nicht-detektionsbasierte Prozesse dennoch zu einer korrekten Auswahl der tatbeteiligten Person führen. Mit der Wahrscheinlichkeit b wird die tatbeteiligte Person gewählt, weil sie aus den anderen Personen der Gegenüberstellung herausragt. In diesem Fall spricht man von einer unfairen Gegenüberstellung. Wird die tatbeteiligte Person nicht aufgrund einer Unfairness in der Gegenüberstellung ausgewählt ($1 - b$), können Augenzeug:innen sich mit der Wahrscheinlichkeit g immer noch dazu entscheiden, eine Person der Gegenüberstellung basierend auf Raten auszuwählen. Die Wahrscheinlichkeit, dass die ratebasierte Auswahl dabei zufällig auf die tatbeteiligte Person fällt, entspricht dem Kehrwert von n, wobei n die Anzahl der Personen in der Gegenüberstellung repräsentiert. In einer aus sechs Personen bestehenden Gegenüberstellung liegt die zufällige Auswahlwahrscheinlichkeit der tatbeteiligten Person beispielsweise bei ⅙. Mit der Gegenwahrscheinlichkeit (hier ⅚) wird eine der Vergleichspersonen gewählt. Erfolgt keine ratebasierte Auswahl ($1 - g$), wird die Gegenüberstellung fälschlicherweise zurückgewiesen. Der untere Modellbaum repräsentiert die postulierten Prozesse, die den Entscheidungen von Augenzeug:innen in Gegenüberstellungen *ohne* tatbeteiligte Person zugrunde liegen. Die Abwesenheit der tatbeteiligten Person wird mit der Wahrscheinlichkeit *dA* detektiert, was zu einer korrekten Zurückweisung der Gegenüberstellung führt. Wird die Abwesenheit der tatbeteiligten Person nicht detektiert ($1 - dA$), werden die gleichen nicht-detektionsbasierten Prozesse angenommen wie in Gegenüberstellungen mit tatbeteiligter Person. Wird die An- oder Abwesenheit der tatbeteiligten Person nicht detektiert, sind Augenzeug:innen nicht in der Lage, zwischen Gegenüberstellungen mit und ohne tatbeteiligte Person zu differenzieren, sodass sich die nicht-detektionsbasierten Prozesse nicht unterscheiden sollten.

Bevor das 2-HT EI-Modell zur Untersuchung neuer rechtspsychologischer Fragestellungen verwendet werden kann, müssen die Parameter des Modells validiert werden. Es muss also zunächst sichergestellt werden, dass die Modellparameter tatsächlich die Prozesse abbilden, die ihrer Beschreibung entsprechen (Erdfelder et al., 2009; Schmidt et al., 2023). Zu diesem Zweck wurden in Winter et al. (2022) neue Experimente durchgeführt, die darauf abzielten, die betreffenden Modellparameter durch geeignete experimentelle Manipulationen in vorhersagbarer Weise zu beeinflussen. Ein Modell sollte sich allerdings nicht nur bei neuen Daten – die explizit für die Modellvalidierung erhoben wurden – bewähren, sondern sollte auch bereits publizierte Daten anderer Forschungsgruppen korrekt beschreiben können. Im Rahmen der vorliegenden Dissertation wurden zu Validierungszwecken deshalb für jeden der vier Parameter des 2-HT EI-Modells zwei bereits publizierte Datensätze reanalysiert, bei denen experimentelle Manipulationen verwendet

wurden, die den betreffenden Parameter in vorhersagbarer Weise beeinflussen sollten (Reanalysen 1a bis 4b). Nach erfolgreicher Modellvalidierung lag das Ziel in der konkreten Anwendung des 2-HT EI-Modells. Anhand von zwei inhaltlich relevanten Fragestellungen zur Fairness (Experimente 1a bis 1d) und zur Größe von Gegenüberstellungen (Experimente 2a und 2b) konnten die Chancen des modellbasierten Ansatzes für die zukünftige Forschung demonstriert werden.

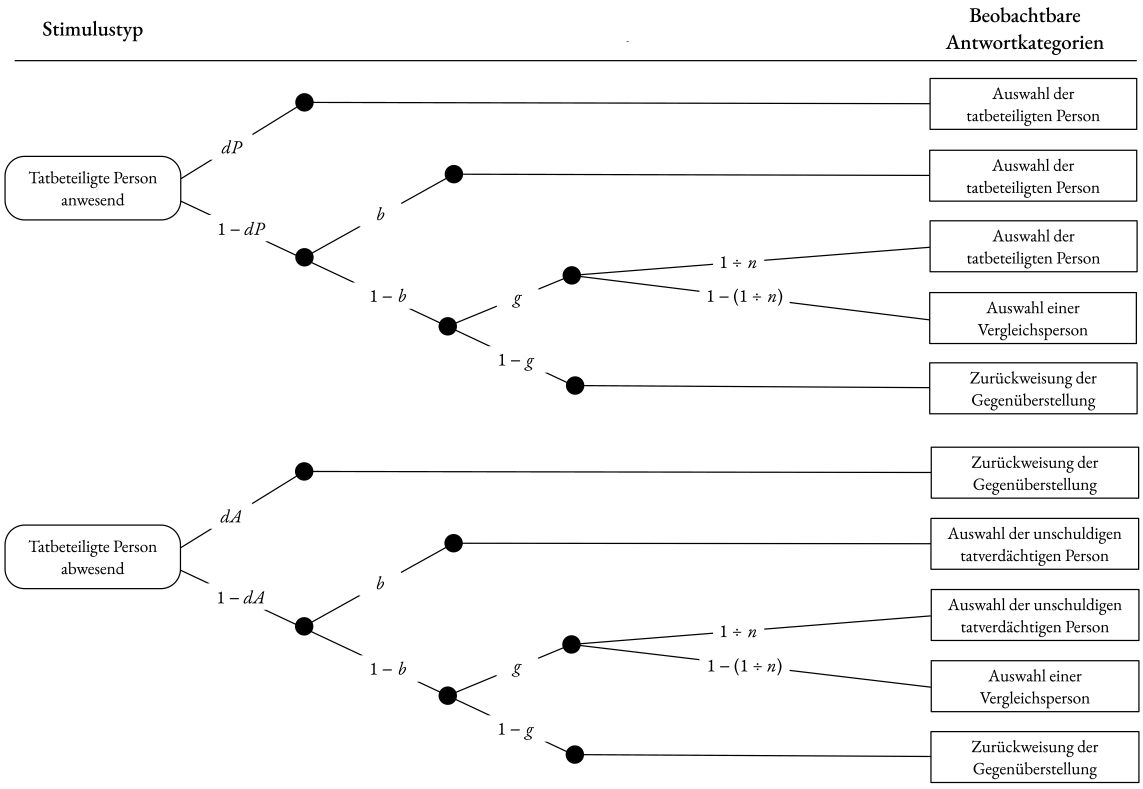

Abbildung 1. Graphische Darstellung des 2-HT EI-Modells. Die abgerundeten Rechtecke auf der linken Seite repräsentieren die Gegenüberstellungen mit und ohne tatbeteiligte Person. Die Rechtecke auf der rechten Seite stellen die beobachtbaren Antwortkategorien dar. Die Parameter entlang der Äste repräsentieren die Wahrscheinlichkeiten der angenommenen latenten Prozesse (*dP* = Wahrscheinlichkeit, die Anwesenheit der tatbeteiligten Person zu detektieren; *b* = Wahrscheinlichkeit einer auf Unfairness basierten Auswahl der tatverdächtigen Person; *g* = Wahrscheinlichkeit einer ratebasierten Auswahl; *dA* = Wahrscheinlichkeit, die Abwesenheit der tatbeteiligten Person zu detektieren). Die Wahrscheinlichkeit, dass die tatverdächtige Person basierend auf Raten aus der Gegenüberstellung ausgewählt wird, entspricht dem Kehrwert von *n*, wobei *n* die Anzahl der Personen in der Gegenüberstellung repräsentiert.

Modellvalidierung anhand publizierter Daten

Das erste Ziel der vorliegenden Dissertation lag in der Validierung der Modellparameter des 2-HT EI-Modells anhand der Reanalyse bereits publizierter Daten (siehe Einzelarbeit 1). Bei der Auswahl geeigneter Studien für die modellbasierten Reanalysen wurden folgende Kriterien angelegt: (1) Der Effekt der experimentellen Manipulation auf den betreffenden Parameter musste möglichst offensichtlich sein. (2) Die Daten mussten ausreichend detailliert vorliegen. (3) Das Studiendesign musste von minimaler Komplexität sein, um die Reanalyse möglichst einfach zu

gestalten. Für jeden der vier Modellparameter wurden die ersten zwei gefundenen Studien, die diese drei Kriterien erfüllten, reanalysiert.

Reanalysen 1a und 1b: Validierung des Parameters *dP*

Zunächst lag der Fokus auf der Validierung des Parameters *dP,* der die Wahrscheinlichkeit repräsentiert, die Anwesenheit der tatbeteiligten Person in einer Gegenüberstellung zu detektieren. Zur Validierung dieses Parameters wurden publizierte Datensätze von Memon et al. (2003) und Smith (2020) herangezogen, bei denen gezielt die Fähigkeit, die Anwesenheit der tatbeteiligten Person zu detektieren, manipuliert wurde. In der Studie von Memon et al. (2003) wurde die Darbietungsdauer des Tatvideos manipuliert. Die Teilnehmenden sahen entweder ein langes (100 Sekunden) oder ein kurzes (67 Sekunden) Video des Tathergangs. Während das Gesicht der tatbeteiligten Person in dem langen Video für 45 Sekunden klar erkennbar war, gewährte das kurze Video lediglich für 12 Sekunden eine klare Sicht auf das Gesicht der tatbeteiligten Person. In der Studie von Smith (2020) wurde hingegen die Qualität des Tatvideos manipuliert. Die Teilnehmenden sahen entweder das Video des Tathergangs in hoher Auflösung oder eine niedrig aufgelöste, überbelichtete Version desselben Videos, bei der es äußerst schwierig war, Details des Gesichts der tatbeteiligten Person zu erkennen. Die modellbasierten Reanalysen bestätigten, dass der Anwesenheits-Detektions-Parameter *dP* erwartungsgemäß auf beide Enkodierungsmanipulationen reagierte: Parameter *dP* war signifikant höher bei langer Darbietungsdauer und hoher Qualität des Tatvideos als bei kurzer Darbietungsdauer und geringer Qualität des Tatvideos.

In beiden Reanalysen zeigte sich ein zusätzlicher Effekt der Enkodierungsmanipulationen auf den Rateparameter *g*. Die Wahrscheinlichkeit einer ratebasierten Auswahl war höher bei geringerer Fähigkeit, die Anwesenheit der tatbeteiligten Person zu detektieren. Konkret war Parameter *g* höher bei kurzer Darbietungsdauer und niedriger Qualität des Tatvideos als bei langer Darbietungsdauer und hoher Qualität des Tatvideos. Im Idealfall werden zur Modellvalidierung experimentelle Manipulationen verwendet, die sich selektiv auf den Zielparameter auswirken und möglichst keinen Einfluss auf andere Modellparameter haben (Batchelder & Riefer, 1999). Häufig ist es jedoch nicht möglich, derart theoretisch starke Manipulationen zu finden, die nur einen Prozess beeinflussen, ohne einen Nebeneffekt auf andere Prozesse zu haben (Schmidt et al., 2023). In diesem Fall ist es wichtig, dass diese Nebeneffekte plausibel erklärt werden können. Der zusätzliche Effekt der Enkodierungsmanipulationen auf den Rateparameter *g* kann hier über den Erklärungsansatz des kompensatorischen Ratens begründet werden. Dieser Ansatz besagt, dass Personen bei schwacher Gedächtnisleistung eher zu einer ratebasierten Auswahl neigen, um die unzureichende Erinnerung zu kompensieren (Batchelder & Batchelder, 2008; Bayen & Kuhlmann, 2011; Ehrenberg & Klauer, 2005; Hirshman, 1995; Küppers & Bayen, 2014; Meiser et al., 2007; Riefer et al., 1994). Auch im Kontext der multinomialen Modellierung konnte bereits in anderen Studien beobachtet werden, dass sich eine Manipulation der Gedächtnisleistung nicht

nur auf die Detektionsparameter auswirkt, sondern auch die Rateprozesse beeinflusst (z. B. Bayen et al., 1996; Meiser et al., 2007).

Reanalysen 2a und 2b: Validierung des Parameters *b*

Die Reanalysen 2a und 2b dienten der Validierung des Parameters *b,* der die Wahrscheinlichkeit repräsentiert, dass eine tatverdächtige Person nur deshalb gewählt wird, weil sie aus der Reihe der Personen in der Gegenüberstellung herausragt. Es wurden publizierte Datensätze von Wetmore et al. (2015) und Colloff et al. (2016) reanalysiert, bei denen gezielt die Fairness der Gegenüberstellung manipuliert wurde. Wetmore et al. (2015) manipulierten die Ähnlichkeit zwischen der tatverdächtigen Person und den Vergleichspersonen. Zur Erstellung einer fairen Gegenüberstellung wurden Vergleichspersonen mit hoher Ähnlichkeit zur tatverdächtigen Person verwendet, während für die Erstellung einer unfairen Gegenüberstellung Vergleichspersonen mit geringer Ähnlichkeit zur tatverdächtigen Person ausgewählt wurden. Colloff et al. (2016) präsentierten den Teilnehmenden hingegen ein Tatvideo, in dem die tatbeteiligte Person ein distinktives Gesichtsmerkmal aufwies (z. B. ein blaues Auge oder eine Tätowierung im Gesicht). Zur Erstellung der fairen Gegenüberstellungen wurden Bildbearbeitungstechniken verwendet, die verhinderten, dass das Bild der tatverdächtigen Person aufgrund des distinktiven Merkmals aus der Gegenüberstellung herausragte. Das distinktive Merkmal wurde entweder zu jedem anderen Gesicht der Vergleichspersonen digital hinzugefügt oder der betroffene Gesichtsbereich wurde bei allen Gesichtern verpixelt beziehungsweise durch einen schwarzen Balken verdeckt. Bei den unfairen Gegenüberstellungen hingegen war die tatverdächtige Person die einzige Person mit dem distinktiven Gesichtsmerkmal, sodass sie deutlich aus der Gegenüberstellung herausragte. In beiden Reanalysen konnte der Unfairness-Parameter *b* durch die Fairness-Manipulationen in vorhersagbarer Weise beeinflusst werden, was die Validität dieses Parameters stützt: Parameter *b* war signifikant höher für unfaire Gegenüberstellungen als für faire Gegenüberstellungen.

In beiden Reanalysen zeigte sich ein zusätzlicher Effekt der Fairness-Manipulationen auf den Rateparameter *g.* Die Wahrscheinlichkeit einer ratebasierten Auswahl war in fairen Gegenüberstellungen signifikant höher als in unfairen Gegenüberstellungen. Aufgrund der geringen Ähnlichkeit zwischen den Vergleichspersonen und der tatverdächtigen Person in unfairen Gegenüberstellungen ist zu erwarten, dass die Teilnehmenden weniger wahrscheinlich eine ratebasierte Auswahl treffen als in fairen Gegenüberstellungen, in denen alle Personen der Gegenüberstellung echte Antwortalternativen darstellen (Bell et al., 2023; Winter et al., 2022). Bei der Reanalyse der Daten von Colloff et al. (2016) zeigte sich zusätzlich ein Effekt der Fairness auf Parameter *dP.* Die Wahrscheinlichkeit, die tatbeteiligte Person zu detektieren, war in unfairen Gegenüberstellungen signifikant höher als in fairen Gegenüberstellungen. Auch dieser Nebeneffekt kann post hoc plausibel erklärt werden. Es ist zu erwarten, dass die distinktiven Gesichtsmerkmale in unfairen Gegenüberstellungen die Aufmerksamkeit der Teilnehmenden auf das Gesicht der tatbeteiligten

Person lenken, was die Detektion erleichtern sollte. Es konnte bereits mehrfach nachgewiesen werden, dass sich mit zunehmender Ähnlichkeit zwischen den Vergleichspersonen und der tatverdächtigen Person die Identifikationsleistung von Augenzeug:innen reduziert (Fitzgerald et al., 2015; Smith et al., 2017; Smith et al., 2018).

Reanalysen 3a und 3b: Validierung des Parameters *g*

Das Ziel der Reanalysen 3a und 3b war die Validierung des Parameters *g,* der die Wahrscheinlichkeit repräsentiert, dass eine Person aus der Gegenüberstellung basierend auf Raten ausgewählt wird. Es wurden publizierte Datensätze von Malpass und Devine (1981) sowie Lampinen et al. (2020) reanalysiert, bei denen gezielt das Rateverhalten der Teilnehmenden über die Instruktionen vor der Gegenüberstellung manipuliert wurde. In beiden Studien wurden entweder zweiseitige oder einseitige Instruktionen verwendet. Zweiseitige Instruktionen weisen darauf hin, dass die tatbeteiligte Person sowohl anwesend als auch abwesend sein kann, und betonen damit nicht nur die Wichtigkeit einer korrekten Auswahl einer tatbeteiligten Person, sondern auch einer korrekten Zurückweisung einer Gegenüberstellung. Einseitige Instruktionen suggerieren hingegen, dass sich die tatbeteiligte Person in der Gegenüberstellung befindet, und erhöhen damit die Bereitschaft der Teilnehmenden, bei Unsicherheit eine Person in der Gegenüberstellung basierend auf Raten auszuwählen (z. B. Brewer & Wells, 2006; Clark, 2005; Keast et al., 2007). Die Manipulation der Instruktionen konnte den Rateparameter *g* in beiden Reanalysen in vorhersagbarer Weise selektiv beeinflussen: Bei der Verwendung einseitiger Instruktionen war Parameter *g* signifikant höher als bei der Verwendung zweiseitiger Instruktionen.

Reanalysen 4a und 4b: Validierung des Parameters *dA*

Zuletzt stand die Validierung des Parameters *dA* an, der die Wahrscheinlichkeit repräsentiert, die Abwesenheit der tatbeteiligten Person in einer Gegenüberstellung zu detektieren. Während für die anderen Modellparameter relativ einfach offensichtliche und triviale Manipulationen in der Literatur gefunden werden konnten, die den betreffenden Parameter in vorhersagbarer Weise beeinflussen sollten, gestaltete sich dies beim Parameter *dA* schwieriger. Winter et al. (2022) konnten nachweisen, dass Parameter *dA* im Vergleich zu einer Kontrollbedingung signifikant erhöht war, wenn alle Personen in einer Gegenüberstellung ohne tatbeteiligte Person (auch die unschuldige tatverdächtige Person) ein auffälliges Gesichtsmerkmal aufwiesen und somit als tatbeteiligte Person leicht ausgeschlossen werden konnten. Eine solch triviale Manipulation erscheint jedoch nur im Kontext einer Modellvalidierung sinnvoll und kann daher in der Literatur nicht gefunden werden. Deshalb wurde der Schwerpunkt auf experimentelle Manipulationen gelegt, die entwickelt wurden, um die Detektion der Abwesenheit der tatbeteiligten Person unter realistischeren Bedingungen zu verbessern.

Karageorge und Zajac (2011) verwendeten eine *Wildcard,* um Kindern das Zurückweisen von Gegenüberstellungen bei Abwesenheit der tatbeteiligten Person zu erleichtern. Bei einer Wildcard handelt es sich um das Bild einer Silhouette des Kopf- und Halsbereichs einer Person mit einem darin platzierten Fragezeichen. In der Wildcard-Bedingung wurde die Wildcard mittig zwischen den anderen Bildern der Gegenüberstellung präsentiert und konnte gewählt werden, um die Gegenüberstellung zurückzuweisen. In der Kontrollbedingung sollten die Kinder der Versuchsleitung hingegen verbal den Hinweis geben, dass sie die Gegenüberstellung zurückweisen wollen.

Wilcock und Bull (2010) verwendeten eine Übungs-Gegenüberstellung ohne tatbeteiligte Person, um älteren Erwachsenen die Detektion der Abwesenheit der tatbeteiligten Person zu erleichtern. Der einen Hälfte der Teilnehmenden wurde vor den eigentlichen Gegenüberstellungen eine Übungs-Gegenüberstellung präsentiert, die aus sechs Bildern berühmter Frauen bestand. Die Teilnehmenden wurden instruiert, die ehemalige Königin von England, Elisabeth II., zu identifizieren (es gab kein Bild der Königin). Anschließend wurden die Teilnehmenden darauf hingewiesen, dass nicht alle Gegenüberstellungen eine tatbeteiligte Person enthalten und dass auch die Polizei Fehler machen könne. Der anderen Hälfte der Teilnehmenden wurde keine Übungs-Gegenüberstellung vor den eigentlichen Gegenüberstellungen präsentiert.

Sowohl die Verwendung einer Wildcard als auch die Verwendung einer Übungs-Gegenüberstellung führte im Vergleich zur jeweiligen Kontrollbedingung zu einer signifikant erhöhten Rate korrekter Zurückweisungen von Gegenüberstellungen ohne tatbeteiligte Person, ohne die Rate falscher Zurückweisungen von Gegenüberstellungen mit tatbeteiligter Person zu beeinflussen (siehe auch Havard et al., 2017; Havard & Memon, 2013; Zajac & Karageorge, 2009). Dieses Ergebnismuster lässt vermuten, dass beide Manipulationen die Detektion der Abwesenheit der tatbeteiligten Person verbessern (Parameter *dA*) und nicht die Wahrscheinlichkeit einer ratebasierten Auswahl reduzieren (Parameter *g*). Wie erwartet, zeigte sich in beiden modellbasierten Reanalysen ein selektiver Einfluss beider Manipulationen auf Parameter *dA*: Die Verwendung einer Wildcard und einer Übungs-Gegenüberstellung erhöhte signifikant Parameter *dA* im Vergleich zur jeweiligen Kontrollbedingung.

Diskussion

In acht Reanalysen bereits publizierter Datensätze konnte gezeigt werden, dass die Parameter des 2-HT EI-Modells sensitiv auf Manipulationen jener Prozesse reagierten, die sie erfassen sollten. In den Reanalysen 1a und 1b konnte der Anwesenheits-Detektions-Parameter *dP* in vorhersagbarer Weise durch Manipulationen der Enkodierungsbedingungen beeinflusst werden (Memon et al., 2003; Smith, 2020): Parameter *dP* war höher bei langer Darbietungsdauer und hoher Qualität des Tatvideos als bei kurzer Darbietungsdauer und geringer Qualität des Tatvideos. In den Reanalysen 2a und 2b reagierte der Unfairness-Parameter *b* erwartungsgemäß auf Manipulationen der Fairness (Colloff et al., 2016; Wetmore et al., 2015): Zum einen war Parameter *b* höher in Gegen-

überstellungen mit geringer als in solchen mit hoher Ähnlichkeit zwischen der tatverdächtigen Person und den Vergleichspersonen. Zum anderen war Parameter *b* höher in Gegenüberstellungen, bei denen die tatverdächtige Person aufgrund eines distinktiven Merkmals aus den anderen Personen herausragte, als in Gegenüberstellungen, bei denen das distinktive Merkmal verdeckt oder auf die Gesichter der anderen Personen repliziert wurde. In den Reanalysen 3a und 3b hatten Instruktionen vor der Gegenüberstellung den erwarteten Effekt auf den Rateparameter *g* (Lampinen et al., 2020; Malpass & Devine, 1981): Einseitige Instruktionen führten zu einem höheren *g*-Parameter als zweiseitige Instruktionen. Zuletzt konnte auch der Abwesenheits-Detektions-Parameter *dA* durch Manipulationen, die gezielt das Zurückweisen von Gegenüberstellungen ohne tatbeteiligte Person erleichtern sollen, hypothesenkonform beeinflusst werden (Karageorge & Zajac, 2011; Wilcock & Bull, 2010): Die Verwendung einer Wildcard oder einer Übungs-Gegenüberstellung erhöhte Parameter *dA* im Vergleich zu einer Kontrollbedingung.

Die modellbasierten Reanalysen zeigen also, dass die manipulierten Prozesse sogar über verschiedene experimentelle Manipulationen, Stichproben und Paradigmen hinweg in den betreffenden Parametern des 2-HT EI-Modells reflektiert werden konnten. Damit ergänzen diese Validierungsbefunde die Ergebnisse von Winter et al. (2022), die in einer Serie von neuen Experimenten die Modellparameter ebenfalls erfolgreich validieren konnten. Da beide Validierungsuntersuchungen das Modell stützen, kann es nun in der Forschung gewinnbringend eingesetzt werden. Mithilfe des Modells kann nicht nur zwischen den Prozessen der Detektion der An- und Abwesenheit der tatbeteiligten Person differenziert werden, sondern es erlaubt zusätzlich zwei nicht-detektionsbasierte Prozesse – eine auf Unfairness oder eine auf Raten basierte Auswahl – getrennt voneinander zu erfassen. Dadurch bietet das Modell erstmals die Möglichkeit, die Auftretenswahrscheinlichkeit des auf Unfairness basierten Auswahlprozesses einer tatverdächtigen Person *direkt* anhand der Daten der Augenzeug:innen zu messen.

Die Fairness von Gegenüberstellungen

Die Fairness von Gegenüberstellungen wurde bislang typischerweise mithilfe des Pseudozeug:innen-Paradigmas erfasst. Pseudozeug:innen – Personen, denen die Identität der tatbeteiligten Person unbekannt ist – wird dabei eine Gegenüberstellung präsentiert mit der Instruktion, diejenige Person zu wählen, die ihrer Meinung nach die von der Polizei tatverdächtige Person darstellt (Doob & Kirshenbaum, 1973). Eine Möglichkeit besteht darin, dass die Pseudozeug:innen eine Beschreibung der tatbeteiligten Person als Grundlage für ihre Entscheidung erhalten (z. B. Humphries et al., 2012; Mansour et al., 2017). Alternativ werden ihnen keine zusätzlichen Informationen präsentiert außer dem Hinweis, dass die tatverdächtige Person aus den anderen Personen der Gegenüberstellung herausragen könnte (z. B. Brigham et al., 1999; Flowe & Humphries, 2011). Damit soll nicht nur überprüft werden, ob die Vergleichspersonen der Beschreibung der tatbeteiligten Person genügen, sondern auch, ob die tatverdächtige Person aufgrund anderer Hinweise

aus der Gegenüberstellung herausragt (Malpass et al., 2007). Das Bild der tatverdächtigen Person stammt beispielsweise oft aus einer anderen Quelle als die Bilder der Vergleichspersonen, sodass allein Bildunterschiede in der Helligkeit, dem Kontrast oder der Farbbalance die tatverdächtige Person hervorstechen lassen können. In einer fairen Gegenüberstellung sollten sich die Wahlen der Pseudozeug:innen gleichmäßig über die Personen in der Gegenüberstellung verteilen. Wählen die Pseudozeug:innen hingegen überzufällig häufig die tatverdächtige Person, kann von einer unfairen Gegenüberstellung gesprochen werden.

Auf den ersten Blick bietet das Pseudozeug:innen-Paradigma eine vermeintlich einfache Lösung, um die Fairness von Gegenüberstellungen zu beurteilen. Allerdings wurde die Validität dieses Verfahrens angesichts entscheidender Unterschiede in der Aufgabe von Augenzeug:innen und Pseudozeug:innen in Frage gestellt (Corey et al., 1999; Lee et al., 2022; Malpass et al., 2007): Erstens müssen Pseudozeug:innen typischerweise eine Person aus der Gegenüberstellung wählen und können die Gegenüberstellung nicht zurückweisen (Doob & Kirshenbaum, 1973). Augenzeug:innen werden hingegen explizit darauf hingewiesen, dass die Gegenüberstellung zurückgewiesen werden soll, wenn keine der Personen wiedererkannt wird (Wells et al., 2020). Zweitens können Pseudozeug:innen im Gegensatz zu Augenzeug:innen keine gedächtnisbasierte Entscheidung treffen, da sie die tatbeteiligte Person nie zuvor gesehen haben. Während Augenzeug:innen die Gedächtnisrepräsentation der tatbeteiligten Person mit jedem Gesicht der Gegenüberstellung abgleichen (Wixted & Mickes, 2014), müssen Pseudozeug:innen die Bilder in der Gegenüberstellung miteinander vergleichen, um jenes Bild einer Person zu finden, das sich von den anderen Bildern abhebt. Im Gegensatz zu Augenzeug:innen werden Pseudozeug:innen demnach darauf aufmerksam gemacht, dass die tatverdächtige Person aus der Gegenüberstellung herausragen könnte, was sie möglicherweise besonders sensitiv für Hinweise einer Unfairness macht. Aufgrund dieser entscheidenden Unterschiede ist es fraglich, ob mithilfe des Pseudozeug:innen-Paradigmas die Fairness in echten Gegenüberstellungen valide vorgesagt werden kann. Das 2-HT EI-Modell bietet bei der Beantwortung dieser Frage einen großen Vorteil: Über Parameter b kann die Auftretenswahrscheinlichkeit des auf Unfairness basierten Auswahlprozesses einer tatverdächtigen Person direkt anhand der Daten der Augenzeug:innen gemessen werden.

In einer Serie von vier Experimenten wurde die Fairness von Gegenüberstellungen mit gemorphten und nicht-gemorphten Vergleichsbildern (fortan als gemorphte beziehungsweise nicht-gemorphte Gegenüberstellungen bezeichnet) erfasst. Mithilfe des Morphing-Verfahrens kann durch die digitale Kombination der Gesichtsbilder zweier Personen ein neues (gemorphtes) Gesichtsbild erstellt werden. In der polizeilichen Praxis bietet dieses Verfahren ein großes Potential. Zum einen ist die Auswahl geeigneter Vergleichsbilder, die der Beschreibung der tatbeteiligten Person genügen, durch limitierte Datenbanken in der Praxis häufig sehr herausfordernd (Steblay & Wells, 2020). Mithilfe des Morphing-Verfahrens kann die Auswahl an potentiellen Vergleichsbildern erweitert werden. Zum anderen können durch das Morphing-Verfahren die Persönlichkeitsrechte der Vergleichspersonen geschützt werden. Beispielsweise müssen die Vergleichsbilder in Nord-

rhein-Westfalen digital so verfremdet werden, dass die Person auf dem Bild zweifelsfrei nicht mehr zu erkennen ist (Ministerium des Innern des Landes Nordrhein-Westfalen, 2024). Allerdings können im Morphing-Prozess gewisse Artefakte (z. B. Geisterartefakte, verzerrte Kantenverläufe, Weichzeichnungseffekte) entstehen (Alley & Cunningham, 1991; Borghi et al., 2021), sodass das einzige nicht-gemorphte Bild der tatverdächtigen Person aus den anderen, gemorphten Bildern der Vergleichspersonen herausragen könnte. Die Abwesenheit von Morphing-Artefakten könnte somit ein entscheidender Hinweis auf die Identität der tatverdächtigen Person sein. In den nachfolgenden Experimenten wurde deshalb der Einfluss der Morphing-Manipulation auf die Fairness von Gegenüberstellungen untersucht (siehe Einzelarbeit 2). Um die Validität des Pseudozeug:innen-Paradigmas zu prüfen, wurde die Fairness von gemorphten und nicht-gemorphten Gegenüberstellungen zum einen über die Entscheidungen von Pseudozeug:innen mithilfe traditioneller Fairnessmaße erfasst (Experiment 1a). Zum anderen wurde mithilfe des Modells die Auftretenswahrscheinlichkeit des auf Unfairness basierten Auswahlprozesses einer tatverdächtigen Person direkt anhand der Daten der Augenzeug:innen gemessen (Experimente 1b bis 1d).

Experiment 1a

In Experiment 1a wurde ein klassisches Pseudozeug:innen-Paradigma verwendet, um die Fairness von gemorphten und nicht-gemorphten Gegenüberstellungen zu beurteilen. Die Teilnehmenden, die nur eine kurze verbale Beschreibung des Tathergangs erhalten hatten, wurden instruiert, das Bild zu wählen, das ihrer Meinung nach das Bild der tatverdächtigen Person sei. Sie wurden zusätzlich darauf hingewiesen, dass das Bild der tatverdächtigen Person aus den anderen Bildern der Gegenüberstellung herausragen könnte. Anschließend wurden nacheinander vier simultane Gegenüberstellungen präsentiert, in denen jeweils ein Bild der tatverdächtigen Person gleichzeitig mit fünf Vergleichsbildern dargeboten wurde. Der einen Hälfte der Teilnehmenden wurden nicht-gemorphte Gesichtsbilder von Vergleichspersonen aus einer Datenbank gezeigt. Der anderen Hälfte wurden gemorphte Gesichtsbilder von Vergleichspersonen präsentiert, die jeweils die biometrischen Informationen zweier Personen in einem Verhältnis von 50:50 abbildeten (für ein Beispiel siehe Abbildungen 2 und 3 in Einzelarbeit 2). Dafür wurden die 20 Gesichtsbilder der Vergleichspersonen der nicht-gemorphten Bedingung jeweils mit einem weiteren Gesichtsbild aus einer Datenbank gemorpht. Dem klassischen Pseudozeug:innen-Paradigma folgend mussten die Teilnehmenden eine Person wählen und konnten die Gegenüberstellungen nicht zurückweisen.

Aus der Verteilung der Wahlen auf die Personen in der Gegenüberstellung wurde die Fairness auf zwei traditionelle Arten ermittelt: Zum einen wurde die effektive Größe der gemorphten und nicht-gemorphten Gegenüberstellungen mithilfe von Tredoux's *E* berechnet (Tredoux, 1998), um eine Schätzung über die Anzahl plausibler Wahlalternativen zu erhalten. Zum anderen wurde der Anteil der Pseudozeug:innen, die die tatverdächtige Person in den gemorphten und nicht-gemorphten Gegenüberstellungen wählten, berechnet (Doob & Kirshenbaum, 1973). Die effektive

Größe der nicht-gemorphten Gegenüberstellungen war größer als die effektive Größe der gemorphten Gegenüberstellungen. Zudem war der Anteil der Pseudozeug:innen, die in den gemorphten Gegenüberstellungen die tatverdächtige Person wählten, signifikant höher als in den nicht-gemorphten Gegenüberstellungen.

Die Ergebnisse des Pseudozeug:innen-Paradigmas legten nahe, dass gemorphte Gegenüberstellungen unfairer waren als nicht-gemorphte Gegenüberstellungen. Basierend auf diesen Befunden würde man demnach davon abraten, Morphing-Techniken zur Erstellung von Vergleichsbildern in polizeilichen Ermittlungen einzusetzen. Offen bleibt jedoch, ob die Morphing-Manipulation die Entscheidungen echter Augenzeug:innen in gleichem Maße beeinflusst wie die Entscheidungen der Pseudozeug:innen. Im Gegensatz zu Augenzeug:innen (1) müssen Pseudozeug:innen eine Person der Gegenüberstellung auswählen und (2) werden darauf hingewiesen, dass eine Person möglicherweise aus der Gegenüberstellung herausragen könnte. Mithilfe des 2-HT EI-Modells wurde deshalb in den nachfolgenden Experimenten überprüft, ob die Morphing-Manipulation die Auftretenswahrscheinlichkeit des auf Unfairness basierten Auswahlprozesses einer tatverdächtigen Person bei Augenzeug:innen beeinflusst, abhängig davon, ob sich die Aufgabe bezüglich dieser beiden Merkmale unterscheidet oder nicht. Zunächst wurde die Aufgabe der Augenzeug:innen möglichst stark an die Aufgabe von Pseudozeug:innen angepasst, um zu prüfen, ob der Morphing-Effekt sich unter den spezifischen Bedingungen des Pseudozeug:innen-Paradigmas auch im Unfairnessparameter *b* zeigt. Außerdem wurden in den nachfolgenden Experimenten nicht nur simultane, sondern auch sequentielle Gegenüberstellungen präsentiert, in denen die Personen nacheinander dargeboten werden. Während das Pseudozeug:innen-Paradigma für die Messung der Fairness in simultanen Gegenüberstellungen entwickelt wurde (Corey et al., 1999; Lindsay et al., 1999), ist das 2-HT EI-Modell auch in der Lage, eine Unfairness zuungunsten der tatverdächtigen Person in sequentiellen Gegenüberstellungen zu erfassen (Winter et al., 2022).

Experiment 1b

In Experiment 1b wurde die Aufgabe der Augenzeug:innen möglichst stark an die Aufgabe von Pseudozeug:innen angepasst. Die Teilnehmenden sahen zunächst ein Video eines Tathergangs und wurden anschließend instruiert, die vier tatbeteiligten Personen in mehreren Gegenüberstellungen zu identifizieren. Wie in Experiment 1a wurden die Teilnehmenden darauf hingewiesen, dass die tatverdächtige Person möglicherweise aus den anderen Personen der Gegenüberstellung herausragen könnte. Zusätzlich wurden die Teilnehmenden instruiert, dass sich in jeder Gegenüberstellung mit hoher Wahrscheinlichkeit eine tatbeteiligte Person befindet, um den Anteil der Zurückweisungen möglichst stark zu reduzieren. Anschließend wurden den Teilnehmenden nacheinander vier simultane oder vier sequentielle Gegenüberstellungen präsentiert. Während zwei Gegenüberstellungen eine tatbeteiligte Person enthielten, beinhalteten die anderen zwei eine unschuldige tatverdächtige Person. Wie in Experiment 1a bestanden die Gegenüberstellungen aus

einem Bild einer tatverdächtigen Person und entweder fünf gemorphten oder nicht-gemorphten Vergleichsbildern. Es wurden die gleichen Vergleichsbilder wie in Experiment 1a verwendet.

Parallel zu Experiment 1a zeigte die modellbasierte Analyse eine signifikant höhere Wahrscheinlichkeit einer auf Unfairness basierten Auswahl einer tatverdächtigen Person (Parameter *b*) in gemorphten simultanen Gegenüberstellungen als in nicht-gemorphten simultanen Gegenüberstellungen. In sequentiellen Gegenüberstellungen war hingegen kein Effekt der Morphing-Manipulation zu beobachten. Dies steht im Einklang mit den Befunden bisheriger Studien, die zeigten, dass sequentielle Gegenüberstellungen der tatverdächtigen Person einen gewissen Schutz vor einer auf Unfairness basierten Auswahl bieten (Carlson et al., 2008; Lindsay et al., 1991; Steblay & Wells, 2020). Die Teilnehmenden können in sequentiellen Gegenüberstellungen die Gesichter nicht direkt nebeneinander vergleichen, sodass Merkmale, die die tatverdächtige Person hervorstechen lassen, nur schwer erkannt werden können.

Wenn den Augenzeug:innen also die Anwesenheit der tatbeteiligten Person suggeriert wurde und sie darauf hingewiesen wurden, dass die tatverdächtige Person möglicherweise aus der Gegenüberstellung herausragen könnte, stimmten die modellbasierten Schlussfolgerungen mit denen aus Experiment 1a überein: Simultane gemorphte Gegenüberstellungen waren unfairer als simultane nicht-gemorphte Gegenüberstellungen. Bislang ist jedoch unklar, ob sich der Morphing-Effekt lediglich unter den spezifischen Bedingungen des Pseudozeug:innen-Paradigmas in den Daten der Augenzeug:innen zeigt. Um dies zu prüfen, wurden in den nachfolgenden Experimenten die Instruktionen sukzessiv an die Standardinstruktionen für Augenzeug:innen angepasst.

Experiment 1c

In Experiment 1c wurde dasselbe Material und dieselbe Prozedur wie in Experiment 1b verwendet, jedoch mit folgender Ausnahme: Den Teilnehmenden wurde dieses Mal nicht suggeriert, dass sich die tatbeteiligte Person mit hoher Wahrscheinlichkeit in der Gegenüberstellung befindet. Stattdessen wurde eine zweiseitige Instruktion präsentiert, die sowohl die Wichtigkeit einer korrekten Auswahl der tatbeteiligten Person als auch die einer korrekten Zurückweisung einer Gegenüberstellung hervorhebt. Wie in den Experimenten 1a und 1b wurden die Teilnehmenden jedoch weiterhin darauf hingewiesen, dass die tatverdächtige Person möglicherweise aus den anderen Personen der Gegenüberstellung herausragen könnte.

Parallel zu Experiment 1b zeigte sich deskriptiv eine höhere Wahrscheinlichkeit einer auf Unfairness basierten Auswahl einer tatverdächtigen Person (Parameter *b*) in gemorphten simultanen Gegenüberstellungen als in nicht-gemorphten simultanen Gegenüberstellungen. Allerdings war dieser Unterschied kleiner als in Experiment 1b und statistisch nicht signifikant. Wie in Experiment 1b ließ sich auch hier kein Effekt der Morphing-Manipulation in sequentiellen Gegenüberstellungen feststellen.

Anders als in Experiment 1b, in dem die Anwesenheit der tatbeteiligten Person suggeriert wurde, fand sich bei Verwendung einer zweiseitigen Instruktion also keine signifikant erhöhte Wahrscheinlichkeit einer auf Unfairness basierten Auswahl in simultanen gemorphten Gegenüberstellungen. Es war zu erwarten, dass die Teilnehmenden verstärkt nach Merkmalen suchen, die die tatverdächtige Person aus der Gegenüberstellung hervorstechen lassen, wenn sie den Hinweis erhalten, dass die tatbeteiligte Person sich höchstwahrscheinlich in der Gegenüberstellung befindet. In Experiment 1d sollte nun überprüft werden, ob sich der Morphing-Effekt unter den Bedingungen einer sachgerecht durchgeführten Gegenüberstellung vollständig eliminieren lässt.

Experiment 1d

In Experiment 1d wurde dasselbe Material und dieselbe Prozedur wie in Experiment 1c verwendet, jedoch mit folgender Ausnahme: Die Teilnehmenden wurden nicht darauf hingewiesen, dass die tatverdächtige Person möglicherweise aus den anderen Personen der Gegenüberstellung herausragen könnte. Dadurch entsprachen die vorgelegten Instruktionen den Standardinstruktionen für Augenzeug:innen, bei denen weder die Anwesenheit der tatbeteiligten Person in der Gegenüberstellung suggeriert noch auf eine mögliche Unfairness hingewiesen wird (Wells et al., 2020).

Sowohl in den simultanen als auch in den sequentiellen Gegenüberstellungen zeigte sich wie in Experiment 1c kein signifikanter Unterschied in der Wahrscheinlichkeit einer auf Unfairness basierten Auswahl einer tatverdächtigen Person (Parameter *b*) zwischen gemorphten und nicht-gemorphten Gegenüberstellungen. Zudem verschwand der in Experiment 1c noch sichtbare deskriptive Unterschied zwischen gemorphten und nicht-gemorphten simultanen Gegenüberstellungen im Parameter *b* vollständig. Unter den Voraussetzungen einer sachgerecht durchgeführten Gegenüberstellung zeigte sich also keine Evidenz für einen Effekt der Morphing-Manipulation auf die Fairness in simultanen und sequentiellen Gegenüberstellungen.

Diskussion

Die Ergebnisse der Experimente 1a bis 1d demonstrieren, wie wichtig es ist, die Fairness einer Gegenüberstellung direkt anhand der Daten der Augenzeug:innen zu erfassen, statt sie indirekt über die Entscheidungen von Pseudozeug:innen zu schätzen. Basierend auf den Entscheidungen der Pseudozeug:innen hätte man voreilig annehmen müssen, dass die Verwendung gemorphter Vergleichsbilder zwangsläufig zu unfairen Gegenüberstellungen führt. Die modellbasierten Analysen der Daten der Augenzeug:innen zeigten jedoch, dass die Verwendung gemorphter Vergleichsbilder lediglich unter den spezifischen Bedingungen des Pseudozeug:innen-Paradigmas zu einer Unfairness führte, jedoch keinen Einfluss auf die Wahrscheinlichkeit einer auf Unfairness basierten Auswahl in sachgerecht durchgeführten Gegenüberstellungen hatte.

Die Ergebnisse stützen die Befunde anderer Studien, die ebenfalls Diskrepanzen zwischen den Daten von Pseudozeug:innen und Augenzeug:innen nachwiesen und damit die Nützlichkeit des Pseudozeug:innen-Paradigmas zur Beurteilung der Fairness von Gegenüberstellungen in Frage stellten (Lee et al., 2022; Lindsay et al., 1999; Steblay & Wells, 2020). Basierend auf den Ergebnissen der Experimente 1a bis 1d kann angenommen werden, dass zwei wesentliche Unterschiede in der Aufgabe von Pseudozeug:innen und Augenzeug:innen verantwortlich für diese divergierenden Befunde sind: Zum einen haben Augenzeug:innen die Möglichkeit, eine Gegenüberstellung zurückzuweisen, während Pseudozeug:innen üblicherweise gebeten werden, in jedem Fall eine Person auszuwählen. Zum anderen werden Pseudozeug:innen im Gegensatz zu Augenzeug:innen darauf aufmerksam gemacht, dass die tatverdächtige Person aus den anderen Personen der Gegenüberstellung herausragen könnte. Augenzeug:innen treffen hingegen eine gedächtnisbasierte Entscheidung, indem sie die Gedächtnisrepräsentation der tatbeteiligten Person mit jedem Gesicht der Gegenüberstellung abgleichen. Aufgrund dieser Unterschiede stellen die Entscheidungen von Pseudozeug:innen keine gute Basis dar, um die Entscheidungen der Augenzeug:innen vorherzusagen und damit die Fairness von Gegenüberstellungen abzuschätzen. Das 2-HT EI-Modell bietet erstmals die Möglichkeit, mithilfe von Parameter *b* die Wahrscheinlichkeit einer auf Unfairness basierten Auswahl einer tatverdächtigen Person direkt anhand der Daten der Augenzeug:innen zu erfassen – unabhängig davon, ob die Personen der Gegenüberstellung simultan oder sequentiell präsentiert werden.

Mithilfe des Modells kann nicht nur die Auftretenswahrscheinlichkeit des auf Unfairness basierten Auswahlprozesses gemessen werden, sondern es erlaubt zusätzlich die getrennte Erfassung von Detektions- und Rateprozessen, die den Entscheidungen von Augenzeug:innen zugrunde liegen. Dadurch lässt sich ein tiefer gehendes Verständnis zahlreicher Phänomene der Gegenüberstellungsforschung erreichen. Dies soll anhand einer weiteren konkreten Modellanwendung demonstriert werden, in der der Einfluss der Größe einer Gegenüberstellung auf die beobachtbaren Entscheidungen von Augenzeug:innen in Beiträge detektionsbasierter und nicht-detektionsbasierter Prozesse dekomponiert wurde.

Die Größe von Gegenüberstellungen

Wie aus dem in Abbildung 1 dargestellten 2-HT EI-Modell ersichtlich wird, beeinflusst nicht nur die Fairness von Gegenüberstellungen die Auswahlrate tatverdächtiger Personen, sondern auch die Größe von Gegenüberstellungen. Je mehr Personen in einer Gegenüberstellung präsentiert werden, desto geringer ist die Wahrscheinlichkeit, dass eine ratebasierte Auswahl zufällig auf die tatverdächtige Person fällt. Mathematisch ausgedrückt verhält sich die Wahrscheinlichkeit einer zufälligen Auswahl der tatverdächtigen Person umgekehrt proportional zur Größe *n* einer Gegenüberstellung ($1 \div n$). Abgesehen von diesem offensichtlichen Effekt größerer Gegenüberstellungen auf die zufällige Auswahlwahrscheinlichkeit tatverdächtiger Personen ist bislang unklar,

inwiefern die Anzahl der präsentierten Personen die latenten Prozesse, die den Entscheidungen von Augenzeug:innen zugrunde liegen, beeinflusst. Die Relevanz dieser Frage wird deutlich, wenn man sich die erheblichen Unterschiede in der empfohlenen Anzahl von Personen in einer Gegenüberstellung zwischen verschiedenen Ländern vor Augen führt. Zum Beispiel werden in den USA typischerweise sechs Personen präsentiert (Police Executive Research Forum, 2013), in Großbritannien werden den Augenzeug:innen hingegen meist neun Personen gezeigt (Home Office, 2017), und in Deutschland müssen es mindestens acht Personen sein (Bundesministerium des Innern und für Heimat, 2023). Ziel der Experimente 2a und 2b war es deshalb, die Effekte der Gegenüberstellungsgröße auf die zugrundeliegenden Prozesse der Entscheidungen von Augenzeug:innen mithilfe des 2-HT EI-Modells zu untersuchen (siehe Einzelarbeit 3). Das Modell bietet hier einen entscheidenden Vorteil: Die inverse Beziehung zwischen der zufälligen Auswahlwahrscheinlichkeit tatverdächtiger Personen und der Gegenüberstellungsgröße wird in Form einer festen Konstante ($1 \div n$) berücksichtigt, sodass die Modellparameter unabhängig davon geschätzt werden können.

In bisherigen Studien reduzierte das Hinzufügen von Personen in Gegenüberstellungen die Auswahlrate unschuldiger tatverdächtiger Personen, ging aber auch mit einer Reduktion der Auswahlrate tatbeteiligter Personen einher (für eine Metanalyse siehe Juncu & Fitzgerald, 2021). Basierend auf den reinen Antwortraten bleibt jedoch unklar, welche Prozesse zu diesem beobachtbaren Antwortmuster geführt haben. Die erhöhte Auswahlrate tatverdächtiger Personen in kleineren im Vergleich zu größeren Gegenüberstellungen könnte ausschließlich durch die erhöhte Auswahlwahrscheinlichkeit in kleineren Gegenüberstellungen erklärt werden, mit der eine ratebasierte Auswahl zufällig auf die tatverdächtige Person fällt. Allerdings könnte die erhöhte Auswahlrate tatbeteiligter Personen in kleineren Gegenüberstellungen auch das Ergebnis einer verbesserten Detektion der Anwesenheit der tatbeteiligten Person sein. Eine verbesserte Detektionsleistung in kleineren Gegenüberstellungen ist zu erwarten, da das Gesicht jeder zusätzlichen Vergleichsperson visuell verarbeitet werden muss und damit möglicherweise Ressourcen beansprucht werden, die ansonsten für die visuelle Verarbeitung des Gesichts der tatverdächtigen Person zur Verfügung gestanden hätten (Vredeveldt et al., 2011; Vredeveldt et al., 2015; Wais et al., 2010). In den nachfolgenden Experimenten wurde demnach getestet, ob kleinere im Vergleich zu größeren Gegenüberstellungen zu einer verbesserten Detektion der Anwesenheit der tatbeteiligten Person führen (Parameter *dP*) oder ob die Detektionsleistung von der Gegenüberstellungsgröße unbeeinflusst bleibt.

Mithilfe des 2-HT EI-Modells war es zusätzlich möglich, den Einfluss der Gegenüberstellungsgröße auf die Wahrscheinlichkeit einer ratebasierten Auswahl zu untersuchen. Die erhöhte Auswahlrate tatverdächtiger Personen in kleineren im Vergleich zu größeren Gegenüberstellungen könnte auf den ersten Blick suggerieren, dass die Wahrscheinlichkeit einer ratebasierten Auswahl mit steigender Anzahl der Personen in der Gegenüberstellung sinkt. Bei Betrachtung der restlichen zur Verfügung stehenden Datenkategorien wird jedoch deutlich, dass mit zunehmender

Größe der Gegenüberstellungen die Auswahlrate der Vergleichspersonen anstieg und die Rate der Zurückweisungen sank (Juncu & Fitzgerald, 2021). Dieses Antwortmuster lässt vermuten, dass der erhöhte Schutz tatverdächtiger Personen in größeren Gegenüberstellungen ausschließlich auf die inverse Beziehung zwischen der zufälligen Auswahlwahrscheinlichkeit tatverdächtiger Personen und der Gegenüberstellungsgröße zurückzuführen ist ($1 \div n$), während die Wahrscheinlichkeit einer ratebasierten Auswahl in größeren im Vergleich zu kleineren Gegenüberstellungen erhöht ist (Parameter g).

Experiment 2a

Das Ziel von Experiment 2a war es, den Einfluss der Gegenüberstellungsgröße auf die beobachtbaren Entscheidungen von Augenzeug:innen mithilfe des 2-HT EI-Modells in Beiträge detektionsbasierter und nicht-detektionsbasierter Prozesse zu dekomponieren. Wie in den Experimenten 1b bis 1d sahen die Teilnehmenden zunächst das Video des Tathergangs und sollten anschließend die tatbeteiligten Personen in vier Gegenüberstellungen identifizieren. Während die eine Hälfte der Teilnehmenden Gegenüberstellungen mit sechs Personen (eine tatverdächtige Person und fünf Vergleichspersonen) sah, wurden der anderen Hälfte Gegenüberstellungen mit drei Personen (eine tatverdächtige Person und zwei Vergleichspersonen) präsentiert. Für die Gegenüberstellungen mit drei Personen wurden die Gesichtsbilder der zwei Vergleichspersonen randomisiert aus den Gesichtsbildern der fünf Vergleichspersonen für jede Gegenüberstellung ausgewählt. Dadurch wurde garantiert, dass der entscheidende Unterschied zwischen den Bedingungen in der Anzahl der Personen und nicht in deren Identität lag. Die vier Gegenüberstellungen wurden entweder simultan oder sequentiell dargeboten. Zwei Gegenüberstellungen enthielten eine tatbeteiligte Person, die anderen zwei beinhalteten hingegen eine unschuldige tatverdächtige Person.

Die Wahrscheinlichkeit, die Anwesenheit der tatbeteiligten Person zu detektieren (Parameter dP), war in Gegenüberstellungen mit drei Personen signifikant höher als in Gegenüberstellungen mit sechs Personen. Wie erwartet war außerdem die Wahrscheinlichkeit einer ratebasierten Auswahl (Parameter g) in Gegenüberstellungen mit sechs Personen signifikant höher als in Gegenüberstellungen mit drei Personen. Um die Robustheit dieser Befunde zu prüfen (vgl. Open Science Collaboration, 2015), diente Experiment 2b als konzeptuelle Replikation von Experiment 2a.

Experiment 2b

In Experiment 2b wurde dasselbe Material und dieselbe Prozedur wie in Experiment 2a verwendet, jedoch mit folgender Ausnahme: Statt Gegenüberstellungen mit drei beziehungsweise sechs Personen, sahen die Teilnehmenden nun Gegenüberstellungen mit entweder zwei (eine tatverdächtige Person und eine Vergleichsperson) oder fünf Personen (eine tatverdächtige Person und vier Vergleichspersonen). Wie in Experiment 2a wurden die Gesichtsbilder der Vergleichsperso-

nen randomisiert aus einer Auswahl von fünf Gesichtsbildern von Vergleichspersonen für jede Gegenüberstellung ausgewählt.

Parallel zu Experiment 2a zeigte sich in Gegenüberstellungen mit zwei Personen eine signifikant höhere Wahrscheinlichkeit, die Anwesenheit der tatbeteiligten Person zu detektieren (Parameter *dP*), als in Gegenüberstellungen mit fünf Personen. Außerdem war die Wahrscheinlichkeit einer ratebasierten Auswahl (Parameter *g*) in Gegenüberstellungen mit fünf Personen signifikant höher als in Gegenüberstellungen mit zwei Personen. Damit konnten die Ergebnisse von Experiment 2a bestätigt werden.

Diskussion

In beiden Experimenten zeigte sich eine verbesserte Detektion der Anwesenheit der tatbeteiligten Person in kleineren als in größeren Gegenüberstellungen. Die modellbasierten Analysen konnten damit demonstrieren, dass die erhöhte Auswahlrate tatbeteiligter Personen in kleineren im Vergleich zu größeren Gegenüberstellungen (vgl. Juncu & Fitzgerald, 2021) nicht nur auf die erhöhte Wahrscheinlichkeit zurückzuführen ist, mit der eine ratebasierte Auswahl auf die tatverdächtige Person fällt. Stattdessen ist dieses Befundmuster auch das Ergebnis einer verbesserten Detektion der Anwesenheit der tatbeteiligten Person in kleineren im Vergleich zu größeren Gegenüberstellungen. Die verminderte Fähigkeit, die tatbeteiligte Person in größeren Gegenüberstellungen zu detektieren, könnte dadurch erklärt werden, dass mit zunehmender Anzahl an Vergleichspersonen immer mehr Ressourcen für die visuelle Verarbeitung der Gesichter der Vergleichspersonen benötigt werden, sodass immer weniger Ressourcen für die visuelle Verarbeitung des Gesichts der tatverdächtigen Person zur Verfügung stehen (Vredeveldt et al., 2011; Vredeveldt et al., 2015; Wais et al., 2010).

Wie erwartet zeigte sich in kleineren Gegenüberstellungen zusätzlich eine reduzierte Wahrscheinlichkeit einer ratebasierten Auswahl im Vergleich zu größeren Gegenüberstellungen (Parameter *g*). Jedes Gesicht einer Gegenüberstellung bietet die Möglichkeit einer ratebasierten Auswahl. Daher erscheint es vollkommen plausibel, dass die Wahrscheinlichkeit, mit der über die gesamte Gegenüberstellung hinweg eine Auswahl basierend auf Raten erfolgt, mit zunehmender Anzahl der Personen in der Gegenüberstellung ansteigt. Wenn allerdings eine ratebasierte Auswahl erfolgt, ist die bedingte Wahrscheinlichkeit, dass diese Auswahl zufällig auf die tatverdächtige Person fällt, in größeren Gegenüberstellungen geringer als in kleineren Gegenüberstellungen ($1 \div n$). Beispielsweise liegt die zufällige Auswahlwahrscheinlichkeit der tatverdächtigen Person in Gegenüberstellungen mit drei Personen bei ⅓, in Gegenüberstellungen mit sechs Personen hingegen nur bei ⅙. Da sich die Wahrscheinlichkeit einer ratebasierten Auswahl einer tatverdächtigen Person aus der Kombination beider Komponenten ergibt ($g \cdot (1 \div n)$), stellt sich nun die Frage, ob die Zunahme im Rateparameter (*g*) in größeren Gegenüberstellungen durch die reduzierte Auswahlwahrscheinlichkeit ($1 \div n$) kompensiert werden kann. Auf der Ebene des beobachtbaren Verhaltens zeigte

sich in beiden Experimenten trotz des erhöhten Rateparameters in den größeren Gegenüberstellungen eine verringerte Auswahlrate tatverdächtiger Personen im Vergleich zu den kleineren Gegenüberstellungen (siehe Tabelle 1 in Einzelarbeit 3). Dieses Befundmuster ist konsistent zu den Ergebnissen bisheriger Studien, die den Einfluss der Gegenüberstellungsgröße auf die Entscheidungen von Augenzeugen:innen untersuchten (Akan et al., 2020; Juncu & Fitzgerald, 2021; Meissner et al., 2005; Wooten et al., 2020).

Zusammenfassend ermöglichen die modellbasierten Analysen ein tieferes Verständnis der Vor- und Nachteile kleinerer und größerer Gegenüberstellungen: Kleinere Gegenüberstellungen verbesserten im Vergleich zu größeren die Detektion tatbeteiligter Personen und reduzierten die Wahrscheinlichkeit einer ratebasierten Auswahl. Wenn jedoch eine ratebasierte Auswahl erfolgt, bieten größere Gegenüberstellungen der tatverdächtigen Person einen höheren Schutz vor einer zufälligen Auswahl als kleinere Gegenüberstellungen.

Allgemeine Diskussion

In der vorliegenden Dissertation wurde ein neues multinomiales Verarbeitungsbaummodell vorgestellt, das die Erfassung der latenten Prozesse ermöglicht, die den Entscheidungen von Augenzeug:innen zugrunde liegen. Im Gegensatz zu den traditionellen Gütemaßen berücksichtigt das 2-HT EI-Modell dabei *alle* zur Verfügung stehenden Datenkategorien einer Gegenüberstellung: die korrekte Auswahl einer tatbeteiligten Person, die falsche Auswahl einer unschuldigen tatverdächtigen Person, die falsche Auswahl einer Vergleichsperson in Gegenüberstellungen mit und ohne tatbeteiligte Person und die falsche und korrekte Zurückweisung. Dadurch ist das Modell in der Lage, zwei detektionsbasierte Prozesse – die Detektion der An- oder Abwesenheit der tatbeteiligten Person – und zwei nicht-detektionsbasierte Prozesse – eine auf Unfairness oder eine auf Raten basierte Auswahl– getrennt voneinander zu erfassen. Bevor das 2-HT EI-Modell jedoch auf neue rechtspsychologische Fragestellungen angewendet werden konnte, musste zunächst nachgewiesen werden, dass die Modellparameter auch tatsächlich die Prozesse abbilden, die ihrer Beschreibung entsprechen. Zu Validierungszwecken wurden in der vorliegenden Dissertation deshalb für jeden der vier Modellparameter zwei bereits publizierte Datensätze reanalysiert, bei denen experimentelle Manipulationen verwendet wurden, die den betreffenden Parameter in vorhersagbarer Weise beeinflussen sollten. Die vier Modellparameter des 2-HT EI-Modells konnten dabei erfolgreich validiert werden: In allen acht modellbasierten Reanalysen konnten die manipulierten Prozesse in den betreffenden Parametern sensitiv reflektiert werden. Auf der Basis dieser Ergebnisse und der Validierungsbefunde von Winter et al. (2022), die in einer Serie von neuen Experimenten die Modellparameter ebenfalls erfolgreich validieren konnten, schien es gerechtfertigt, das 2-HT EI-Modell zur Beantwortung neuer rechtspsychologischer Fragestellungen einzusetzen. Die Chancen des Modells für die Gegenüberstellungsforschung wurden in der vorliegenden Dissertation anhand von zwei konkreten Anwendungen verdeutlicht.

In den Experimenten 1a bis 1d wurde überprüft, ob die Verwendung gemorphter Vergleichsbilder die Fairness von Gegenüberstellungen beeinflusst. Die Fairness von gemorphten und nicht-gemorphten Gegenüberstellungen wurde dabei zum einen über die Entscheidungen von Pseudozeug:innen mithilfe traditioneller Fairnessmaße erfasst. Zum anderen wurde mithilfe des Modells die Auftretenswahrscheinlichkeit des auf Unfairness basierten Auswahlprozesses einer tatverdächtigen Person direkt anhand der Daten der Augenzeug:innen gemessen. Basierend auf den Ergebnissen des Pseudozeug:innen-Paradigmas hätte man schlussfolgern müssen, dass die Verwendung gemorphter Vergleichsbilder zu unfairen Gegenüberstellungen führt. In den modellbasierten Analysen der Daten der Augenzeug:innen zeigte sich hingegen nur dann ein signifikanter Effekt der Morphing-Manipulation auf den Unfairnessparameter *b,* wenn (1) die Anwesenheit der tatbeteiligten Person suggeriert wurde und (2) die Augenzeug:innen darauf hingewiesen wurden, dass die tatverdächtige Person aus den anderen Personen der Gegenüberstellung herausragen könnte. Waren diese spezifischen Bedingungen des Pseudozeug:innen-Paradigmas nicht gegeben und wurde stattdessen eine Standardinstruktion für Gegenüberstellungen präsentiert, zeigte sich kein Effekt der Morphing-Manipulation auf die Wahrscheinlichkeit einer auf Unfairness basierten Auswahl einer tatverdächtigen Person. Diese Befunde demonstrieren die entscheidenden Unterschiede zwischen der Aufgabe von Pseudozeug:innen und Augenzeug:innen und unterstreichen damit die Wichtigkeit, die Fairness einer Gegenüberstellung direkt anhand der Daten der Augenzeug:innen zu messen. Die indirekte Erfassung der Fairness über die Entscheidungen von Pseudozeug:innen hätte in diesem Fall dazu geführt, dass eine eigentlich unproblematische Technik für die Erstellung von Gegenüberstellungen in der Praxis irrtümlich verworfen worden wäre.

Mithilfe des 2-HT EI-Modells kann nicht nur die Auftretenswahrscheinlichkeit des auf Unfairness basierten Auswahlprozesses gemessen werden, sondern es ermöglicht auch, andere Prozesse der Detektion und des Ratens sichtbar zu machen. So konnte in den Experimenten 2a und 2b der Einfluss der Gegenüberstellungsgröße auf die beobachtbaren Entscheidungen von Augenzeug:innen in Beiträge detektionsbasierter und nicht-detektionsbasierter Prozesse dekomponiert werden. Die modellbasierten Analysen zeigten eine verbesserte Detektion der Anwesenheit der tatbeteiligten Person mit abnehmender Gegenüberstellungsgröße. Zudem ließ sich in kleineren Gegenüberstellungen eine geringere Wahrscheinlichkeit einer ratebasierten Auswahl im Vergleich zu größeren Gegenüberstellungen feststellen. Wenn jedoch eine ratebasierte Auswahl erfolgt, ist die bedingte Wahrscheinlichkeit, dass diese Auswahl zufällig auf die tatverdächtige Person fällt, in größeren Gegenüberstellungen geringer als in kleineren Gegenüberstellungen. Aufgrund dieser reduzierten Auswahlwahrscheinlichkeit tatverdächtiger Personen in größeren Gegenüberstellungen resultierte auf der Ebene des beobachtbaren Verhaltens trotz des erhöhten Rateparameters eine geringere Auswahlrate tatverdächtiger Personen in größeren als in kleineren Gegenüberstellungen. Die modellbasierten Analysen ermöglichen damit ein tiefergehendes Verständnis über die Vor- und Nachteile kleinerer und größerer Gegenüberstellungen auf der Ebene

der zugrundeliegenden Prozesse. Diese neuen Erkenntnisse können wertvolle Impulse für die weitere Entwicklung von Gegenüberstellungsprozeduren liefern. Beispielsweise konnte bereits demonstriert werden, dass mithilfe von Instruktionen die Wahrscheinlichkeit einer ratebasierten Auswahl reduziert werden kann, ohne die Detektion der Anwesenheit der tatbeteiligten Person zu beeinflussen (Winter et al., 2022). Es könnte demnach überprüft werden, ob mithilfe solcher Instruktionen die Wahrscheinlichkeit einer ratebasierten Auswahl (Parameter g) in kleineren Gegenüberstellungen so stark reduziert werden kann, dass die erhöhte zufällige Auswahlwahrscheinlichkeit tatverdächtiger Personen ($1 \div n$) – und damit der Nachteil kleinerer Gegenüberstellungen – kompensiert werden kann.

Anhand der vorgestellten konkreten Anwendungsbeispiele konnten die vielfältigen Einsatzmöglichkeiten des 2-HT EI-Modells aufgezeigt werden. Allerdings weist das in Abbildung 1 dargestellte Modell eine offenkundige Grenze auf: Das 2-HT EI-Modell ist nicht in der Lage, die subjektiven Konfidenzurteile von Augenzeug:innen zu berücksichtigen, die typischerweise in der Forschung und auch in der Praxis nach einer Entscheidung erfasst werden (Wells et al., 2020). Tatsächlich bestimmen subjektive Konfidenzurteile maßgeblich das Gewicht, das den Entscheidungen von Augenzeug:innen im Gerichtsverfahren beigemessen wird (Sauer et al., 2017; Semmler et al., 2012), sodass die Möglichkeit einer systematischen Untersuchung dieser Urteile wünschenswert erscheint. In der Vergangenheit konnten bereits einige multinomiale Verarbeitungsbaummodelle erfolgreich um subjektive Konfidenzurteile erweitert werden (z. B. Bröder et al., 2013; Erdfelder & Buchner, 1998; Province & Rouder, 2012). Eine in diesem Sinne erweiterte Variante des 2-HT EI-Modells wird in Abbildung 2 vorgestellt. Das hier dargestellte erweiterte Modell ist für $n = 3$ Konfidenzkategorien vorgesehen, lässt sich aber problemlos für jede beliebige Anzahl von n Konfidenzkategorien anpassen. Im Unterschied zum originalen 2-HT EI-Modell beinhaltet das erweiterte Modell zusätzliche c-Parameter, die die bedingte Auftretenswahrscheinlichkeit der Konfidenzkategorie i (1 = niedrige Konfidenz, 2 = mittlere Konfidenz, 3 = hohe Konfidenz) innerhalb eines bestimmten latenten Zustands abbilden. Die Parameter sind wie folgt definiert:

- Die c_{di}-Parameter repräsentieren die Wahrscheinlichkeit, dass die Konfidenzkategorie i gewählt wird, sofern die An-oder Abwesenheit der tatbeteiligten Person detektiert wurde.
- Die c_{bi}-Parameter repräsentieren die Wahrscheinlichkeit, dass die Konfidenzkategorie i gewählt wird, sofern eine tatverdächtige Person basierend auf einer Unfairness in der Gegenüberstellung ausgewählt wurde.
- Die c_{gi}-Parameter repräsentieren die Wahrscheinlichkeit, dass die Konfidenzkategorie i gewählt wird, sofern eine Person basierend auf Raten ausgewählt wurde.
- Die c_{gri}-Parameter repräsentieren die Wahrscheinlichkeit, dass die Konfidenzkategorie i gewählt wird, sofern die Gegenüberstellung aus einem Zustand der Unsicherheit zurückgewiesen wurde.

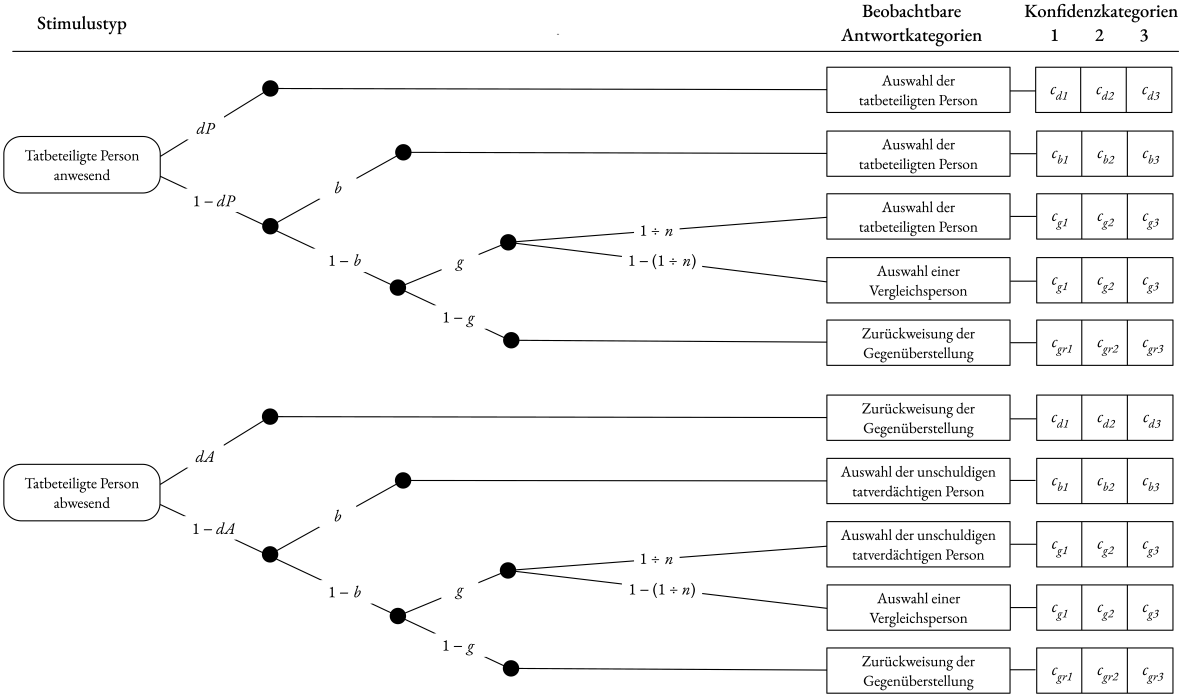

Abbildung 2. Graphische Darstellung des um Konfidenzurteile erweiterten 2-HT EI-Modells. Die zusätzlichen *c*-Parameter auf der rechten Seite repräsentieren die Wahrscheinlichkeit, dass die Konfidenzkategorie *i* (1 = niedrige Konfidenz, 2 = mittlere Konfidenz, 3 = hohe Konfidenz) gewählt wird, wenn a) die An- oder Abwesenheit der tatbeteiligten Person detektiert wurde (c_{di}), b) eine tatverdächtige Person basierend auf einer Unfairness in der Gegenüberstellung ausgewählt wurde (c_{bi}), c) eine Person basierend auf Raten ausgewählt wurde (c_{gi}) oder d) die Gegenüberstellung aus einem Zustand der Unsicherheit zurückgewiesen wurde (c_{gri}).

Aus der Literatur ist bereits bekannt, dass Konfidenzurteile nicht nur die Stärke der Gedächtnisrepräsentation reflektieren, sondern auch durch zahlreiche andere Faktoren wie beispielsweise den Instruktionen oder der Bezeichnung der Konfidenzkategorien beeinflusst werden können (z. B. Bröder et al., 2013; Leippe et al., 2009; Province & Rouder, 2012; Schwarz et al., 1991). Um Verzerrungen aufgrund von Antworttendenzen entgegenzuwirken, werden Teilnehmende beispielsweise häufig instruiert, die gesamte Breite der Konfidenzskala zu verwenden (Moosbrugger & Augustin, 2020). Allein diese Instruktion kann dazu führen, dass Personen auch im Zustand der Detektion eine geringe Konfidenz angeben (Province & Rouder, 2012). Aus diesem Grund beinhaltet das erweiterte 2-HT EI-Modell keine kritischen theoretischen Annahmen über die Skalennutzung im Zustand der Detektion oder Nicht-Detektion. Nimmt man jedoch zumindest einen gewissen Zusammenhang zwischen Konfidenz und Korrektheit an, so würde man im Zustand der Detektion höhere Konfidenzen erwarten als im Zustand der Nicht-Detektion (vgl. Bröder et al., 2013; Erdfelder & Buchner, 1998; Province & Rouder, 2012).

Um das in Abbildung 2 dargestellte Modell einem ersten Test zu unterziehen, wurden die Daten von Winter et al. (2022, Experiment 4), die explizit für die Validierung des *dA*-Parameters erhoben wurden, reanalysiert. Da in den Experimenten 1a bis 1d sowie 2a und 2b der Schwerpunkt auf den Parametern *b*, *dP* und *g* lag, soll abschließend der Parameter *dA* in den Fokus gerückt

werden. Parameter *dA* bietet erstmals die Möglichkeit, den Prozess der Detektion der Abwesenheit der tatbeteiligten Person zu erfassen. Wie zuvor beschrieben, wurde in der Studie von Winter et al. (2022, Experiment 4) die Einfachheit manipuliert, mit der die Abwesenheit der tatbeteiligten Person detektiert werden konnte. Der einen Hälfte der Teilnehmenden wurden in Gegenüberstellungen ohne tatbeteiligte Person ausschließlich Personen mit einem auffälligen Muttermal präsentiert, die somit leicht als tatbeteiligte Person ausgeschlossen werden konnten. Der anderen Hälfte der Teilnehmenden wurden Personen ohne Muttermal präsentiert, die als tatbeteiligte Person durchaus in Frage kamen. Zusätzlich wurde die Form der Gegenüberstellung manipuliert. Die Teilnehmenden sahen entweder simultane oder sequentielle Gegenüberstellungen. Nach jeder Entscheidung sollten die Teilnehmenden auf einer Skala von 0 % (*nicht sicher*) bis 100 % (*sehr sicher*) angeben, wie sicher sie sich in ihrer Entscheidung waren. Zur Reanalyse der Daten mithilfe des erweiterten 2-HT EI-Modells wurden die Konfidenzurteile in drei Konfidenzkategorien eingeteilt (Kategorie 1: 0 % bis 30 %, Kategorie 2: 31 % bis 70 %, Kategorie 3: 71 % bis 100 %).

Sowohl die Modellpassung als auch die Parameterschätzer wurden mit dem Programm *multiTree* (Moshagen, 2010) berechnet. Das Basismodell enthielt die gleichen Restriktionen wie in Winter et al. (2022): Da die Gegenüberstellungen aus sechs Personen bestanden, wurde die Konstante $1 \div n$ in allen vier Bedingungen auf .16667 gesetzt. Außerdem wurden die *b*-Parameter über alle vier Bedingungen hinweg gleichgesetzt, da sich die Unfairness zwischen den Bedingungen nicht unterscheiden sollte. Aus dem gleichen Grund wurden zusätzlich auch die c_{bi}-Parameter jeweils über die Bedingungen hinweg gleichgesetzt. Das erweiterte Modell mit diesen Restriktionen passte gut auf die Daten, $G^2(25) = 23.46$, $p = .551$. Die Schätzungen der zentralen Modellparameter (*dP*, *dA*, *b*, *g*) des erweiterten Modells waren nahezu identisch mit denen des bereits erfolgreich validierten originalen Modells (siehe Tabelle 1). Die Schätzungen für die *c*-Parameter sind in Abbildung 3 dargestellt[1]. Parallel zu den bisherigen Befunden in der Literatur (Bröder et al., 2013; Erdfelder & Buchner, 1998; Province & Rouder, 2012) wurden im Zustand der Detektion durchschnittlich höhere Konfidenzen angegeben als im Zustand der Nicht-Detektion. Während im Zustand der Detektion erwartungsgemäß am häufigsten die höchste Konfidenzkategorie gewählt wurde, gaben Teilnehmende im Zustand der Nicht-Detektion am häufigsten eine mittlere Konfidenz an. Wie bereits aus Abbildung 3 ersichtlich wird, hatte die Einfachheit, mit der eine Gegenüberstellung ohne tatbeteiligte Person zurückzuweisen war, weder einen signifikanten Einfluss auf die c_{di}-Parameter, $\Delta G^2(4) = 1.57$, $p = .815$, noch auf die c_{gi}-Parameter, $\Delta G^2(4) = 1.37$, $p = .849$, oder die c_{gri}-Parameter, $\Delta G^2(4) = 1.53$, $p = .822$. Dies war zu erwarten, da die Validierungsmanipulation ausschließlich einen Einfluss darauf haben sollte, die Abwesenheit der tatbeteiligten Person zu detektieren, nicht aber die Verteilung der Konfidenzen innerhalb eines latenten Zustandes der Detektion oder Nicht-Detektion beeinflussen sollte (Province & Rouder,

[1] Da in dieser Studie nahezu perfekt faire Gegenüberstellungen verwendet wurden, fielen extrem wenige Daten in die Kategorien, die für die Berechnung der c_{bi}- Parameter entscheidend sind. Da die c_{bi}- Parameter auf dieser schwachen Datenlage nicht aussagekräftig geschätzt werden können, wird hier auf die Darstellung dieser Parameter verzichtet.

2012). Die Form der Gegenüberstellung hatte jedoch einen signifikanten Effekt auf die c_{gi}-Parameter, $\Delta G^2(4) = 55.01$, $p < .001$, und die c_{gri}-Parameter, $\Delta G^2(4) = 23.31$, $p < .001$, nicht aber auf die c_{di}-Parameter, $\Delta G^2(4) = 3.26$, $p = .516$. Im Zustand der Nicht-Detektion führten sequentielle Gegenüberstellungen zu höheren Konfidenzen als simultane Gegenüberstellungen. Dieses Befundmuster stimmt mit den Ergebnissen von Dobolyi und Dodson (2013) überein, die ebenfalls höhere Konfidenzen für falsche Entscheidungen in sequentiellen als in simultanen Gegenüberstellungen fanden. Die Form der Gegenüberstellung hatte auch dort hingegen keinen Einfluss auf die Konfidenzen bei korrekten Entscheidungen.

	Simultane Gegenüberstellungen		Sequentielle Gegenüberstellungen	
	Schwer zurückzuweisen	Einfach zurückzuweisen	Schwer zurückzuweisen	Einfach zurückzuweisen
Parameter *dP*				
Originales Modell	.30 (.03)	.32 (.03)	.29 (.03)	.23 (.03)
Erweitertes Modell	.29 (.03)	.31 (.03)	.28 (.03)	.22 (.03)
Parameter *b*				
Originales Modell		.00 (.01)		
Erweitertes Modell		.01 (.01)		
Parameter *g*				
Originales Modell	.52 (.03)	.51 (.03)	.71 (.03)	.73 (.03)
Erweitertes Modell	.52 (.02)	.52 (.03)	.71 (.03)	.71 (.03)
Parameter *dA*				
Originales Modell	.02 (.08)	.13 (.08)	.00 (.05)	.20 (.04)
Erweitertes Modell	.03 (.04)	.15 (.05)	.01 (.04)	.18 (.04)

Tabelle 1. Schätzer der zentralen Modellparameter des originalen 2-HT EI-Modells und des um Konfidenzurteile erweiterten Modells für die in Winter et al. (2022, Experiment 4) berichteten Daten als Funktion der Form der Gegenüberstellung (simultan vs. sequentiell) und der Einfachheit, mit der eine Gegenüberstellung ohne tatbeteiligte Person zurückzuweisen war (schwer zurückzuweisen vs. einfach zurückzuweisen). Die Werte in Klammern repräsentieren die Standardfehler. *dP* = Wahrscheinlichkeit, die Anwesenheit der tatbeteiligten Person zu detektieren; *b* = Wahrscheinlichkeit einer auf Unfairness basierten Auswahl der tatverdächtigen Person; *g* = Wahrscheinlichkeit einer ratebasierten Auswahl; *dA* = Wahrscheinlichkeit, die Abwesenheit der tatbeteiligten Person zu detektieren.

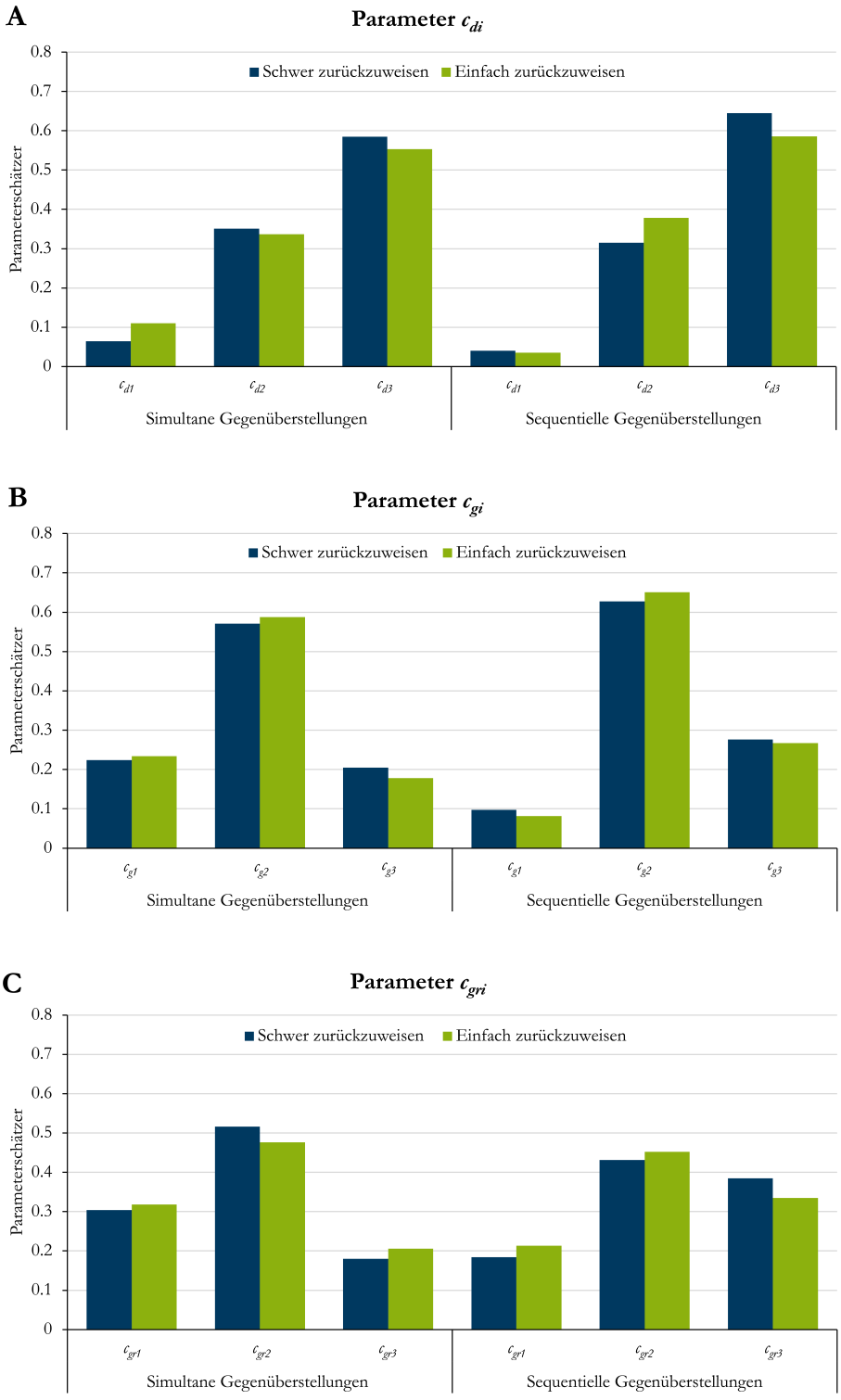

Abbildung 3. Schätzer der *c*-Parameter des um Konfidenzurteile erweiterten 2-HT EI-Modells für die in Winter et al. (2022, Experiment 4) berichteten Daten als Funktion der Form der Gegenüberstellung (simultan vs. sequentiell) und der Einfachheit, mit der eine Gegenüberstellung ohne tatbeteiligte Person zurückzuweisen war (schwer zurückzuweisen vs. einfach zurückzuweisen). Die *c*-Parameter repräsentieren die Wahrscheinlichkeit, dass die Konfidenzkategorie *i* (1 = niedrige Konfidenz; 2 = mittlere Konfidenz, 3 = hohe Konfidenz) gewählt wird, wenn (A) die An-oder Abwesenheit der tatbeteiligten Person detektiert wurde (c_{di}), (B) eine Person basierend auf Raten ausgewählt wurde (c_{gi}) oder (C) die Gegenüberstellung aus einem Zustand der Unsicherheit zurückgewiesen wurde (c_{gri}).

Angesichts dieser Ergebnisse hat das erweiterte 2-HT EI-Modell eine erste Bewährungsprobe bestanden. Der nächste notwendige Schritt im Rahmen der Modellvalidierung ist jedoch die experimentelle Validierung der c-Parameter. Zu diesem Zweck wird eine Manipulation benötigt, die sich auf die Konfidenzurteile der Teilnehmenden auswirkt, jedoch idealerweise keinen Einfluss auf die anderen zentralen Modellparameter hat. Bröder et al. (2013) beispielsweise manipulierten die Bezeichnungen der Endpunkte der Skala (*absolut sicher [unsicher]* vs. *ziemlich sicher [unsicher]*), um selektiv die Konfidenzurteile der Teilnehmenden zu beeinflussen. Nur wenn es zukünftig gelingt, die Validität der c-Parameter erfolgreich nachzuweisen, kann das erweiterte 2-HT EI-Modell für weitere Anwendungen empfohlen werden.

Fazit

Zusammenfassend wurde in dieser Arbeit ein multinomiales Verarbeitungsbaummodell vorgestellt, das unter Berücksichtigung aller Datenkategorien einer Gegenüberstellung die Erfassung der latenten Prozesse ermöglicht, die den Entscheidungen von Augenzeug:innen zugrunde liegen. In acht Reanalysen bereits publizierter Daten konnten die Modellparameter erfolgreich validiert werden. Damit kann das 2-HT EI-Modell zur Beantwortung inhaltlich relevanter Forschungsfragen eingesetzt werden. Die Chancen des modellbasierten Ansatzes für zukünftige Forschungsvorhaben wurden anhand von zwei konkreten Fragestellungen zur Fairness und zur Größe von Gegenüberstellungen demonstriert. Ausblickend auf zukünftige Forschung wurde abschließend eine erweiterte Variante des Modells vorgeschlagen, die eine Berücksichtigung der Konfidenzurteile von Augenzeug:innen ermöglicht. Insgesamt eröffnet das 2-HT EI-Modell neue Perspektiven für die Analyse der Entscheidungen von Augenzeug:innen in Gegenüberstellungen, die die Forschung gewinnbringend vorantreiben können.

Literatur

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

Alley, T. R., & Cunningham, M. R. (1991). Article commentary: Averaged faces are attractive, but very attractive faces are not average. *Psychological Science, 2*(2), 123–125. https://doi.org/10.1111/j.1467-9280.1991.tb00113.x

Batchelder, W. H., & Batchelder, E. (2008). Metacognitive guessing strategies in source monitoring. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 211–244). Psychology Press.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*(1), 57–86. https://doi.org/10.3758/BF03210812

Bayen, U. J., & Kuhlmann, B. G. (2011). Influences of source–item contingency and schematic knowledge on source monitoring: Tests of the probability-matching account. *Journal of Memory and Language, 64*(1), 1–17. https://doi.org/10.1016/j.jml.2010.09.001

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

Bell, R., Menne, N. M., Mayer, C., & Buchner, A. (2023). On the advantages of using AI-generated images of filler faces for creating fair lineups [Manuscript submitted for publication].

Borghi, G., Franco, A., Graffieti, G., & Maltoni, D. (2021). Automated artifact retouching in morphed images with attention maps. *IEEE Access, 9*, 136561–136579. https://doi.org/10.1109/ACCESS.2021.3117718

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

Brigham, J. C., Meissner, C. A., & Wasserman, A. W. (1999). Applied issues in the construction and expert assessment of photo lineups. *Applied Cognitive Psychology, 13*(S1), S73–S92. https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+<S73::AID-ACP631>3.0.CO;2-4

Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory, 21*(8), 916–944. https://doi.org/10.1080/09658211.2013.767348

Bröder, A., & Meiser, T. (2007). Measuring source memory. *Zeitschrift für Psychologie/Journal of Psychology, 215*(1), 52–60. https://doi.org/10.1027/0044-3409.215.1.52

Bundesministerium des Innern und für Heimat. (2023, März). *Richtlinien für das Strafverfahren und das Bußgeldverfahren (RiStBV)*. https://www.verwaltungsvorschriften-im-internet.de/bsvwvbund_28032023_BMJRB3313104000060001.htm

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

Clark, S. E. (2005). A re-examination of the effects of biased lineup instructions in eyewitness identification. *Law and Human Behavior, 29*(5), 575–604. https://doi.org/10.1007/s10979-005-7121-1

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

Corey, D., Malpass, R. S., & McQuiston, D. E. (1999). Parallelism in eyewitness and mock witness identifications. *Applied Cognitive Psychology, 13*(S1), S41–S58. https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+<S41::AID-ACP632>3.0.CO;2-A

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

Doob, A. N., & Kirshenbaum, H. M. (1973). Bias in police lineups – Partial remembering. *Journal of Police Science and Administration, 1*(3), 287–293.

Ehrenberg, K., & Klauer, K. C. (2005). Flexible use of source information: Processing components of the inconsistency effect in person memory. *Journal of Experimental Social Psychology, 41*(4), 369–387. https://doi.org/10.1016/j.jesp.2004.08.001

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie/Journal of Psychology, 217*(3), 108–124. https://doi.org/10.1027/0044-3409.217.3.108

Erdfelder, E., & Buchner, A. (1998). Comment: Process-dissociation measurement models: Threshold theory or detection theory? *Journal of Experimental Psychology: General, 127*(1), 83–96. https://doi.org/10.1037/0096-3445.127.1.83

Erdfelder, E., Cüpper, L., Auer, T.-S., & Undorf, M. (2007). The four-states model of memory retrieval experiences. *Zeitschrift für Psychologie/Journal of Psychology, 215*(1), 61–71. https://doi.org/10.1027/0044-3409.215.1.61

Fitzgerald, R. J., Oriet, C., & Price, H. L. (2015). Suspect filler similarity in eyewitness lineups: A literature review and a novel methodology. *Law and Human Behavior, 39*(1), 62–74. https://doi.org/10.1037/lhb0000095

Flowe, H. D., & Humphries, J. E. (2011). An examination of criminal face bias in a random sample of police lineups. *Applied Cognitive Psychology, 25*(2), 265–273. https://doi.org/10.1002/acp.1673

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

Havard, C., Laybourn, P., & Klecha, B. (2017). The mystery man can increase the reliability of eyewitness identifications for older adult witnesses. *Journal of Police and Criminal Psychology, 32*(3), 214–224. https://doi.org/10.1007/s11896-016-9214-9

Havard, C., & Memon, A. (2013). The mystery man can help reduce false identification for child witnesses: Evidence from video line-ups. *Applied Cognitive Psychology, 27*(1), 50–59. https://doi.org/10.1002/acp.2870

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 302–313. https://doi.org/10.1037/0278-7393.21.2.302

Home Office. (2017, February). *Police and criminal evidence act 1984 (PACE) code D*. https://assets.-publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/903812/pace-code-d-2017.pdf

Humphries, J. E., Holliday, R. E., & Flowe, H. D. (2012). Faces in motion: Age-related changes in eyewitness identification performance in simultaneous, sequential, and elimination video lineups. *Applied Cognitive Psychology, 26*(1), 149–158. https://doi.org/10.1002/acp.1808

Innocence Project. (2020, April). *How eyewitness misidentification can send innocent people to prison*. https://innocenceproject.org/how-eyewitness-misidentification-can-send-innocent-people-to-prison/

Juncu, S., & Fitzgerald, R. J. (2021). A meta-analysis of lineup size effects on eyewitness identification. *Psychology, Public Policy, and Law, 27*(3), 295–315. https://doi.org/10.1037/law0000311

Karageorge, A., & Zajac, R. (2011). Exploring the effects of age and delay on children's person identifications: Verbal descriptions, lineup performance, and the influence of wildcards. *British Journal of Psychology, 102*(2), 161–183. https://doi.org/10.1348/000712610X507902

Keast, A., Brewer, N., & Wells, G. L. (2007). Children's metacognitive judgments in an eyewitness identification task. *Journal of Experimental Child Psychology, 97*(4), 286–314. https://doi.org/10.1016/j.jecp.2007.01.007

Kroneisen, M., & Heck, D. W. (2020). Interindividual differences in the sensitivity for consequences, moral norms, and preferences for inaction: Relating basic personality traits to the CNI model. *Personality and Social Psychology Bulletin, 46*(7), 1013–1026. https://doi.org/10.1177/0146167219893991

Küppers, V., & Bayen, U. J. (2014). Inconsistency effects in source memory and compensatory schema-consistent guessing. *Quarterly Journal of Experimental Physiology, 67*(10), 2042–2059. https://doi.org/10.1080/17470218.2014.904914

Lampinen, J. M., Race, B., Wolf, A. P., Phillips, P., Moriarty, N., & Smith, A. M. (2020). Comparing detailed and less detailed pre-lineup instructions. *Applied Cognitive Psychology, 34*(2), 409–424. https://doi.org/10.1002/acp.3627

Lee, J., Mansour, J. K., & Penrod, S. D. (2022). Validity of mock-witness measures for assessing lineup fairness. *Psychology, Crime & Law, 28*(3), 215–245. https://doi.org/10.1080/1068316X.2021.1905811

Leippe, M. R., Eisenstadt, D., & Rauch, S. M. (2009). Cueing confidence in eyewitness identifications: Influence of biased lineup instructions and pre-identification memory feedback under varying lineup conditions. *Law and Human Behavior, 33*(3), 194–212. https://doi.org/10.1007/s10979-008-9135-y

Lindsay, R. C. L., Lea, J. A., Nosworthy, G. J., Fulford, J. A., Hector, J., LeVan, V., & Seabrook, C. (1991). 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. C. L., Smith, S. M., & Pryke, S. (1999). Measures of lineup fairness: Do they postdict identification accuracy? *Applied Cognitive Psychology, 13*(S1), S93–S107. https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+<S93::AID-ACP633>3.0.CO;2-X

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

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

Malpass, R. S., Tredoux, C. G., & McQuiston-Surrett, D. E. (2007). Lineup construction and lineup fairness. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology: Memory for people* (Vol. 2, pp. 155–178). Lawrence Erlbaum Associates.

Mansour, J. K., Beaudry, J. L., Kalmet, N., Bertrand, M. I., & Lindsay, R. C. L. (2017). Evaluating lineup fairness: Variations across methods and measures. *Law and Human Behavior, 41*(1), 103–115. https://doi.org/10.1037/lhb0000203

Meiser, T., & Bröder, A. (2002). Memory for multidimensional source information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(1), 116–137. https://doi.org/10.1037/0278-7393.28.1.116

Meiser, T., Sattler, C., & Von Hecker, U. (2007). Metacognitive inferences in source memory judgements: The role of perceived differences in item recognition. *Quarterly Journal of Experimental Psychology, 60*(7), 1015–1040. https://doi.org/10.1080/17470210600875215

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 and Cognition, 33*(5), 783–792. https://doi.org/10.3758/BF03193074

Memon, A., Hope, L., & Bull, R. (2003). Exposure duration: Effects on eyewitness accuracy and confidence. *British Journal of Psychology, 94*(3), 339–354. https://doi.org/10.1348/000712603767876262

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

Ministerium des Innern des Landes Nordrhein-Westfalen. (2024, Januar). *Wahllichtbildvorlage im Strafverfahren*. https://recht.nrw.de/lmi/owa/br_bes_text?anw_nr=1&bes_id=9147&aufgehoben=N

Moosbrugger, H., & Augustin, K. (2020). *Testtheorie und Fragebogenkonstruktion* (3. Aufl.). Springer. https://doi.org/10.1007/978-3-662-61532-4

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42*(1), 42–54. https://doi.org/10.3758/BRM.42.1.42

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

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

Police Executive Research Forum. (2013, March). *A national survey of eyewitness identification procedures in law enforcement agencies*. https://www.policeforum.org/assets/docs/Free_Online_Documents/Eyewitness_Identification/a%20national%20survey%20of%20eyewitness%20identification%20procedures%20in%20law%20enforcement%20agencies%202013.pdf

Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences, 109*(36), 14357–14362. https://doi.org/10.1073/pnas.1103880109

Riefer, D. M., Hu, X., & Batchelder, W. H. (1994). Response strategies in source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(3), 680–693. https://doi.org/10.1037/0278-7393.20.3.680

Sauer, J. D., Palmer, M. A., & Brewer, N. (2017). Mock-juror evaluations of traditional and ratings-based eyewitness identification evidence. *Law and Human Behavior, 41*(4), 375–384. https://doi.org/10.1037/lhb0000235

Schmidt, O., Erdfelder, E., & Heck, D. W. (2023). How to develop, test, and extend multinomial processing tree models: A tutorial. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000561

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*(4), 570–582. https://doi.org/10.1086/269282

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

Smith, A. M. (2020). Why do mistaken identification rates increase when either witnessing or testing conditions get worse? *Journal of Applied Research in Memory and Cognition, 9*(4), 495–507. https://doi.org/10.1016/j.jarmac.2020.08.002

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

Smith, R. E., & Bayen, U. J. (2004). A multinomial model of event-based prospective memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(4), 756–777. https://doi.org/10.1037/0278-7393.30.4.756

Steblay, N. K., & Wells, G. L. (2020). Assessment of bias in police lineups. *Psychology, Public Policy, and Law, 26*(4), 393–412. https://doi.org/10.1037/law0000287

Tredoux, C. G. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior, 22*(2), 217–237. https://doi.org/10.1023/A:1025746220886

Vredeveldt, A., Hitch, G. J., & Baddeley, A. D. (2011). Eyeclosure helps memory by reducing cognitive load and enhancing visualisation. *Memory and Cognition, 39*(7), 1253–1263. https://doi.org/10.3758/s13421-011-0098-8

Vredeveldt, A., Tredoux, C. G., Nortje, A., Kempen, K., Puljević, C., & Labuschagne, G. N. (2015). A field evaluation of the Eye-Closure Interview with witnesses of serious crimes. *Law and Human Behavior, 39*(2), 189–197. https://doi.org/10.1037/lhb0000113

Wais, P. E., Rubens, M. T., Boccanfuso, J., & Gazzaley, A. (2010). Neural mechanisms underlying the impact of visual distraction on retrieval of long-term memory. *Journal of Neuroscience, 30*(25), 8541–8550. https://doi.org/10.1523/JNEUROSCI.1478-10.2010

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. C. L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88*(3), 776–784. https://doi.org/10.1037/0033-2909.88.3.776

Wells, G. L., & Luus, C. A. E. (1990). Police lineups as experiments: Social methodology as a framework for properly conducted lineups. *Personality and Social Psychology Bulletin, 16*(1), 106–117. https://doi.org/10.1177/01461672901610

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

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (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., Smith, A. M., & Smalarz, L. (2015). ROC analysis of lineups obscures information that is critical for both theoretical understanding and applied purposes. *Journal of Applied Research in Memory and Cognition, 4*(4), 324–328. https://doi.org/10.1016/j.jarmac.2015.08.010

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*(1), 8–14. https://doi.org/10.1016/j.jarmac.2014.07.003

Wilcock, R., & Bull, R. (2010). Novel lineup methods for improving the performance of older eyewitnesses. *Applied Cognitive Psychology, 24*(5), 718–736. https://doi.org/10.1002/acp.1582

Winter, K., Menne, N. M., Bell, R., & Buchner, A. (2022). Experimental validation of a multinomial processing tree model for analyzing eyewitness identification decisions. *Scientific Reports, 12*, 15571. https://doi.org/10.1038/s41598-022-19513-w

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

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

Zajac, R., & Karageorge, A. (2009). The wildcard: A simple technique for improving children's target-absent lineup performance. *Applied Cognitive Psychology, 23*(3), 358–368. https://doi.org/10.1002/acp.1511

Einzelarbeiten

Einzelarbeit 1

Die Einzelarbeit 1 beinhaltet die Reanalysen 1a, 1b, 2a, 2b, 3a, 3b, 4a und 4b.

Menne, N. M., Winter, K., Bell, R., & Buchner, A. (2022). A validation of the two-high threshold eyewitness identification model by reanalyzing published data. *Scientific Reports*, *12*, 13379. https://doi.org/10.1038/s41598-022-17400-y

Einzelarbeit 2

Die Einzelarbeit 2 beinhaltet die Experimente 1a, 1b, 1c und 1d.

Menne, N. M., Winter, K., Bell, R., & Buchner, A. (2023). Measuring lineup fairness from eyewitness identification data using a multinomial processing tree model. *Scientific Reports*, *13*, 6290. https://doi.org/10.1038/s41598-023-33101-6

Einzelarbeit 3

Die Einzelarbeit 3 beinhaltet die Experimente 2a und 2b.

Menne, N. M., Winter, K., Bell, R., & Buchner, A. (2023). The effects of lineup size on the processes underlying eyewitness decisions. *Scientific Reports*, *13*, 17190. https://doi.org/10.1038/s41598-023-44003-y

scientific reports

Check for updates

OPEN

A validation of the two-high threshold eyewitness identification model by reanalyzing published data

Nicola Marie Menne✉, Kristina Winter, Raoul Bell & Axel Buchner

The two-high threshold (2-HT) eyewitness identification model serves as a new measurement tool to measure the latent cognitive processes underlying eyewitness identification performance. By simultaneously taking into account correct culprit identifications, false innocent-suspect identifications, false filler identifications in culprit-present and culprit-absent lineups as well as correct and false lineup rejections, the model capitalizes on the full range of data categories that are observed when measuring eyewitness identification performance. Thereby, the model is able to shed light on detection-based and non-detection-based processes underlying eyewitness identification performance. Specifically, the model incorporates parameters for the detection of culprit presence and absence, biased selection of the suspect and guessing-based selection among the lineup members. Here, we provide evidence of the validity of each of the four model parameters by applying the model to eight published data sets. The data sets come from studies with experimental manipulations that target one of the underlying processes specified by the model. Manipulations of encoding difficulty, lineup fairness and pre-lineup instructions were sensitively reflected in the parameters reflecting culprit-presence detection, biased selection and guessing-based selection, respectively. Manipulations designed to facilitate the rejection of culprit-absent lineups affected the parameter for culprit-absence detection. The reanalyses of published results thus suggest that the parameters sensitively reflect the manipulations of the processes they were designed to measure, providing support of the validity of the 2-HT eyewitness identification model.

The lineup procedure is an essential tool for assessing eyewitness identifications. In a lineup, a suspect is presented among fillers (known distractors that are not suspected of having committed the crime) to an eyewitness to test the hypothesis that the suspect is the culprit against the hypothesis that the suspect is innocent. Although eyewitness identifications can be a powerful and indispensable form of evidence, the problem of misidentifications has been well documented through DNA-based exonerations[1]. When the lineup includes the culprit (*culprit-present lineup*), the witness may correctly identify the culprit (*correct culprit identification*), but there is also the risk of an incorrect response in that the witness may identify a filler (*false filler identification*) or reject the lineup (*false lineup rejection*). When the suspect is innocent (*culprit-absent lineup*), the witness may either correctly reject the lineup (*correct lineup rejection*) or incorrectly identify the innocent suspect (*false innocent-suspect identification*) or a filler (*false filler identification*). An important goal of eyewitness identification research is to understand the latent cognitive processes underlying these decisions.

When two lineup procedures are compared, the simplest case is that one procedure is clearly superior to the other by yielding both a higher rate of correct culprit identifications in culprit-present lineups and a higher rate of correct lineup rejections of culprit-absent lineups. From such a data pattern, one may conclude that the superior lineup procedure provides better conditions for the process of detecting the culprit's face in the lineup. However, an increase in the correct culprit identifications is often accompanied by an increase in the false identifications of innocent suspects[2]. This demonstrates that understanding eyewitness identification performance in lineup procedures is complex because the selection of a suspect may not only be caused by the detection of the culprit but also by the biased selection of a suspect who stands out from the fillers, by guessing-based selection among

Department of Experimental Psychology, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. ✉email: Nicola.marie.menne@hhu.de

2 × 2 data structure of the standard signal-detection task			
	Identification		Rejection
Signal present	Correct identification		False rejection
Signal absent	False identification		Correct rejection
2 × 3 data structure of the typical eyewitness identification task			
	Suspect identification	Filler identification	Rejection
Culprit present	Correct culprit identification	False filler identification	False lineup rejection
Culprit absent	False innocent-suspect identification	False filler identification	Correct lineup rejection

Table 1. Comparison of the data structures of the standard signal-detection task and the eyewitness identification task when confronted with a typical lineup.

the lineup members or by any combination of these processes. To disentangle the contributions of these different processes to eyewitness identification performance, it is useful to apply a measurement model to the eyewitness identification data. Here, we introduce a novel *two-high threshold (2-HT) eyewitness identification model* for measuring the processes involved in eyewitness identification decisions. The aim of the present article is to test the validity of this model by reanalyzing published data. In an accompanying validation study[3], we applied the model to novel data that we had collected specifically for the purpose of testing the model's validity. We consider these two approaches to model validation as complementary: A model should not only be able to account for novel data generated with the model in mind but also for published data of other research groups. To anticipate, both approaches support the validity of the model by showing that the model's parameters sensitively respond to manipulations of the processes they were designed to measure, suggesting that the model parameters sensitively reflect these processes. Before we describe the 2-HT eyewitness identification model in more detail, we first provide a brief overview of the important discussion about the strengths and weaknesses of different methods for analyzing lineup data that has informed the development of the new measurement tool.

Measures of eyewitness identification performance in previous research

For some decades, eyewitness identification performance has been measured by using the *diagnosticity ratio* (or other measures of probative value), which is defined as the ratio of the proportion of correct culprit identifications to the proportion of false innocent-suspect identifications[4]. Larger ratios indicate a higher likelihood that an identified suspect is guilty[5], which may be interpreted to suggest that a lineup procedure that consistently generates a higher diagnosticity ratio should be preferred over one that does not. However, it has been argued that this measure is affected not only by the ability to discriminate culprits from innocent suspects but also by the witness's response bias, which reflects the overall conservative or liberal tendency to choose someone from the lineup[6–10]. More specifically, the diagnosticity ratio has been shown to increase as a function of an increasingly conservative response bias[9,11]. This discussion has resulted in the application of signal detection theory[12] to eyewitness identifications. Specifically, Mickes et al.[13] introduced *Receiver Operating Characteristic* (ROC) analyses to the field of eyewitness identification research.

ROC analyses have the advantage of yielding a measure of discriminability which is not confounded by response bias[13]. In the context of lineups, the term discriminability has been interpreted to refer to a witness's ability to distinguish culprits from innocent suspects[10,14]. An ROC curve is created by plotting the hit rate (i.e., the proportion of culprit identifications in culprit-present lineups) against the false alarm rate (i.e., the proportion of innocent-suspect identifications in culprit-absent lineups) at different levels of liberal or conservative responding, the latter of which is typically inferred from the witnesses' post-decision confidence judgements. It has been argued that the lineup procedure associated with the higher ROC curve—indicating higher hit rates and lower false alarm rates—is associated with superior discrimination between the culprit and an innocent suspect and should therefore be preferred[13,15].

ROC analyses have been developed to account for simple detection tasks with a 2 (signal present, signal absent) × 2 (identification, rejection) data structure (upper half of Table 1). In these tasks, only correct and false identifications are needed for measuring performance because the remaining two data categories (correct and false rejections) are redundant and provide no further information (false rejection rate = 1− correct identification rate; correct rejection rate = 1− false identification rate [12]). Lineups differ from these simple detection tasks in that they include not only a culprit or an innocent suspect but also fillers. Therefore, in each of the two types of lineups (culprit-present, culprit-absent), witnesses can make one of three responses (suspect identification, filler identification, lineup rejection), resulting in the 2 × 3 data structure displayed in the lower half of Table 1 [14,16]. In ROC analyses, filler identifications are treated like lineup rejections and the two data categories are combined to transform the 2 × 3 data structure of lineups into a 2 × 2 data structure. This is justified by noting that filler identifications and lineup rejections have the same legal consequences for the suspect. Irrespective of whether the witness identifies a filler or rejects the lineup, the suspect is not further incriminated by the eyewitness procedure. Therefore, it has been argued that, for the purpose of deciding which of two lineup procedures is superior, it is sufficient to analyze the rate of correct culprit identifications and the rate of false innocent-suspect identifications[13].

However, if the aim is to understand qualitatively different latent processes underlying eyewitness identification decisions, the two data categories that are combined for ROC analyses can yield important information when analyzed separately. This is so because the underlying processes may differ between identifying a filler and

rejecting a lineup. This is already obvious from the fact that the identification of a filler in a culprit-absent lineup is a false response, while the rejection of a culprit-absent lineup is a correct response. In culprit-absent lineups, few correct lineup rejections and many false filler identifications thus indicate poor eyewitness performance, whereas many correct lineup rejections and few false filler identifications indicate good eyewitness performance. Hence, taking into account suspect and filler identifications separately can yield important information about the latent processes underlying eyewitness identification performance[14,16,17].

The 2-HT eyewitness identification model

Here, we introduce a new measurement model for eyewitness identification performance that capitalizes on the full range of the data categories observed in typical lineup procedures within *one* model. The model belongs to the class of multinomial processing tree (MPT) models. Models from this class of formal measurement models for categorical data have been successfully applied to different areas within psychology[18,19] such as memory[20–23] or decision making[24–26]. Wagenaar and Boer[27] have already successfully introduced MPT models to the area of eyewitness memory research in a study in which they investigated the processes underlying the misinformation effect. MPT models are based on the assumption that observed response frequencies in a finite set of response categories can arise from combinations of latent processes that can be depicted in a tree-like structure[28]. Each cognitive process is represented by a model parameter that serves to measure the probability with which the process occurs[19,29]. Parameter estimation is achieved by employing the expectation-maximization algorithm proposed by Hu and Batchelder[30]. The algorithm aims to determine a set of parameters that minimize the distance between the observed response frequencies and the response frequencies predicted by the model, as measured by the log-likelihood ratio goodness-of-fit statistic G^2[30–32]. If the deviation between the frequencies predicted by the model and the observed response frequencies is not statistically significant, it can be assumed that the model fits the data[19,29]. A model is called identifiable when a unique set of parameter estimates provides an optimal fit for a given set of observed response frequencies[19,29]. Under these circumstances, MPT models can be used to test hypotheses directly at the level of the model parameters. Hypotheses are tested by imposing theoretically motivated restrictions on model parameters. If the restricted model provides a significantly worse fit to the data than the model without the restriction (measured by the ΔG^2 difference statistic), then the assumption implied by the restriction has to be rejected[18,19,29]. Thereby, this method provides insights into the latent processes underlying observable behavior, which is a major advantage of MPT models given that psychological theories often involve hypotheses about cognitive processes. Parameter estimation and statistical tests can be performed with freely available computer programs[28,32,33].

The 2-HT eyewitness identification model is illustrated in Fig. 1. The structure of the model is congenial to that of MPT models designed to measure the processes underlying performance in other recognition paradigms [e.g.,[34,35]]. The model comprises two trees, one for each of the two possible types of lineups presented to the witnesses (culprit-present and culprit-absent lineups). If the culprit is present (upper tree in Fig. 1), the presence of the culprit may be detected with probability dP (for *detection of the presence of the culprit*), which results in the correct identification of the culprit. The detection state is based on the witnesses' memory of the culprit. If witnesses fail to detect the culprit, which occurs with the complementary probability $1 - dP$, then two different types of non-detection-based processes may still lead to correct culprit identifications. First, the lineup may be unfair in that the suspect stands out from the fillers so that it can be inferred who the suspect is without relying on memory. One telling example is the case of Marvin Lamont Anderson who served fifteen years in prison for a rape that he did not commit. The selection of Anderson's face in the police lineup was most likely due to the fact that the lineup was unfair: The witness was shown a color identification card of Anderson along with six black-and-white mug shots that served as fillers[36]. The false identification of Anderson as the culprit was not based on a culprit-detection process but most likely due to the biased selection of the color photo of Anderson that stood out from the black-and-white photos of the fillers. The process of *biased selection* of the suspect in unfair lineups is reflected in parameter b. In culprit-present lineups, the biased selection of the suspect yields a correct identification of the culprit. Second, if a biased selection of the suspect does not occur (with probability $1 - b$), then witnesses may still select one of the lineup members based on guessing with probability g (for *guessing-based selection*). Guessing-based selection leads to the identification of the culprit with a probability that is equal to 1 ÷ lineup size. For instance, in a lineup with six persons, the probability that guessing-based selection will lead to the identification of the culprit is 1/6. With the complementary probability (e.g., 5/6 in a six-person lineup), the witnesses will identify a filler. Alternatively (with probability $1 - g$), the witnesses may abstain from selecting a lineup member based on guessing, which leads them to incorrectly reject the culprit-present lineup.

The lower tree illustrates the cognitive processes in response to lineups in which the culprit is absent. Witnesses may correctly detect that the culprit is absent and that no other person in the lineup can possibly be the culprit with probability dA (for *detection of the absence of the culprit*), which results in the correct rejection of the lineup. This may occur, for instance, if all persons in the lineup have a birthmark and a witness then remembers that the culprit did not have a birthmark. When the absence of the culprit is not detected (with probability $1 - dA$), the witnesses rely on the same non-detection-based biased and guessing-based processes as in culprit-present lineups. With probability b, biased selection may occur if the innocent suspect stands out from the fillers in an unfair lineup. Biased selection leads to the false identification of the innocent suspect. If no biased selection occurs (with probability $1 - b$), the witnesses may select one of the lineup members based on guessing with probability g. The probability with which guessing-based selection leads to the identification of the innocent suspect or one of the fillers depends on the lineup size (in a six-person lineup, the probability of identifying the innocent suspect is 1/6 and the probability of choosing one of the fillers is 5/6). If the witnesses abstain from selecting a lineup member based on guessing (with probability $1 - g$), the culprit-absent lineup is correctly rejected.

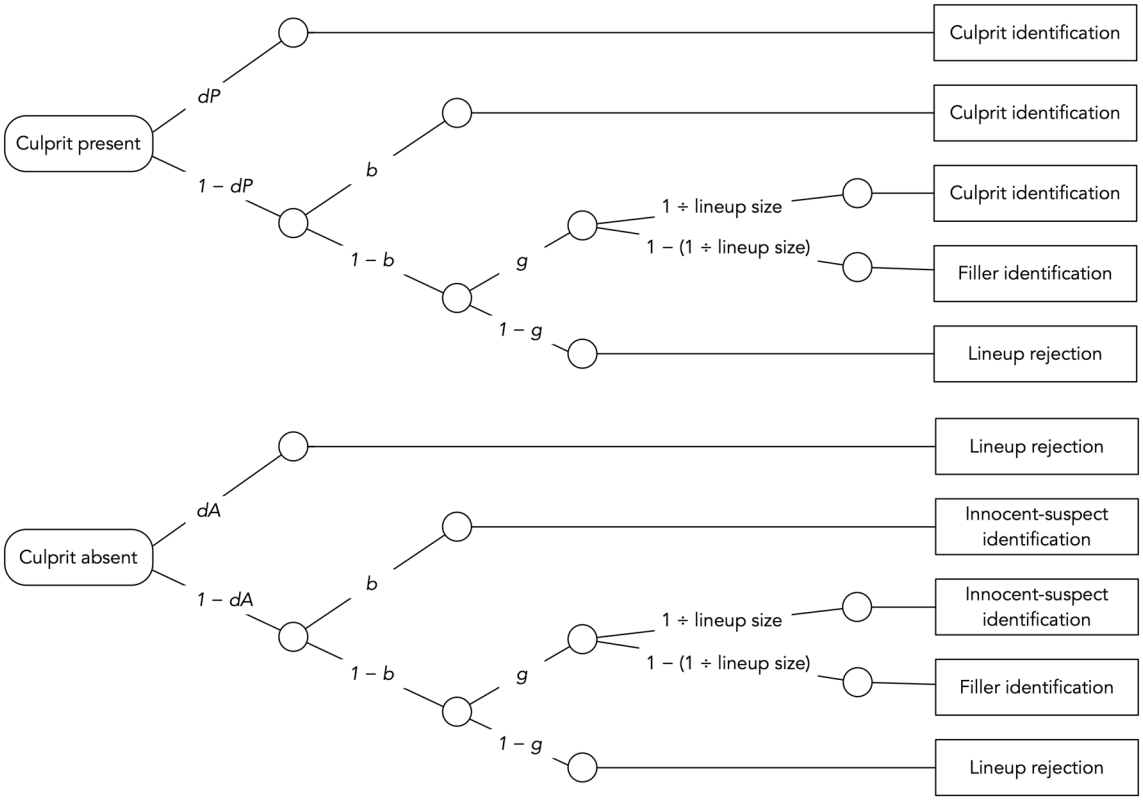

Figure 1. Graphical illustration of the 2-HT eyewitness identification model. Rounded rectangles on the left represent the two types of lineups presented to the participants (culprit-present and culprit-absent). Rectangles on the right represent the observable response categories. The letters attached to the branches represent the probabilities of the latent cognitive processes postulated by the model (*dP* = probability of detecting the presence of the culprit; *b* = probability of biased selection of the suspect; *g* = probability of guessing-based selection among the lineup members; lineup size = the number of persons in the lineup; *dA* = probability of detecting the absence of the culprit).

The 2-HT eyewitness identification model belongs to the class of 2-HT recognition models [cf.[35,37–39]]. Parameter *dP* indicates the probability of crossing the threshold for culprit-presence detection, whereas *dA* indicates the probability of crossing the threshold for detecting the absence of the culprit. To make 2-HT models in standard recognition tests identifiable, it is often assumed that the probability of detecting the presence of a signal is equal to the probability of detecting the absence of the signal [e.g.,[37]]. Here, we have the advantage that both detection parameters (*dP* and *dA*) can vary freely due to the increased number of independent observable data categories in lineups compared to simple old-new recognition tests [cf.[40]]. The 2-HT eyewitness identification model can be transformed into a one-high threshold model by setting parameter *dA* to zero[41]. The fact that *dA* does not have to be set to a certain value to obtain model identifiability implies that it is possible to empirically test whether witnesses are able to detect the absence of a culprit or not. To anticipate, the results of novel validation studies[3] and the present reanalyses of existing data consistently indicate that witnesses do not always spontaneously succeed in detecting the absence of the culprit but that there are circumstances in which witnesses can in fact detect the absence of the culprit (see the section on the validation of parameter *dA* below). When no further restrictions are employed, then the 2-HT eyewitness identification model has zero degrees of freedom, implying that there are as many free parameters as there are independent data categories to fit.

Importantly, the parameters of a novel MPT model need to be validated[19,29]. A crucial first step of model validation entails testing whether the parameters of the model measure the processes they were designed to measure. This is achieved by testing, for each model parameter, whether a parameter intended to represent a specific cognitive process is affected by experimental manipulations that target this process[18]. We decided to take two complementary approaches to model validation. In one approach[3], novel data were collected to target the model parameters one by one. Here, we test whether the parameters of the 2-HT eyewitness identification model are valid given published data from other research groups. We see these approaches as complementary. The following criteria were employed to select the data sets for the present reanalyses: First, the data had to be reported in sufficient detail, which is not the case for a surprisingly large number of studies. Second, the effect of the experimental manipulation on a model parameter had to be as obvious as possible a priori. In validation experiments, manipulations are needed for which there is a straightforward relationship between the factors manipulated and the cognitive processes that can be expected to be influenced by those manipulations. Third, the study design had to be of minimal complexity to keep the reanalysis as simple as possible. For each of the four parameters of the 2-HT eyewitness identification model, the first two studies that fulfilled these requirements were analyzed.

	Culprit-present lineups			Culprit-absent lineups		
	Culprit identifications	Filler identifications	Lineup rejections	Innocent-suspect identifications	Filler identifications	Lineup rejections
Memon et al.[45]						
Exposure duration						
Younger participants						
Long	20	1	0	2	7	12
Short	6	9	6	3	16	2
Older participants						
Long	17	2	1	2	8	10
Short	7	9	4	3	13	4
Smith[46]						
Viewing conditions						
Clear	79	9	20	9	43	153
Degraded	17	35	47	18	89	96

Table 2. Response frequencies as reconstructed from Table 1 of Memon et al. [[45], p. 345] and from Table 1 of Smith [[46], p. 500], see text for details.

First, we validated the culprit-presence detection parameter *dP* by testing whether the duration of the exposure to the culprit's face and the viewing conditions at encoding affect the participants' ability to detect the culprit's presence in the lineup. Second, we validated the biased-suspect-selection parameter *b* by reanalyzing the data of two studies in which lineup fairness was manipulated. Third, we tested whether the guessing-based selection parameter *g* differs between one-sided and two-sided pre-lineup instructions. One-sided instructions insinuate that the culprit is in the lineup, whereas two-sided instructions emphasize that the culprit might or might not be in the lineup. Finally, we completed the model validation by showing that techniques developed to help children and older adults to reject culprit-absent lineups selectively affect the culprit-absence detection parameter *dA*.

Manipulations of culprit-presence detection: Validation of parameter *dP*

Our first aim was to validate parameter *dP*, which represents the probability of detecting the presence of the culprit. The probability of detecting the culprit's face can be expected to increase with the duration of exposure to the culprit's face based on the results of recognition [for a meta-analysis, see[42]] and staged-crime studies[43,44] showing better detection performance following longer exposure. Here, we tested whether these effects of exposure duration are sensitively reflected in parameter *dP* by reanalyzing the data of Memon et al.[45] who had manipulated culprit-exposure duration at encoding. We also applied the 2-HT eyewitness identification model to the data of Smith [[46], Experiment 1] who had manipulated the viewing conditions at encoding. When the viewing conditions at encoding are poor, the culprit's face should provide a weaker match to memory, which should hinder culprit-presence detection.

Effects of exposure time on culprit-presence detection: Reanalysis of Memon et al.[45]. In the study of Memon et al.[45], participants viewed a simulated staged-crime video in which they saw the culprit's face for either a long or a short duration. In line with previous research[43,44], Memon et al. found that participants were better able to identify the culprit under the long than under the short exposure duration. The model-based reanalysis of these data should show that the manipulation of exposure duration affects parameter *dP*. More precisely, parameter *dP* should be significantly higher for the long compared to the short exposure condition.

Method. Memon et al.[45] randomly assigned younger (age: 17 to 25 years, $n = 84$) and older (age: 59 to 81 years, $n = 80$) participants to one of the four conditions resulting from a 2 (exposure duration: short vs. long) × 2 (culprit presence: present vs. absent) between-subjects design. Participants saw a staged-crime video depicting a robbery. The video was long (100 s) or short (67 s). The long video involved a clear exposure to the face of the main culprit for 45 s. The short video provided a full-face and profile-view exposure to the face of the main culprit for only 12 s. After having completed several filler questionnaires, participants were given standard two-sided pre-lineup instructions. Specifically, participants were informed that the culprit may or may not be in the lineup. Participants were then asked to identify one of the lineup members as the culprit or to indicate that the culprit was absent. Half of the participants saw a culprit-present black-and-white photo lineup consisting of six faces in a 3 × 2 array. For the remaining half of the participants, the culprit was replaced by another filler to construct a culprit-absent lineup. All fillers matched the culprit's general description. Both the culprit and the filler that replaced the culprit were positioned in the top right-hand corner of the array [for more details, see[45]].

Results. For all analyses reported in this article, parameter estimates were obtained and likelihood-ratio tests were performed using *multiTree*[32]. The α level was set to 0.05. The observed response frequencies (see the upper half of Table 2) were reconstructed from the data reported in Memon et al.'s[45] Table 1 in which the exact number of participants for each condition is not provided. We therefore divided the total number of participants in each

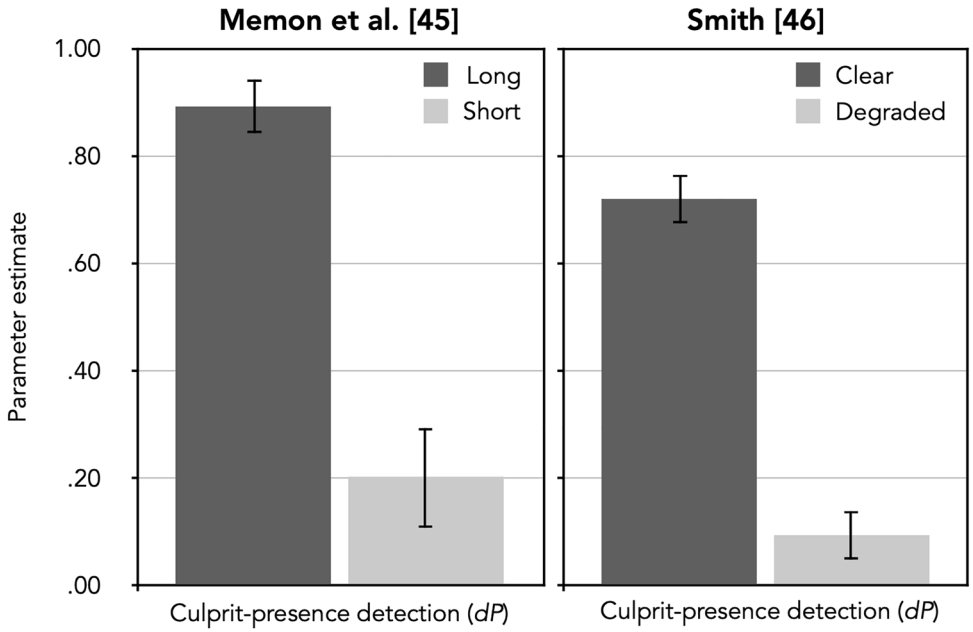

Figure 2. Estimates of parameter *dP* (representing the probability of detecting the presence of the culprit) of the 2-HT eyewitness identification model when applied to the data reported by Memon et al.[45] and by Smith [[46], Experiment 1] as a function of exposure duration (long vs. short; left panel) and viewing conditions at encoding (clear vs. degraded; right panel). The error bars represent standard errors.

Memon et al.[45]				Smith[46]			
	Parameter estimates				Parameter estimates		
Exposure duration	*b*	*g*	*dA*	Viewing conditions	*b*	*g*	*dA*
Long	.02 (.04)	.48 (.09)	.00 (.12)	Clear	.00 (.01)	.26 (.04)	.00 (.12)
Short		.78 (.06)		Degraded		.51 (.05)	

Table 3. Estimates of parameters *b*, *g* and *dA* of the 2-HT eyewitness identification model for the data reported by Memon et al.[45] and by Smith [[46], Experiment 1]. Values in parentheses represent standard errors. Within the model used as a comparison standard, parameter *g* was estimated separately for long and short exposure conditions[45] and for clear and degraded viewing conditions [[46]

age group by four (the number of conditions in each age group) to estimate the number of participants in each condition. Following standard practice [cf.[47,48]], the number of innocent-suspect identifications was estimated by dividing the total number of false identifications in culprit-absent lineups by the number of lineup members (in this case, six). Likewise, the number of filler identifications in culprit-absent lineups was estimated by subtracting the number of innocent-suspect identifications from the total number of false identifications. To test how the manipulation of exposure duration affects parameter *dP*, four sets of the processing trees displayed in Fig. 1 were needed, one for the long and one for the short exposure duration, for both the younger and the older participants, respectively. Parameters *dA* and *b* were each set to be equal among the four conditions because there was no reason to assume that these parameters should differ as a function of the conditions. To simplify the output, we also aggregated the data for younger and older participants by imposing the equality restrictions that culprit-presence detection (*dP*) and guessing-based selection (*g*) did not differ between the two age groups in each exposure condition. These assumptions were based on the results of Memon et al. who had found no significant effect of age on lineup performance. The model incorporating these restrictions was used as a comparison standard for the subsequent nested likelihood-ratio tests, $G^2(10) = 8.19$, $p = 0.610$. The estimates of parameter *dP* as a function of exposure duration are shown in the left panel of Fig. 2. Table 3 shows the estimates of parameters *b*, *g* and *dA*.

As explicated above, the 2-HT eyewitness identification model is based on the full 2×3 structure of lineup data. To use the model for the present reanalysis and that of all other data sets without a designated innocent suspect, it is thus necessary to make the assumption that the culprit-absent lineups contained an innocent suspect and that the lineups were fair, which implies that the innocent suspects were selected with the same probability as each of the fillers. Given that the proportion of innocent-suspect identifications is determined from the proportion of filler identifications, one could argue that, technically speaking, experiments without a designated innocent suspect have one fewer independent data category than experiments with a designated innocent suspect.

As noted by a reviewer during the review process, this would result in a loss of degrees of freedom in the model used as a comparison standard. However, the same applies to all nested models generated by parameter restrictions (see below). Therefore, none of the subsequent likelihood-ratio tests, all of which pertain to the *differences* between the models used as comparison standards and the more restricted, nested models that are central to the model validation are affected by this issue.

The culprit-presence detection parameter *dP* was clearly affected by exposure duration. Participants in the long-exposure group were significantly more likely to detect the culprit than those in the short-exposure group, $\Delta G^2(1) = 34.33$, $p < 0.001$. Furthermore, the probability of selecting a lineup member based on guessing (parameter *g*) was significantly higher when exposure duration was short than when it was long, $\Delta G^2(1) = 10.24$, $p = 0.001$, likely reflecting the well-established phenomenon of compensatory guessing, that is, the phenomenon that participants tend to rely more on guessing when memory is poor[23,49–54].

Effects of viewing conditions on culprit-presence detection: Reanalysis of Smith[46]. Smith [[46], Experiment 1] manipulated the viewing conditions at encoding by presenting participants with either a clear or a degraded version of a simulated-crime video. In line with the results of Smith et al. [[55], Experiment 1], degraded viewing conditions decreased culprit detection performance. If the 2-HT eyewitness identification model is valid, the manipulation of viewing conditions should affect the culprit-presence detection parameter *dP*. Specifically, parameter *dP* should be significantly higher for the clear compared to the degraded viewing conditions.

Method. Smith [[46], Experiment 1] randomly assigned 615 participants to one of the four experimental conditions resulting from a 2 (viewing conditions: clear vs. degraded) × 2 (culprit presence: present vs. absent) between-subjects design. Half of the participants viewed the video of the culprit in high-resolution so that the culprit's facial features were clearly visible. The remaining participants saw a low-resolution, overexposed version of the same video in which it was extremely difficult to perceive facial details. In both versions of the 90-s video, a scene at an airport was shown in which the culprit switched his suitcase with that of another person. After having completed a 4-min anagram task, participants were given standard two-sided pre-lineup instructions and were then presented with a simultaneous lineup. The data from the culprit-absent lineups were crucial for testing Smith's hypothesis. Therefore, approximately two-thirds of the participants were presented with the culprit-absent lineup containing six fillers. About one-third of the participants saw a six-person culprit-present lineup that included the culprit together with five of the six fillers who were randomly selected. All fillers matched the culprit's general description. The lineup members were presented in random order [for more details, see[46], Experiment 1].

Results. The observed response frequencies (see the lower half of Table 2) of lineup identifications and rejections were taken from Table 1 in Smith[46]. As in the previous analysis, we estimated the number of innocent-suspect identifications using the standard procedure of dividing the total number of false identifications in culprit-absent lineups by the numb er of lineup members [cf.[47,48]]. For the model-based reanalysis, we needed two sets of the processing trees depicted in Fig. 1, one for the clear and one for the degraded viewing conditions. As in the previous analysis, parameters *dA* and *b* were each set to be equal between both conditions because there was no reason to assume that these parameters should differ as a function of the viewing conditions. The model incorporating these restrictions was used as a comparison standard for the subsequent nested likelihood-ratio tests, $G^2(2) = 1.81$, $p = 0.405$. The estimates of culprit-presence detection parameter *dP* as a function of viewing conditions are shown in the right panel of Fig. 2. Table 3 shows the estimates of parameters *b*, *g* and *dA*.

The model-based reanalysis confirmed that parameter *dP* sensitively reflected the effect of the viewing conditions at encoding. The probability of correctly detecting the presence of the culprit was significantly higher under clear than under degraded viewing conditions, $\Delta G^2(1) = 74.73$, $p < 0.001$. In addition, parameter *g* was significantly higher when viewing conditions were poor, $\Delta G^2(1) = 32.02$, $p < 0.001$, which can be attributed to compensatory guessing[23,49–54].

Discussion. In reanalyzing data obtained from the literature[45,46], we first focused on parameter *dP* of the 2-HT eyewitness identification model. Parameter *dP* reflects the detection of the presence of the culprit in a lineup. Both exposure duration and viewing conditions can be predicted clearly and unambiguously to affect the ability to detect the culprit in culprit-present lineups. Parameter *dP* sensitively reflected the two different manipulations of culprit-presence detection. The results of the first reanalysis[45] confirmed that parameter *dP* was significantly higher under long than short exposure to the culprit at encoding. The results of the second reanalysis [[46], Experiment 1] demonstrated that the culprit-presence detection parameter *dP* was significantly higher when viewing conditions at encoding were good than when they were poor. It can thus be concluded that parameter *dP* sensitively reflects manipulations of culprit-presence detection in the predicted directions, which is consistent with the theoretical interpretation of the results in the original studies[45,46]. This suggests that the validation of parameter *dP* was successful.

In both reanalyses, the guessing-based selection parameter *g* was also affected by the manipulations of encoding difficulty. Participants were more likely to select one of the lineup members based on guessing when culprit-presence detection was poor. Ideally, the procedure of model validation entails experimental manipulations that only influence the target parameter (*dP* in this instance) in the expected direction without affecting other parameters. However, it is often not possible to find strong manipulations that influence only a single cognitive process without side effects on other processes[29]. In such a case it is ideal if there is a plausible explanation for these side effects. In the present case, the side effect on the guessing-based selection parameter *g* can be explained by compensatory guessing, which refers to the well-established phenomenon that participants rely more on

guessing when memory is poor[23,49–54]. The effect of compensatory guessing may thus be linked to the fact that the differences in memory between the conditions was rather strong in the reanalyzed studies of Memon et al.[45] and of Smith[46].

Manipulations of biased suspect selection: Validation of parameter *b*

Our second objective was to investigate the validity of parameter *b*, which serves to measure the probability of biased selection of the suspect from a lineup if the suspect stands out from the other lineup members. Thus, the estimate of *b* should primarily be determined by the degree of lineup fairness. Specifically, parameter *b* should be higher for unfair lineups than for fair lineups. If a lineup is perfectly fair, there is no way to tell the suspect apart from the other lineup members without memory of the culprit. Under these conditions, biased selection of the suspect should not be possible and parameter *b* should be indistinguishable from zero. Note that this is what happened in the reanalyses of Memon et al.[45] and Smith[46] in which we followed the standard procedure of dividing the total number of false identifications in culprit-absent lineups by the number of lineup members [cf.[47,48]] to obtain an estimate of the innocent-suspect identifications in culprit-absent lineups. However, it is well known that real-world lineups are often not perfectly fair and there is strong evidence of a biased selection of suspects in unfair lineups [e.g.,[56–58]]. If the 2-HT eyewitness identification model is valid, then the biased selection of suspects in unfair lineups should be sensitively reflected in parameter *b*. This was tested by reanalyzing data from two large studies [48,59] in which lineup fairness was manipulated. Specifically, Wetmore et al.[59] observed an effect on lineup fairness when manipulating the similarity between the suspect and the fillers. The study of Colloff et al.[48] provides a complementary approach to manipulating the biased selection of suspects by eliminating, via digital photo manipulation, distinctive features that caused the suspect to stand out from the fillers.

Effects of filler-suspect similarity on biased suspect selection: Reanalysis of Wetmore et al.[59]. Wetmore et al.[59] manipulated the degree of similarity between the lineup fillers and the suspect. Good fillers (high similarity to the suspect) were used to create a fair lineup, whereas bad fillers (low similarity to the suspect) were used to create an unfair lineup. Wetmore et al. found significantly more suspect identifications when the lineup was unfair than when it was fair. The theoretical interpretation of this finding was that the higher rate of suspect identifications in unfair lineups was due to biased suspect selection. Therefore, Wetmore et al.'s lineup fairness manipulation should be reflected in the 2-HT eyewitness identification model's parameter *b*. Specifically, parameter *b* should be significantly larger for the unfair lineup than for the fair lineup if the interpretation of parameter *b* in terms of biased suspect selection is valid.

Method. Wetmore et al.[59] randomly assigned 1584 participants to one of the 18 conditions of a 3 (identification task: showup vs. fair lineup vs. unfair lineup) × 3 (suspect: guilty vs. Innocent 1 vs. Innocent 2) × 2 (delay: immediate vs. delayed) between-subjects design. Participants watched a 105-s video of a man stealing a woman's purse. Participants completed a distractor task that consisted of solving 20 anagrams either immediately or 48 h after the simulated crime had been shown. Standard two-sided pre-lineup instructions were given before the participants proceeded to the identification task. Participants in the lineup condition viewed a six-person simultaneous lineup composed of two rows of three photos. The lineups included the culprit (culprit-present lineup) or an innocent suspect (culprit-absent lineup) who matched the culprit's description. Wetmore et al. distinguished between two innocent suspects that had been taken from an earlier study by Gronlund et al.[60]. The innocent suspects were intended to be equally good matches to the culprit. For simplicity, we did not distinguish between the two innocent suspects. The participants saw either a fair or an unfair lineup. To create unfair lineups in which the suspect stood out, five poor fillers were selected who, apart from being white men, shared only one characteristic with the culprit. In contrast, fair lineups contained good fillers who, apart from being white men, shared five characteristics with the culprit [for more details, see[59]].

Results. The observed response frequencies (see the upper half of Table 4) were reconstructed from the proportions presented by Wetmore et al.[59] in their Table 2. For simplicity, we limited our reanalysis to the lineup data to focus on the most relevant comparison (fair vs. unfair) for testing the validity of parameter *b*. Thus, four sets of the trees shown in Fig. 1 were needed, one for immediate fair, one for immediate unfair, one for delayed fair and another for delayed unfair lineups. The *dA* parameters were set to be equal among the four conditions because there was no reason to assume that the detection of culprit absence should differ as a function of the conditions. Given that Wetmore et al. had found no effect of delay on identification performance, we reduced the model complexity by assuming that biased suspect selection (*b*), culprit-presence detection (*dP*) and guessing-based selection (*g*) did not differ between immediate and delayed lineups. The model incorporating these restrictions was used as a comparison standard for the subsequent nested likelihood-ratio tests, $G^2(9) = 14.10$, $p = 0.119$. The estimates of the biased-suspect-selection parameter *b* as a function of lineup fairness are shown in the left panel of Fig. 3. Table 5 shows the estimates of parameters *dP*, *g* and *dA*.

The fairness manipulation affected parameter *b* as predicted under the assumption that this parameter represents biased suspect selection. The probability of biased suspect selection was significantly higher in the unfair than in the fair-lineup condition, $\Delta G^2(1) = 31.84$, $p < 0.001$. In addition, guessing-based selection was more prevalent when the fillers matched the description of the culprit than when they did not match the description: Parameter *g* was decreased in the unfair in comparison to the fair-lineup condition, $\Delta G^2(1) = 15.42$, $p < 0.001$. This is to be expected because guessing-based selection among the lineup members should have been discouraged in the unfair lineup due to the poor match of the fillers to the description of the culprit. Parameter *dP*, which measures the ability to detect the presence of the culprit, was not affected by lineup fairness, $\Delta G^2(1) = 0.19$, $p = 0.665$.

	Culprit-present lineups			Culprit-absent lineups		
	Culprit identifications	Filler identifications	Lineup rejections	Innocent-suspect identifications	Filler identifications	Lineup rejections
Wetmore et al.[59]						
Lineup fairness						
Immediate						
Fair	41	6	13	31	74	54
Unfair	57	2	11	56	37	58
Delay						
Fair	59	13	14	22	59	38
Unfair	54	4	14	48	31	39
Colloff et al.[48]						
Lineup type						
Block	323	390	414	101	503	534
Pixelation	320	411	414	102	512	510
Replication	347	382	396	105	523	513
Unfair	629	206	275	364	219	434

Table 4. Response frequencies as reconstructed from Table 2 of Wetmore et al. [59, p. 11] and from Table 2 of Colloff et al. [48, p. 1235], see text for details.

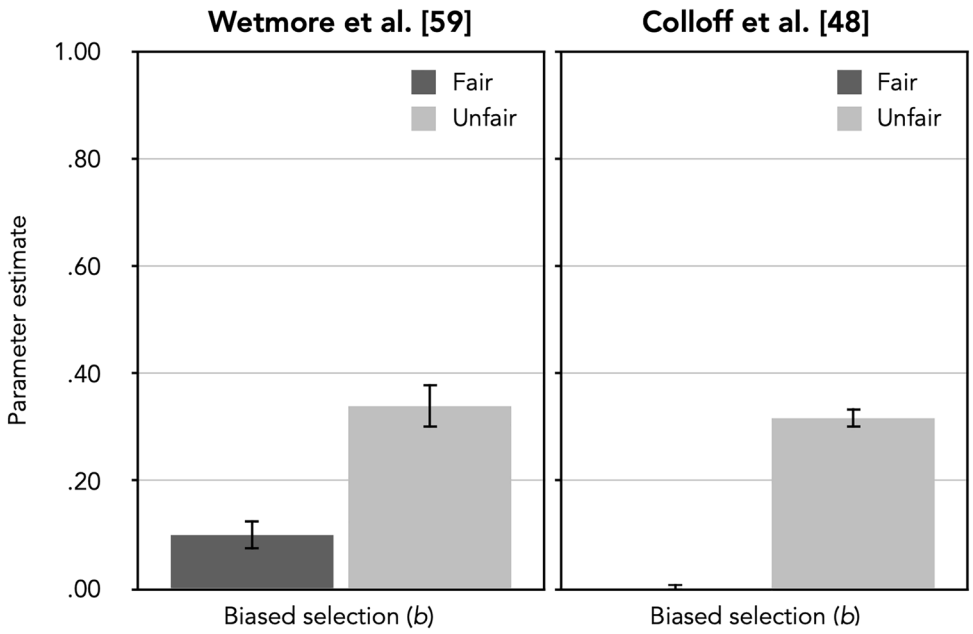

Figure 3. Estimates of parameter *b* (representing biased selection of the suspect) of the 2-HT eyewitness identification model when applied to the data reported by Wetmore et al.[59] and by Colloff et al.[48] as a function of lineup fairness (fair vs. unfair). The error bars represent standard errors.

Effects of the elimination of distinctive features of the suspect on biased suspect selection: Reanalysis of Colloff et al.[48]. Colloff et al.[48] presented participants with staged-crime videos in which the culprits had distinctive facial features. To create fair lineups, Colloff et al. tested three possible techniques to prevent distinctive suspects from standing out: blocking, pixelating or replicating the feature on all faces in the lineup using digital photo-manipulation technology. Colloff et al. also included an unfair lineup in which only the suspect had the distinctive feature. The results showed that if a suspect stood out, identifications shifted from the fillers to the suspect. Thus, unfair lineups led to significantly more suspect identifications than fair lineups. If the 2-HT eyewitness identification model is valid, the biased-suspect-selection parameter *b* should be affected by the manipulation of fairness. Specifically, parameter *b* should be higher for the unfair than for the fair lineups.

Method. Colloff et al.[48] randomly assigned 8925 participants to one of the eight conditions of a 4 (lineup type: block vs. pixelation vs. replication vs. unfair) × 2 (culprit presence: present vs. absent) between-subjects design.

Wetmore et al.[59]				Colloff et al.[48]			
	Parameter estimates				Parameter estimates		
Lineup fairness	*dP*	*g*	*dA*	Lineup fairness	*dP*	*g*	*dA*
Fair	.61 (.05)	.61 (.05)	.00 (.07)	Fair	.22 (.01)	.55 (.01)	.02 (.02)
Unfair	.64 (.06)	.42 (.05)		Unfair	.32 (.03)	.43 (.02)	

Table 5. Estimates of parameters *dP*, *g* and *dA* of the 2-HT eyewitness identification model for the data reported by Wetmore et al.[59] and by Colloff et al.[48]. Values in parentheses represent standard errors. Within the model used as a comparison standard, parameters *dP* and *g* were estimated separately for fair and unfair lineups. Parameter *dA* was set to be equal between the fair and the unfair-lineup conditions. *dP* = probability of detecting the presence of the culprit; *g* = probability of guessing-based selection among the lineup members; *dA* = probability of detecting the absence of the culprit.

The participants viewed one of four 30-s crime videos (carjacking, graffiti attack, mugging, theft) depicting four different culprits, each with a unique and distinctive facial feature (scar on the left cheek, bruising around the right eye, nose piercings in the left nostril or facial tattoo on the right cheek). After an 8-min retention interval, participants were presented with a simultaneous six-person lineup composed of two rows of three photos. The lineup consisted of either one culprit and five fillers (culprit-present lineup) or six fillers (culprit-absent lineup). The fillers were randomly drawn from a pool of 40 description-matched fillers created for each culprit. Depending on the condition, the distinctive feature of the culprit was treated differently. Three treatments were intended to produce fair lineups by (a) concealing the distinctive feature with a solid black rectangle on the culprit and by covering the equivalent area on each filler (block), (b) pixelating the distinctive feature on the culprit and the equivalent area on each filler (pixelation) or (c) digitally adding the distinctive feature to each filler (replication). In the unfair condition, participants saw a lineup in which the suspect stood out due to them being the only person with a distinctive feature. For unfair culprit-present lineups, Colloff et al. left the distinctive feature on the culprit unaltered. For each unfair culprit-absent lineup, a replication filler photo with the culprit's distinctive feature was randomly selected to create an innocent suspect with the culprit's distinctive feature while the other five filler photos remained unedited. For all lineups, the position of the culprit (in culprit-present lineups) and innocent suspect (in unfair culprit-absent lineups) was randomly determined. Before viewing the lineup, standard two-sided pre-lineup instructions were presented [for more details, see[48]].

Results. The observed response frequencies (see the lower half of Table 4) were provided by Colloff et al.[48] in their Table 2 (rounded to the next integer, where applicable). Four sets of the model trees depicted in Fig. 1 were necessary to reanalyze the data, three for the fair-lineup conditions (block, pixelation, replication) and one for the unfair condition. The *dA* parameters were set to be equal among the four conditions because there was no reason to assume that the detection of culprit absence should differ as a function of the conditions. In Colloff et al., the three fair-lineup conditions were all associated with similar levels of performance. To keep the model as simple as possible, we therefore assumed that biased suspect selection (*b*), culprit-presence detection (*dP*) and guessing-based selection (*g*) did not differ among the three fair-lineup conditions. The model incorporating these restrictions was used as a comparison standard for the subsequent nested likelihood-ratio tests, $G^2(9) = 12.92$, $p = 0.166$. The estimates of the biased-suspect-selection parameter *b* as a function of lineup fairness are shown in the right panel of Fig. 3. Table 5 shows the estimates of parameters *dP*, *g* and *dA*.

Parameter *b* adequately reflected the fairness manipulation. The probability of biased suspect selection was significantly higher in the unfair-lineup condition compared to the fair-lineup condition, $\Delta G^2(1) = 418.89$, $p < 0.001$. In addition, parameter *g*, which represents the probability of selecting one of the lineup members based on guessing, was significantly decreased in the unfair-lineup condition in comparison to the fair-lineup condition, $\Delta G^2(1) = 48.99$, $p < 0.001$. This is to be expected given that guessing-based selection among the lineup members was discouraged by the fact that none of the fillers shared the culprit's distinctive facial feature. This is parallel to what was observed in the data of Wetmore et al.[59] considered in the previous section. The ability to detect the presence of the culprit, captured by parameter *dP*, was significantly higher in unfair than in fair lineups, $\Delta G^2(1) = 8.37$, $p = 0.004$, which can be explained by attention being drawn to faces with the culprit's distinctive facial features.

Discussion. The results support the validity of parameter *b* representing the process of biased suspect selection. In the two experiments that were reanalyzed, unfair lineups were created either by using fillers with low similarity to the suspect[59] or by using a suspect with an uncovered distinctive facial feature that makes the suspect stand out from the fillers[48]. Both model-based reanalyses demonstrated parameter *b* to be significantly higher for unfair lineups than for fair lineups. Parameter *b* thus sensitively reflects manipulations of biased suspect selection.

When applying the 2-HT eyewitness identification model to Colloff et al.'s[48] data, the culprit-presence detection parameter *dP* was also affected by lineup fairness. Better detection of culprit presence in unfair lineups can be easily explained by distinctive facial features drawing attention to the culprit's face. This finding is in line with the well-established principle that the presence of highly similar distractors at test decreases memory performance [cf.[14,61,62]]. In both reanalyses, parameter *dP* was higher in unfair than in fair lineups, but this difference reached significance only in the reanalysis of the Colloff et al. data. Given that sample size was much larger in the study

of Colloff et al. than in the study of Wetmore et al.[59], this discrepancy is expected if the secondary effect of lineup fairness on culprit-presence detection is more subtle than the primary effect of lineup fairness on biased selection and can thus only be detected when the sample size provides the statistical power to detect a subtle effect. At first glance, better culprit detection in unfair lineups may seem unexpected given that unfair lineups were associated with an impaired ability to discriminate between culprits and innocent suspects in the analyses of Colloff et al. and Wetmore et al. However, their conclusions relied on ROC analyses and thus only on the correct culprit identifications and false innocent-suspect identifications. As Smith et al.[62] have already pointed out, the finding of increased culprit-presence detection in Colloff et al.'s unfair lineups is already apparent at the surface level of the raw response frequencies when considering the full 3×2 matrix of lineup data. The effect of the fairness manipulation on biased selection is obvious when looking at the false responses: Participants chose the innocent suspect in culprit-absent lineups more often than a filler when the lineup was unfair but chose one of the fillers more often than the innocent suspect when the lineup was fair. In the 2-HT eyewitness identification model, this effect is reflected in the biased-suspect-selection parameter *b*. However, when looking at the correct responses, it is clear that culprit-presence detection was better in unfair lineups than in fair lineups: In culprit-present lineups, participants made more correct identifications when the lineups were unfair than when they were fair, while the correct rejections in unfair lineups stayed at about the same rate. This aspect of the data is captured in the culprit-presence detection parameter *dP* in the 2-HT eyewitness identification model. Furthermore, the results of both reanalyses[48,59] showed a decreased guessing-based selection among the lineup members (parameter *g*) in unfair compared to fair lineups. Even without using a formal model of the processes underlying eyewitness performance, the surface-level data already suggest that guessing-based selection among the lineup members occurred less frequently when the lineup was unfair than when it was fair: Participants in both studies produced fewer filler identifications when lineups were unfair than when they were fair. Guessing-based selection among lineup members may have been discouraged in unfair lineups due to the poor match of the fillers to the culprit.

The 2-HT eyewitness identification model is able to capture all of these changes in performance because its parameters are based on the full 3×2 data structure to distinguish between culprit-presence detection, culprit-absence detection, biased selection and guessing-based selection so that the model is able to capture changes in those data categories that are often ignored when lineup performance is analyzed. Our findings thus suggest that fair lineups produce better outcomes than unfair lineups because they decrease biased selection of the suspect and not because they improve culprit-presence detection [in line with the conclusions of[14,62]].

Manipulations of guessing-based selection: Validation of parameter *g*

The next step was to test the validity of parameter *g*, which reflects the probability of selecting one of the lineup members based on guessing, a process that occurs alarmingly frequently not only in the laboratory but also in real-world lineups in which selecting a lineup member as the culprit may have serious consequences[63]. A straightforward and reliable way to manipulate guessing-based selection is to use instructions designed to manipulate the participants' expectations about what they will encounter [e.g.,[34]]. In the context of lineups, so-called 'biased' pre-lineup instructions insinuate that the culprit is in the lineup and thus increase participants' willingness to select one of the lineup members based on guessing when they are uncertain about whether or not the culprit is in the lineup[64–66]. The term 'biased instructions' is often used to refer to *one-sided* instructions that emphasize selectively the importance of selecting the culprit. The term 'unbiased instructions' is often used to refer to *two-sided* instructions that make participants aware of the fact that the culprit may or may not be in the lineup so that it is equally important to identify the culprit in culprit-present lineups and to reject culprit-absent lineups. Manipulating pre-lineup instructions therefore can be expected to affect the guessing-based selection parameter *g* and not the biased-suspect-selection parameter *b*. To avoid confusion, we therefore reserve the term 'bias' for the biased selection of the suspect in unfair lineups. We use the term one-sided instructions for instructions that emphasize selectively the need to identify the culprit and the term two-sided instructions for instructions that emphasize both the need to identify the culprit in culprit-present lineups and the need to reject culprit-absent lineups. One-sided instructions should encourage guessing-based selection, while two-sided instructions should discourage guessing-based selection. Here we reanalyzed datasets of Malpass and Devine [67] and of Lampinen et al. [[68], Experiment 1] who had used one-sided and two-sided pre-lineup instructions.

Effects of pre-lineup instructions on guessing-based selection: Reanalysis of Malpass and Devine[67]. Malpass and Devine[67] influenced their participants' guessing behavior by manipulating pre-lineup instructions that either insinuated or did not insinuate that the culprit was in the lineup. Malpass and Devine found that participants were more likely to choose one of the lineup members when one-sided instructions were given than when two-sided instructions were given, reflecting a higher prevalence of guessing-based selection after one-sided instructions. If parameter *g* of the 2-HT eyewitness identification model validly reflects guessing-based selection, then parameter *g* should be higher under one-sided than under two-sided instructions.

Method. Malpass and Devine[67] randomly assigned 100 students to one of the four conditions of a 2 (lineup instruction: one-sided vs. two-sided) $\times$ 2 (culprit presence: present vs. absent) between-subjects design. The students witnessed a staged act of vandalism during a biofeedback demonstration at the university. They were exposed to a male confederate (visible for 85 s) who damaged the electrical equipment before he fled the room. On one of three evenings following the act of vandalism, participants viewed a simultaneous live lineup consisting of five persons who were lined up against the wall of a room. Half of the lineups included the culprit and four innocent fillers (culprit-present lineups), while the remaining lineups consisted of five innocent fillers (culprit-absent lineups). All fillers matched the appearance of the culprit. The position of each lineup member was counterbalanced. Before viewing the lineup, participants read either one-sided or two-sided printed lineup

	Culprit-present lineups			Culprit-absent lineups		
	Culprit identifications	Filler identifications	Lineup rejections	Innocent-suspect identifications	Filler identifications	Lineup rejections
Malpass and Devine[67]						
Pre-lineup instructions						
One-sided	21	7	0	3	14	5
Two-sided	19	0	4	2	7	18
Lampinen et al.[68]						
Pre-lineup instructions						
One-sided	68	88	10	25	126	15
Standard two-sided	60	85	22	19	97	50
Detailed two-sided	56	76	32	20	100	46

Table 6. Response frequencies as reconstructed from Table 1 of Malpass and Devine [67, p. 485] and from Table 1 of Lampinen et al. [68, p. 412], see text for details.

instructions. The one-sided instructions led participants to believe that the culprit was present. The students were instructed to choose one of five numbers (one number for each lineup member). There was no obvious option for rejecting the lineup. Instead, participants had to ask how to indicate such a response. In contrast, the two-sided instructions explicitly stated that the culprit may or may not be present and provided the participants with an option (circling number 0) to reject the lineup [for more details, see[67]].

Results. The observed response frequencies (see the upper half of Table 6) were reconstructed from the proportions reported by Malpass and Devine[67] in their Table 1. Again, the number of innocent-suspect identifications was estimated by dividing the total number of false identifications in culprit-absent lineups by the number of lineup members [cf.[47,48]]. For the model-base d r e analysis, two sets of the trees shown in Fig. 1 were needed, one for the one-sided and one for the two-sided pre-lineup instructions. Parameters *dA* and *b* were each set to be equal between the conditions because there was no reason to assume that these parameters should differ as a function of the pre-lineup instructions. The model incorporating these restrictions was used as a compariso n standard for the subsequent nested likelihood-ratio tests, $G^2(2) = 3.35$, $p = 0.187$. The estimates of the guessing-based selection parameter *g* as a function of pre-lineup instructions are shown in the left panel of Fig. 4. Table 7 shows the estimates of parameters *dP*, *b* and *dA*.

The guessing-based selection parameter was clearly affected by the manipulation of the pre-lineup instructions. Parameter *g* representing the tendency to select, in a state of uncertainty, one of the lineup members based on guessing was significantly higher under one-sided than under two-sided instructions, $\Delta G^2(1) = 20.95$, $p < 0.001$. By contrast, the culprit-presence detection parameter *dP* remained unaffected by the instructions, $\Delta G^2(1) = 0.88$, $p = 0.347$.

Effects of pre-lineup instructions on guessing-based selection: Reanalysis of Lampinen et al.[68]. Lampinen et al. [68, Experiment 1] manipulated participants' guessing behavior in a similar fashion as Malpass and Devine[67]. While one-sided pre-lineup instructions were presented to encourage guessing-based selection, two versions of two-sided pre-lineup instructions were given to discourage guessing-based selection. The findings showed that the two-sided instructions significantly reduced inaccurate identifications[68]. Thus, in terms of the 2-HT eyewitness identification model, the guessing-based selection parameter *g* should be higher under one-sided instructions than under two-sided instructions.

Method. Lampinen et al. [68, Experiment 1] randomly assigned 995 students to one of the six experimental conditions resulting from a 3 (lineup instruction: one-sided vs. standard two-sided vs. detailed two-sided) × 2 (culprit presence: present vs. absent) between-subjects design. Participants viewed a 15-s video showing a woman stealing a backpack. After completing a 5-min distractor task, participants were given a paper copy of one of three types of instructions to read while the experimenter simultaneously read the instructions out loud. The one-sided instructions simply required the participants to identify the culprit in the lineup. The standard two-sided instructions contained the additional statement that the culprit may or may not be in the lineup. The detailed two-sided instructions were formulated according to the recommendations of major United States and international law enforcement agencies and were supplemented by statements that (a) it is just as important to clear an innocent person as it is to identify the culprit and (b) the police will continue to investigate the crime regardless of whether an identification is made or the lineup is rejected[69]. Participants were subsequently shown a simultaneous lineup consisting of six faces in a 3 × 2 array in which the culprit was either present or absent. Lampinen et al. created six versions of the culprit-present lineups by replacing one of the fillers with the culprit. All fillers matched the culprit's description [for more details, see[68], Experiment 1].

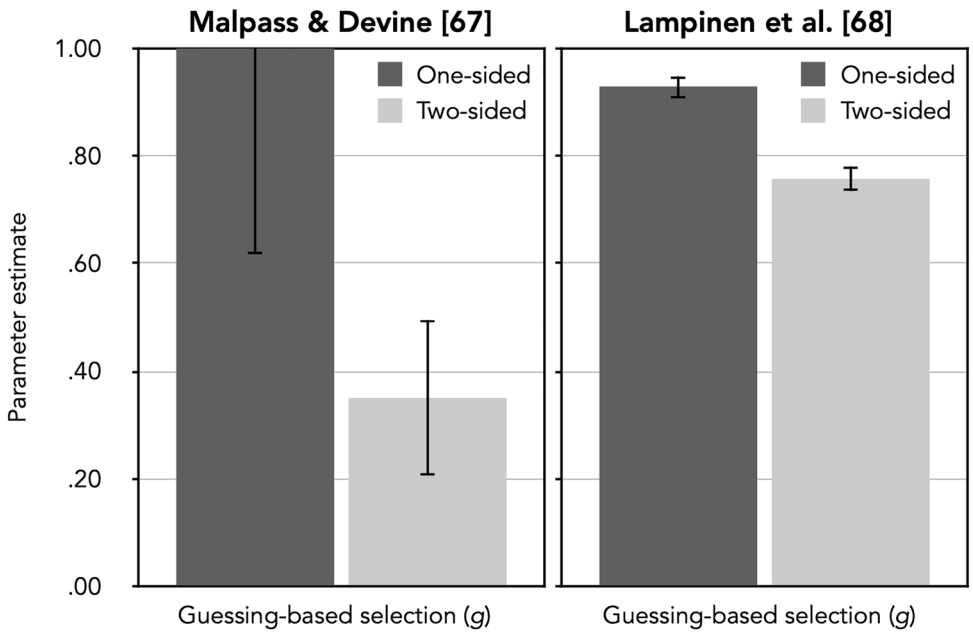

Figure 4. Estimates of parameter *g* (representing the probability of guessing-based selection among the lineup members) of the 2-HT eyewitness identification model when applied to the data reported by Malpass and Devine[67] and by Lampinen et al. [[68], Experiment 1] as a function of pre-lineup instructions (one-sided vs. two-sided). The error bars represent standard errors.

Malpass and Devine[67]				Lampinen et al.[68]			
	Parameter estimates				Parameter estimates		
Pre-lineup instructions	*dP*	*b*	*dA*	Pre-lineup instructions	*dP*	*b*	*dA*
One-sided	.68 (.11)	.01 (.06)	.21 (.30)	One-sided	.30 (.05)	.00 (.02)	.04 (.03)
Two-sided	.81 (.09)			Two-sided	.26 (.03)		

Table 7. Estimates of parameters *dP, b* and *dA* of the 2-HT eyewitness identification model for the data reported by Malpass and Devine[67] and by Lampinen et al. [[68], Experiment 1]. Values in parentheses represent standard errors. Within the model used as a comparison standard, parameter *dP* was estimated separately for one-sided and two-sided pre-lineup instructions. Parameters *b* and *dA* were each set to be equal between the one-sided and the two-sided pre-lineup instruction conditions. *dP* = probability of detecting the presence of the culprit; *b* = probability of biased suspect selection; *dA* = probability of detecting the absence of the culprit.

Results. The observed response frequencies (see the lower half of Table 6) were calculated from the proportions reported by Lampinen et al.[68] in their Table 1. Because the exact number of participants for each condition was not provided, we assumed that participants were assigned in equal numbers to the conditions and divided the total number of participants by six (the number of conditions). The number of innocent-suspect identifications was estimated by dividing the total number of false identifications in culprit-absent lineups by the number of lineup members [cf.[47,48]]. Three sets of the trees shown in Fig. 1 were used, one for the one-sided instructions condition, one for the standard two-sided instructions condition and one for the detailed two-sided instructions condition. Parameters *dA* and *b* were each set to be equal among the conditions because there was no reason to assume that these parameters should differ as a function of the lineup instructions. For the sake of simplicity, the processing trees for the two types of two-sided instructions were combined by additionally assuming that guessing-based selection (*g*) and culprit-presence detection (*dP*) did not differ between more and less detailed two-sided instructions. This assumption was based on the results of Lampinen et al., who had found no differences in identification performance between the two types of two-sided lineup instructions. The model incorporating these restrictions was used as a comparison standard for the subsequent nested likelihood-ratio tests, $G^2(6) = 4.86$, $p = 0.562$. The estimates of the guessing-based selection parameter *g* as a function of pre-lineup instructions are shown in the right panel of Fig. 4. Table 7 shows the estimates of parameters *dP*, *b* and *dA*.

The estimate of parameter *g,* which represents the probability of selecting, in a state of uncertainty, one of the lineup members based on guessing, was higher for one-sided instructions than for two-sided instructions. This difference was statistically significant, $\Delta G^2(1) = 36.39$, $p < 0.001$. By contrast, the culprit-presence detection parameter *dP* did not significantly differ between the types of instructions, $\Delta G^2(1) = 0.69$, $p = 0.406$.

Discussion. The model-based results of both reanalyses[67,68] showed that parameter *g*, representing the probability of selecting one of the lineup members based on guessing, were consistently higher for one-sided than for two-sided pre-lineup instructions. The fact that the guessing-based selection parameter sensitively reflected the experimental manipulations of the pre-lineup instructions in both reanalyses further supports the validity of the 2-HT eyewitness identification model.

Manipulations of culprit-absence detection: Validation of parameter *dA*

The final step in the model validation presented here concerns parameter *dA,* which represents the probability of detecting the absence of the culprit. In order to validate this parameter, an experimental manipulation is needed that affects the proportion of correct rejections of culprit-absent lineups. To this end, Winter et al.[3] constructed culprit-absent lineups in which all members had conspicuous birthmarks, which was not the case for the culprit. Given that none of the members in the culprit-absent lineups resembled the culprit, it was relatively easy to detect the absence of the culprit and thus to reject the culprit-absent lineups. As expected, parameter *dA* was significantly higher when the culprit-absent lineups were easy to reject. The obvious and downright trivial manipulation of the detection of culprit absence used by Winter et al. is ideal for the purpose of the model validation. However, it seems also interesting to explore whether there are factors that facilitate culprit-absence detection in more realistic settings. In fact, it would be highly desirable to find methods that actually improve the witnesses' ability to detect the absence of the culprit and to specifically reject culprit-absent lineups, ideally without affecting guessing-based selection. Two such methods are reported below.

The first method is the use of a wildcard, that is, a silhouette with a question mark that represents a 'mystery man' that can be chosen instead of the suspect or one of the fillers. The wildcard option was introduced to make it easier for children to reject culprit-absent lineups by providing an option to reject the lineup that is more equivalent to a positive response when choosing the suspect[70]. In the majority of studies available to date, an intriguing pattern of results has emerged: A wildcard decreases the rate of false identifications in culprit-absent lineups without increasing the rate of false rejections of culprit-present lineups[71–73]. This pattern of results suggests that the effect of using a wildcard does not affect guessing-based selection—in which case it should have increased both correct *and* false lineup rejections—but may specifically improve the detection of the absence of the culprit.

The second method is the use of a culprit-absent practice lineup[74] that has been introduced to facilitate the rejecting of culprit-absent lineups for older adults. Just as the wildcard, the culprit-absent practice lineup has led to an increase in correct culprit-absent lineup rejections without affecting the rate of false culprit-present lineup rejections, suggesting that the underlying process is a facilitation of the detection of the culprit absence and not a decrease in guessing-based selection.

If the 2-HT eyewitness identification model is valid, then the effects of wildcards[70] and culprit-absent practice lineups[74] should be reflected in parameter *dA,* which was designed to measure the detection of the absence of the culprit. We tested this assumption by reanalyzing the data obtained by Karageorge and Zajac[70] and by Wilcock and Bull [[74], Experiment 2].

Effects of a wildcard on culprit-absence detection: Reanalysis of Karageorge and Zajac[70]. The results of a number of studies have shown that children appear to have considerable difficulty rejecting lineups even if the culprit is absent [for a review, see[75]]. Karageorge and Zajac[70] aimed to enhance children's ability to reject culprit-absent lineups by inserting a wildcard within the lineup that could be chosen instead of one of the lineup members. Children were more likely to reject the culprit-absent lineup when a wildcard was provided than when no such option was provided. Interestingly, the rate of correct rejections of culprit-absent lineups increased, whereas the rate of false rejections of culprit-present lineups did not. If a wildcard would simply affect the probability of selecting one of the lineup members based on guessing, it should have affected correct and false rejections equally. The selective effect of a wildcard on correct lineup rejections thus can only be caused by an increased detection of the absence of the culprit. In the 2-HT eyewitness identification model, a wildcard manipulation thus should selectively affect the culprit-absence detection parameter *dA*. Specifically, parameter *dA* should be higher in the wildcard condition than in the control condition.

Method. Karageorge and Zajac[70] randomly assigned younger (age: 5 to 7 years, $n = 101$) and older (age: 8 to 11 years, $n = 109$) children to one of the eight conditions of a 2 (wildcard condition: wildcard vs. control) × 2 (culprit presence: present vs. absent) × 2 (delay: 1 to 2 days vs. 2 weeks) between-subjects design. During a visit to a fire station, the children were exposed for 30 to 45 s to a male confederate (henceforth referred to as the culprit) sliding down a fire pole. Either 1 to 2 days or 2 weeks after the event, all children who stated that they remembered the visit to the fire station ($n = 204$) were presented with a six-person culprit-present or culprit-absent lineup. The culprit-absent lineup contained the same five fillers as the culprit-present lineup, but the culprit was replaced by an innocent suspect who most resembled the culprit while the other fillers also shared basic characteristics with the culprit. The lineup photos were placed on a table in two rows of three. In the wildcard condition, a photo of a silhouette with a superimposed question mark was placed between the two rows. Prior to viewing the photos, all children were given standard two-sided pre-lineup instructions. Children in the control condition were instructed to point to the photo of the culprit if it was present and to tell the experimenter if it was not. Children in the wildcard condition were instructed to point to the photo of the culprit if it was present and to the silhouette (denoted as "this special photo") if it was not [for more details, see[70]].

Results. The observed response frequencies (see the upper half of Table 8) were reconstructed from the proportions reported by Karageorge and Zajac[70] in their Table 1 which the data were already collapsed over the delay conditions (1 to 2 days vs. 2 weeks). Given that the total number of children in each age group in the final sample

	Culprit-present lineups			Culprit-absent lineups		
	Culprit identifications	Filler identifications	Lineup rejections	Innocent-suspect identifications	Filler identifications	Lineup rejections
Karageorge and Zajac[70]						
Wildcard condition						
Wildcard	31	10	7	2	6	41
Control	36	8	9	25	13	16
Wilcock and Bull[74]						
Pre-lineup procedure						
Practice	24	21	5	2	12	36
Control	20	20	10	7	36	7

Table 8. Response frequencies as reconstructed from Table 1 of Karageorge and Zajac [70, p. 173] and from Table 3 of Wilcock and Bull [74, p. 730], see text for details.

was not specified, it was not possible to analyze the data for younger and older children separately. Therefore, we collapsed the data across age groups. According to Karageorge and Zajac [70, p. 168], half of the children were presented with the culprit-present lineup, while the remaining children were presented with the culprit-absent lineup. More specifically, in the control condition, 53 of the 107 children saw the culprit-present lineup and 54 children saw the culprit-absent lineup, whereas in the wildcard condition, 48 of the 97 children saw the culprit-present lineup and 49 children saw the culprit-absent lineup. For the model-based reanalysis, we needed two sets of the model trees depicted in Fig. 1, one for the wildcard condition and one for the control condition. The *b* parameters were set to be equal between the conditions because there was no reason to assume that these parameters should differ as a function of the presence or absence of the wildcard. The model incorporating these restrictions was used as a comparison standard for the subsequent nested likelihood-ratio tests, $G^2(1) = 3.58$, $p = 0.059$. The estimates of the culprit-absence detection parameter *dA* as a function of the wildcard condition are shown in the left panel of Fig. 5. Table 9 shows the estimates of parameters *dP*, *b* and *g*.

The wildcard manipulation affected the culprit-absence detection parameter *dA* as expected. The probability of detecting the absence of the culprit was significantly higher when a wildcard was presented than when it was not, $\Delta G^2(1) = 29.79$, $p < 0.001$. The wildcard manipulation did not affect the probability of detecting the culprit as measured by parameter *dP*, $\Delta G^2(1) = 0.21$, $p = 0.646$. The guessing-based selection parameter *g* was not affected by the wildcard manipulation either, $\Delta G^2(1) = 1.44$, $p = 0.229$.

Effects of a culprit-absent practice lineup on culprit-absence detection: Reanalysis of Wilcock and Bull[74]. Just like children, older adults are less likely to correctly reject a culprit-absent lineup than younger adults [for a meta-analysis, see[76]]. Wilcock and Bull [74, Experiment 2] examined the effect of a culprit-absent practice lineup on correct lineup rejections. Participants in the culprit-absent practice lineup condition correctly rejected culprit-absent lineups more often than participants in the control condition without a culprit-absent practice lineup. Interestingly, the culprit-absent practice lineup did not increase false rejections of culprit-present lineups. This pattern of results suggests that guessing-based selection cannot be responsible for the effect of the culprit-absent practice lineup because guessing-based selection would have decreased the rate of rejections of both culprit-present and culprit-absent lineups. Instead, the culprit-absent practice lineup must have improved the detection of the absence of the culprit because only this explanation is consistent with a selective increase in correct rejections of culprit-absent lineups. Therefore, it can be predicted that the culprit-absent practice manipulation should affect the culprit-absence detection parameter *dA*. Specifically, parameter *dA* should be significantly higher in the culprit-absent practice lineup condition than in the control condition.

Method. Wilcock and Bull [74, Experiment 2] randomly assigned 100 older participants to one of two groups (culprit-absent practice: culprit-absent practice lineup vs. control). Culprit presence (present vs. absent) was a within-subjects factor. Participants were shown a 110-s video of two men breaking into a house. After a 30-min retention interval, half of the participants were presented with a culprit-absent practice lineup consisting of six color pictures of famous women. The participants were asked to identify the Queen of England and were also informed that her face may or may not be present (there was no picture of the Queen). All participants correctly rejected the lineup. After rejecting the lineup, participants were again warned that not all police lineups include the culprit and that even the police can make mistakes. Participants were then given standard two-sided pre-lineup instructions before viewing the real lineups. Wilcock and Bull constructed a culprit-present and a culprit-absent lineup for each of the two culprits consisting of six faces in a 3×2 array. Participants were shown one culprit-present and one culprit-absent lineup (i.e., the first participant saw a culprit-present lineup for culprit 1, followed by a culprit-absent lineup for culprit 2, the second participant saw a culprit-absent lineup for culprit 1 and a culprit-present lineup for culprit 2 and so on). For the sake of simplicity, we did not distinguish between the two culprits but aggregated the data. All fillers matched the culprits' descriptions. In the culprit-absent lineups, the culprit was replaced by the filler who was rated as most similar-looking to the culprit. The culprit and the culprit replacement were randomly placed across all six lineup positions [for more details, see[74], Experiment 2].

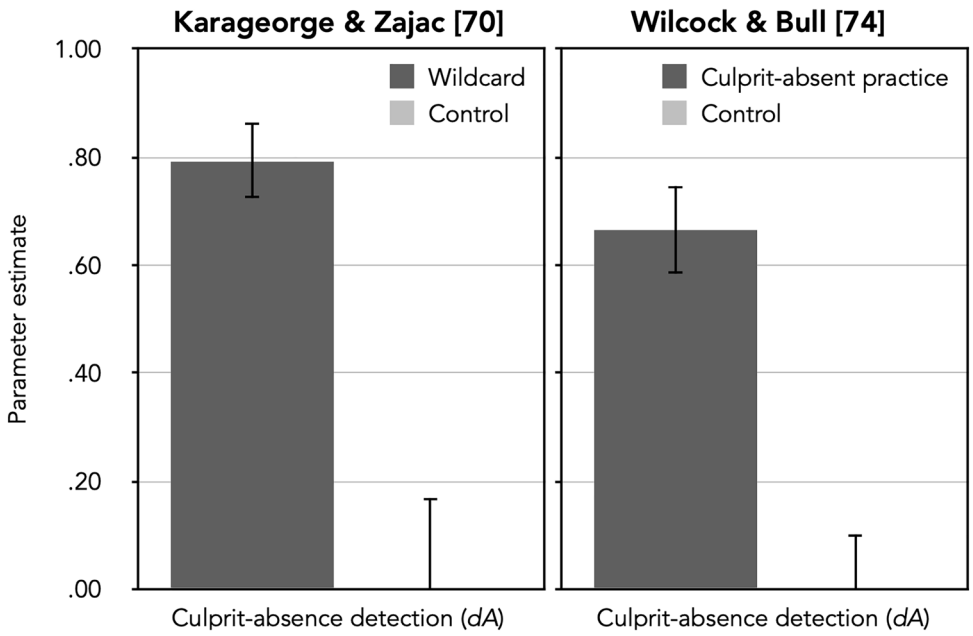

Figure 5. Estimates of parameter *dA* (representing the probability of detecting the absence of the culprit) of the 2-HT eyewitness identification model when applied to the data reported by Karageorge and Zajac[70] and by Wilcock and Bull [74, Experiment 2] as a function of the wildcard condition (wildcard vs. control; left panel) and the pre-lineup procedure (culprit-absent practice vs. control; right panel). The error bars represent standard errors.

Karageorge and Zajac[70]				Wilcock and Bull[74]			
	Parameter estimates				Parameter estimates		
Wildcard condition	*dP*	*b*	*g*	Pre-lineup procedure	*dP*	*b*	*g*
Wildcard	.37 (.15)	.37 (.09)	.67 (.10)	Practice	.40 (.09)	.00 (.05)	.83 (.07)
Control	.45 (.14)		.51 (.11)	Control	.31 (.09)		.80 (.06)

Table 9. Estimates of parameters *dP*, *b* and *g* of the 2-HT eyewitness identification model for the data reported by Karageorge and Zajac[70] and by Wilcock and Bull [74, Experiment 2]. Values in parentheses represent standard errors. Within the model used as a comparison standard, parameters *dP* and *g* were estimated separately for the wildcard and the control conditions[70] and for the culprit-absent practice and the control conditions [74, Experiment 2]. Parameter *b* was set to be equal between the experimental conditions. *dP* = probability of detecting the presence of the culprit; *b* = probability of biased suspect selection; *g* = probability of guessing-based selection among the lineup members.

Results. The observed response frequencies (see the lower half of Table 8) were taken from Table 3 of Wilcock and Bull[74]. As in the previous reanalyses, we estimated the number of innocent-suspect identifications using the standard procedure of dividing the total number of false identifications in culprit-absent lineups by the number of lineup members[47,48]. Two sets of the model trees depicted in Fig. 1 were needed, one for the culprit-absent practice lineup condition and one for the control condition. The *b* parameters were set to be equal between the conditions because there was no reason to assume that these parameters should differ as a function of the culprit-absent practice manipulation. The model incorporating these restrictions was used as a comparison standard for the subsequent nested likelihood-ratio tests, $G^2(1) = 2.92$, $p = 0.087$. The estimates of culprit-absence detection parameter *dA* as a function of the culprit-absent practice manipulation are shown in the right panel of Fig. 5. Table 9 shows the estimates of parameters *dP*, *b* and *g*.

The culprit-absence detection parameter *dA* was significantly higher in the culprit-absent practice lineup condition than in the control condition, $\Delta G^2(1) = 23.42$, $p < 0.001$. By contrast, the detection of the culprit presence reflected in parameter *dP* remained unaffected by the culprit-absent practice manipulation, $\Delta G^2(1) = 0.58$, $p = 0.444$. The same was true for the guessing-based selection parameter *g*, $\Delta G^2(1) = 0.18$, $p = 0.673$.

Discussion. The two final reanalyses demonstrated that manipulations designed to facilitate culprit-absence detection selectively affected parameter *dA* of the 2-HT eyewitness identification model. Karageorge and Zajac[70] inserted a wildcard within the lineup. Wilcock and Bull [74, Experiment 2] presented a culprit-absent practice

lineup. Both procedures selectively increased the rate of correct rejections of culprit-absent lineups but did not affect the rate of false rejections of culprit-present lineups, suggesting that introducing these procedures did not induce decreased guessing-based selection. The latter conclusion is consistent with the fact that guessing-based selection parameter *g* was not affected by either the wildcard or the culprit-absent practice lineup.

General discussion

Here, we report a validation of the novel 2-HT eyewitness identification model using published data. The model simultaneously takes into account all of the data categories that are observed in lineups, that is, correct culprit identifications, false innocent-suspect identifications, false filler identifications in culprit-present and culprit-absent lineups, false rejections of culprit-present lineups and correct rejections of culprit-absent lineups. Based on these data, the model yields measures of latent processes that underlie eyewitness identification performance. Specifically, the model is designed to distinguish between two types of detection processes—detection of the presence of the culprit and detection of the absence of the culprit—as well as two different types of non-detection-based decision processes—biased selection of a suspect that stands out from the fillers in unfair lineups and selecting a lineup member based on guessing. We hope that distinguishing between these qualitatively different latent processes helps to improve the clarity of the interpretation of lineup data. A typical approach is to try to infer the underlying processes indirectly from surface-level data by comparing the rate of correct culprit identifications and false innocent-suspect identifications between different conditions, but often the problem arises that the same manipulation can simultaneously affect qualitatively different processes such as, for example, both culprit-presence detection and guessing-based selection[8,13]. The 2-HT eyewitness identification model serves to disentangle the effects of manipulations on the processes underlying eyewitness identifications by yielding separate measures of these latent processes and by enabling researchers to test hypotheses directly at the level of the parameters representing these latent processes. In that way, the model can be used to tackle important research questions such as, for example, the question of whether the process of culprit-presence and culprit-absence detection, biased selection and guessing-based selection differ between simultaneous and sequential lineups. However, before a measurement model such as the 2-HT eyewitness identification model can be used to tackle new and unresolved empirical questions, it is important to demonstrate that the model parameters sensitively reflect the processes they are intended to measure[19,29].

In a separate empirical contribution, Winter et al.[3] have reported fresh experiments that were designed with the sole purpose of testing whether the parameters of the 2-HT eyewitness identification model sensitively reflect specific manipulations that were carefully crafted to affect the latent processes underlying eyewitness identifications postulated by the model; these tests were successful. The analyses reported here complement the results reported by Winter et al. by demonstrating that the model parameters also sensitively reflect manipulations of the processes they were designed to measure in published studies obtained with a wide variety of different experimental manipulations, samples, paradigms and laboratory as well as real-world settings. First, it was demonstrated that parameter *dP* reacts sensitively to manipulations of encoding conditions that can be expected to affect the detection of the presence of the culprit. Specifically, the culprit-presence detection parameter *dP* was higher in response to long as opposed to short culprit exposure and under good as opposed to poor viewing conditions at encoding. Second, manipulations of lineup fairness affected the estimate of parameter *b*, designed to reflect the process of the biased selection of a suspect who stands out from the fillers. As expected under the assumption that this model parameter is valid, parameter *b* was higher for lineups with low as opposed to high suspect-filler similarity and for lineups in which the suspects had unique facial features distinguishing them from the fillers as opposed to lineups in which these distinctive features were concealed in all photos or replicated in the photos of the fillers. Third, manipulations of pre-lineup instructions had the predicted effects on the parameter representing the selection of a lineup member based on guessing. Parameter *g* was consistently higher when one-sided than when two-sided instructions were used. One-sided instructions implicate that the culprit is in the lineup, whereas two-sided instructions emphasize that the culprit may be present or absent. It should be uncontroversial that difficult encoding conditions should affect the process of culprit-presence detection, that biased suspect selection should be increased in unfair lineups and that participants can be discouraged to select a lineup member based on guessing by two-sided lineup instructions. Selecting manipulations that obviously affect the detection of the absence of the culprit is somewhat more difficult. Winter et al. have provided evidence showing that the detection of culprit absence is enhanced when all of the lineup members in the culprit-absent lineup can be easily ruled out based on salient perceptual features. Such a manipulation is unlikely to be found in the literature. This is so because such an experiment makes sense only in the context of model validation in which the ideal manipulation is obvious and trivial in the sense that there is broad agreement on the manipulation's effect on certain cognitive processes. Therefore, our reanalysis-based test of the validity of the culprit-absence detection parameter concentrated on manipulations that were specifically designed to help children and older adults to reject culprit-absent lineups[70,74]. In previous studies these manipulations led to an increase in correct rejections of culprit-absent lineups while the false rejections of culprit-present lineups remained unaffected, suggesting that the affected process was the detection of the absence of the culprit and not the selection of a lineup member based on guessing. The present reanalysis confirms this conclusion and thereby provides further evidence for the process of culprit-absence detection: The use of a wildcard procedure and a culprit-absent practice lineup increased the estimates of culprit-absence detection parameter *dA* compared to standard lineup procedures.

The 2-HT eyewitness identification model distinguishes between two non-detection-based judgements: a process of selecting a lineup member based on guessing (parameter *g*) and a process of biased selection of a suspect who stands out from the fillers (parameter *b*). Traditionally, eyewitness researchers have relied on the mock-witness task to determine the degree to which a lineup is unfair. This task involves presenting a lineup and a description of a culprit to participants who did not witness the crime—so called mock witnesses—and

then asking the mock witnesses to identify the person who best matches the description[77]. Based on the answers of the mock witnesses, various lineup fairness measures can be calculated, reflecting either the lineup size (i.e., how many lineup members have plausibility as the culprit?) or biased selection of the suspect (i.e., to what extent does the suspect stand out from other lineup members?[78,79]). However, some researchers have cautioned that these measures may suffer from low validity and reliability and have suggested to consider alternative methods for estimating lineup fairness[80]. The 2-HT eyewitness identification model offers such an alternative method for measuring and testing hypotheses about biased suspect selection. In fact, the 2-HT eyewitness identification model measures lineup fairness directly from the witnesses' identification data—without relying on a separate paradigm involving mock witnesses. Based on this direct measurement, one can draw conclusions about the cognitive processes that determine the decisions of eyewitnesses that may, to some degree, differ from those of mock witnesses[81].

The model also makes it possible to distinguish between two detection processes: the detection of the presence of the culprit (parameter *dP*) and the detection of the absence of the culprit (parameter *dA*). Traditional measures of lineup performance often provide only a single accuracy index that simultaneously accounts for the witness's performance in culprit-present and culprit-absent lineups. However, it is possible to argue that detecting the presence of a culprit might be achieved by a different underlying process than detecting the absence of a culprit. The first piece of evidence supporting this argument is that the process of detecting the presence of the culprit varies as a function of manipulations that leave the process of detecting the absence of the culprit unaffected (see Fig. 2 and Table 3 above) and vice versa (see Fig. 5 and Table 9 above). The second piece of evidence is that in most of the data sets presented here, the probability of detecting the presence of the culprit was quite high, while the probability of detecting the absence of the culprit was considerably lower (the statistical test of a difference between *dP* and *dA* results in $p < 0.05$ for all reanalyses reported in the present article). Interestingly, it seems plausible that techniques developed to help children and older adults to reject culprit-absent lineups specifically affect the process of culprit-absence detection. However, without the help of a model-based analysis, Karageorge and Zajac[70] and Wilcock and Bull[74] had to rely on the observation that a wildcard procedure or a culprit-absent lineup practice selectively increased correct rejections of culprit-absent lineups with no effect on false rejections of culprit-present lineups, which indirectly suggests that these procedures may help to detect culprit absence and do not decrease guessing-based selection. With the 2-HT eyewitness identification model it is less indirect and more straightforward to conclude that these procedures enhance the detection of the culprit absence (parameter *dA*) while leaving guessing-based selection (parameter *g*) unaffected. We hope that by including a separate parameter for culprit-absence detection, the 2-HT eyewitness identification model will stimulate more research on techniques that specifically improve witnesses' ability to detect the absence of a culprit, which seems highly desirable.

A limitation of the present reanalyses is that many studies in the eyewitness literature used only fillers in culprit-absent lineups [e.g.,[45,46,67,68,74]]. However, the model-based analyses make use of the full 2×3 data structure, which includes innocent-suspect identifications. To be able to perform a model-based analysis of experiments without a designated innocent suspect in culprit-absent lineups, we followed the standard procedure for estimating the rate of innocent-suspect identifications from the filler identifications by dividing the total number of false identifications in culprit-absent lineups by the number of lineup members[47,48]. This method rests on the assumption that the culprit-absent lineups contained an innocent suspect that was selected with the same probability as each of the fillers, which implies that the lineups were fair. Importantly, this assumption does not seem to affect the major conclusions that can be drawn from the data as the present reanalysis is consistent with the experimental validation study by Winter et al.[3] in which we used designated innocent suspects in culprit-absent lineups. However, to validly measure lineup fairness and to increase the ecological validity of the analysis, we encourage researchers to include a designated innocent suspect in culprit-absent lineups in future studies. In the real world, the photographs of the suspects whose guilt or innocence is unknown to the police are often taken from different sources (e.g., social media) than the photographs of the fillers (e.g., a database). Photographs from different sources may differ systematically in certain characteristics. It thus greatly improves the ecological validity of the results to have a designated innocent suspect whose photograph deviates from the photographs of the fillers in the culprit-absent lineups in the same way as the photograph of the culprit deviates from the photographs of the fillers in the culprit-present lineups [cf.[3]].

Due to their mathematical and conceptual simplicity, the class of MPT models is ideally suited to develop simple formal measurement models to assess latent processes involved in specific empirical paradigms[18,19,29]. However, a more controversial property of MPT models is that they involve a threshold concept, thus assuming that recognition judgements result from discrete memory states rather than continuously distributed signal strength (as the SDT assumes). While some researchers argued that the threshold assumption is inconsistent with the available empirical evidence[82,83], others have shown that threshold models and SDT-based models can both account for recognition memory performance[37,84,85]. The theoretical and practical usefulness of both approaches in the field of eyewitness identifications has to be further investigated in future research.

Given the well-documented evidence illustrating the fallibility of eyewitness testimonies, it remains an important goal to advance our knowledge about the latent processes underlying eyewitness identification decisions. We hope to contribute to this advancement by presenting a multinomial model for analyzing lineup performance, the 2-HT eyewitness identification model. By incorporating the entire 2×3 data structure of responses in lineup identification tasks, the model enables inferences about latent cognitive processes that are not accessible when standard measures of eyewitness identification accuracy are used. Here, we tested the validity of the model by applying it to published data. A series of eight reanalyses provides evidence in support of a successful model validation. Validations of culprit-presence detection (*dP*), biased suspect selection (*b*), guessing-based selection among the lineup members (*g*) and culprit-absence detection (*dA*) showed that the parameters sensitively reflected experimental manipulations of the processes they were designed to measure. We conclude that the

2-HT eyewitness identification model is promising and can complement existing tools to analyze eyewitness identifications in lineups.

Data availability

The multiTree equations and data files of all reanalyses are available in the OSF repository, https://osf.io/pjc5b.

Received: 18 February 2022; Accepted: 25 July 2022
Published online: 04 August 2022

References

1. Innocence Project. *Exonerate the Innocent.* https://innocenceproject.org/exonerate/ (2022).
2. Clark, S. E. Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspect. Psychol. Sci.* **7**, 238–259. https://doi.org/10.1177/1745691612439584 (2012).
3. Winter, K., Menne, N. M., Bell, R., Buchner, A. Experimental validation of a multinomial processing tree model for analyzing eyewitness identification decisions. Manuscript submitted for publication (2022).
4. Wells, G. L. & Lindsay, R. C. L. On estimating the diagnosticity of eyewitness nonidentifications. *Psychol. Bull.* **88**, 776–784. https://doi.org/10.1037/0033-2909.88.3.776 (1980).
5. Wells, G. L. Eyewitness identification: Probative value, criterion shifts, and policy regarding the sequential lineup. *Curr. Dir. Psychol. Sci.* **23**, 11–16. https://doi.org/10.1177/0963721413504781 (2014).
6. Meissner, C. A., Tredoux, C. G., Parker, J. F. & MacLin, O. H. Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Mem. Cognit.* **33**, 783–792. https://doi.org/10.3758/BF03193074 (2005).
7. Gronlund, S. D. *et al.* Showups versus lineups: An evaluation using ROC analysis. *J. Appl. Res. Mem. Cogn.* **1**, 221–228. https://doi.org/10.1016/j.jarmac.2012.09.003 (2012).
8. Palmer, M. A. & Brewer, N. Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law Hum. Behav.* **36**, 247–255. https://doi.org/10.1037/h0093923 (2012).
9. Gronlund, S. D., Wixted, J. T. & Mickes, L. Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Curr. Dir. Psychol. Sci.* **23**, 3–10. https://doi.org/10.1177/0963721413498891 (2014).
10. Wixted, J. T. & Mickes, L. A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychol. Rev.* **121**, 262–276. https://doi.org/10.1037/a0035940 (2014).
11. Wixted, J. T. & Mickes, L. The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspect. Psychol. Sci.* **7**, 275–278. https://doi.org/10.1177/1745691612442906 (2012).
12. Macmillan, N. A. & Creelman, C. D. *Detection Theory: A User's Guide* (Lawrence Erlbaum Associates, Mahwah, 2005).
13. Mickes, L., Flowe, H. D. & Wixted, J. T. Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *J. Exp. Psychol. Appl.* **18**, 361–376. https://doi.org/10.1037/a0030609 (2012).
14. Smith, A. M., Wells, G. L., Lindsay, R. C. L. & Penrod, S. D. Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law Hum. Behav.* **41**, 127–145. https://doi.org/10.1037/lhb0000219 (2017).
15. Mickes, L., Moreland, M. B., Clark, S. E. & Wixted, J. T. Missing the information needed to perform ROC analysis? Then compute d′, not the diagnosticity ratio. *J. Appl. Res. Mem. Cogn.* **3**, 58–62. https://doi.org/10.1016/j.jarmac.2014.04.007 (2014).
16. Wells, G. L., Smalarz, L. & Smith, A. M. ROC analysis of lineups does not measure underlying discriminability and has limited value. *J. Appl. Res. Mem. Cogn.* **4**, 313–317. https://doi.org/10.1016/j.jarmac.2015.08.008 (2015).
17. Wells, G. L., Smith, A. M. & Smalarz, L. ROC analysis of lineups obscures information that is critical for both theoretical understanding and applied purposes. *J. Appl. Res. Mem. Cogn.* **4**, 324–328. https://doi.org/10.1016/j.jarmac.2015.08.010 (2015).
18. Batchelder, W. H. & Riefer, D. M. Theoretical and empirical review of multinomial process tree modeling. *Psychon. Bull. Rev.* **6**, 57–86. https://doi.org/10.3758/BF03210812 (1999).
19. Erdfelder, E. *et al.* Multinomial processing tree models: A review of the literature. *Z. Psychol./J. Psychol.* **217**, 108–124. https://doi.org/10.1027/0044-3409.217.3.108 (2009).
20. Meiser, T. & Bröder, A. Memory for multidimensional source information. *J. Exp. Psychol. Learn. Mem. Cogn.* **28**, 116–137. https://doi.org/10.1037/0278-7393.28.1.116 (2002).
21. Smith, R. E. & Bayen, U. J. A multinomial model of event-based prospective memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **30**, 756–777. https://doi.org/10.1037/0278-7393.30.4.756 (2004).
22. Stahl, C. & Klauer, K. C. A simplified conjoint recognition paradigm for the measurement of gist and verbatim memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **34**, 570–586. https://doi.org/10.1037/0278-7393.34.3.570 (2008).
23. Bayen, U. J. & Kuhlmann, B. G. Influences of source–item contingency and schematic knowledge on source monitoring: Tests of the probability-matching account. *J. Mem. Lang.* **64**, 1–17. https://doi.org/10.1016/j.jml.2010.09.001 (2011).
24. Erdfelder, E. & Buchner, A. Decomposing the hindsight bias: A multinomial processing tree model for separating recollection and reconstruction in hindsight. *J. Exp. Psychol. Learn. Mem. Cogn.* **24**, 387–414. https://doi.org/10.1037/0278-7393.24.2.387 (1998).
25. Unkelbach, C. & Stahl, C. A multinomial modeling approach to dissociate different components of the truth effect. *Conscious. Cogn.* **18**, 22–38. https://doi.org/10.1016/j.concog.2008.09.006 (2009).
26. Mieth, L., Buchner, A. & Bell, R. Moral labels increase cooperation and costly punishment in a Prisoner's Dilemma game with punishment option. *Sci. Rep.* **11**, 10221. https://doi.org/10.1038/s41598-021-89675-6 (2021).
27. Wagenaar, W. A. & Boer, J. P. A. Misleading postevent information: Testing parameterized models of integration in memory. *Acta Psychol.* **66**, 291–306. https://doi.org/10.1016/0001-6918(87)90040-0 (1987).
28. Singmann, H. & Kellen, D. MPTinR: Analysis of multinomial processing tree models in R. *Behav. Res. Methods* **45**, 560–575. https://doi.org/10.3758/s13428-012-0259-0 (2013).
29. Schmidt, O., Erdfelder, E., Heck, D. W. Tutorial on multinomial processing tree modeling: How to develop, test, and extend MPT models. *PsyArXiv*, https://doi.org/10.31234/osf.io/gh8md. (2022).
30. Hu, X. & Batchelder, W. H. The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika* **59**, 21–47. https://doi.org/10.1007/BF02294263 (1994).
31. Riefer, D. M. & Batchelder, W. H. Multinomial modeling and the measurement of cognitive processes. *Psychol. Rev.* **95**, 318–339. https://doi.org/10.1037/0033-295X.95.3.318 (1988).
32. Moshagen, M. multiTree: A computer program for the analysis of multinomial processing tree models. *Behav. Res. Methods* **42**, 42–54. https://doi.org/10.3758/BRM.42.1.42 (2010).
33. Stahl, C. & Klauer, K. C. HMMTree: A computer program for latent-class hierarchical multinomial processing tree models. *Behav. Res. Methods* **39**, 267–273. https://doi.org/10.3758/BF03193157 (2007).
34. Buchner, A., Erdfelder, E. & Vaterrodt-Plünnecke, B. Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *J. Exp. Psychol. Gen.* **124**, 137–160. https://doi.org/10.1037/0096-3445.124.2.137 (1995).
35. Bayen, U. J., Murnane, K. & Erdfelder, E. Source discrimination, item detection, and multinomial models of source monitoring. *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 197–215. https://doi.org/10.1037/0278-7393.22.1.197 (1996).

36. Innocence Project. *Marvin Anderson*. https://innocenceproject.org/cases/marvin-anderson/ (2022).
37. Snodgrass, J. G. & Corwin, J. Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *J. Exp. Psychol. Gen.* **117**, 34–50. https://doi.org/10.1037//0096-3445.117.1.34 (1988).
38. Erdfelder, E., Cüpper, L., Auer, T.-S. & Undorf, M. The four-states model of memory retrieval experiences. *Z. Psychol./J. Psychol.* **215**, 61–71. https://doi.org/10.1027/0044-3409.215.1.61 (2007).
39. Bell, R., Mieth, L. & Buchner, A. Emotional memory: No source memory without old–new recognition. *Emotion* **17**, 120–130. https://doi.org/10.1037/emo0000211 (2017).
40. Bröder, A., Kellen, D., Schütz, J. & Rohrmeier, C. Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory* **21**, 916–944. https://doi.org/10.1080/09658211.2013.767348 (2013).
41. Swets, J. A. Is there a sensory threshold? When the effects of the observer's response criterion are isolated, a sensory limitation is not evident. *Science* **134**, 168–177. https://doi.org/10.1126/science.134.3473.168 (1961).
42. Bornstein, B. H., Deffenbacher, K. A., Penrod, S. D. & McGorty, E. K. Effects of exposure time and cognitive operations on facial identification accuracy: A meta-analysis of two variables associated with initial memory strength. *Psychol. Crime Law* **18**, 473–490. https://doi.org/10.1080/1068316X.2010.508458 (2012).
43. Valentine, T., Pickering, A. & Darling, S. Characteristics of eyewitness identification that predict the outcome of real lineups. *Appl. Cogn. Psychol.* **17**, 969–993. https://doi.org/10.1002/acp.939 (2003).
44. Palmer, M. A., Brewer, N., Weber, N. & Nagesh, A. The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *J. Exp. Psychol. Appl.* **19**, 55–71. https://doi.org/10.1037/a0031602 (2013).
45. Memon, A., Hope, L. & Bull, R. Exposure duration: Effects on eyewitness accuracy and confidence. *Br. J. Psychol.* **94**, 339–354. https://doi.org/10.1348/000712603767876262 (2003).
46. Smith, A. M. Why do mistaken identification rates increase when either witnessing or testing conditions get worse?. *J. Appl. Res. Mem. Cogn.* **9**, 495–507. https://doi.org/10.1016/j.jarmac.2020.08.002 (2020).
47. Mickes, L. Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *J. Appl. Res. Mem. Cog.* **4**, 93–102. https://doi.org/10.1016/j.jarmac.2015.01.003 (2015).
48. Colloff, M. F., Wade, K. A. & Strange, D. Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychol. Sci.* **27**, 1227–1239. https://doi.org/10.1177/0956797616655789 (2016).
49. Riefer, D. M., Hu, X. & Batchelder, W. H. Response strategies in source monitoring. *J. Exp. Psychol. Learn. Mem. Cogn.* **20**, 680–693. https://doi.org/10.1037/0278-7393.20.3.680 (1994).
50. Hirshman, E. Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *J. Exp. Psychol. Learn. Mem. Cogn.* **21**, 302–313. https://doi.org/10.1037/0278-7393.21.2.302 (1995).
51. Ehrenberg, K. & Klauer, K. C. Flexible use of source information: Processing components of the inconsistency effect in person memory. *J. Exp. Soc. Psychol.* **41**, 369–387. https://doi.org/10.1016/j.jesp.2004.08.001 (2005).
52. Meiser, T., Sattler, C. & Von Hecker, U. Metacognitive inferences in source memory judgements: The role of perceived differences in item recognition. *Q. J. Exp. Psychol.* **60**, 1015–1040. https://doi.org/10.1080/17470210600875215 (2007).
53. Batchelder, W. H. & Batchelder, E. Metacognitive guessing strategies in source monitoring. In *Handbook of Metamemory and Memory* (eds Dunlosky, J. & Bjork, R. A.) 211–244 (Psychology Press, London, 2008).
54. Küppers, V. & Bayen, U. J. Inconsistency effects in source memory and compensatory schema-consistent guessing. *Q. J. Exp. Psychol.* **67**, 2042–2059. https://doi.org/10.1080/17470218.2014.904914 (2014).
55. Smith, A. M., Wilford, M. M., Quigley-McBride, A. & Wells, G. L. Mistaken eyewitness identification rates increase when either witnessing or testing conditions get worse. *Law Hum. Behav.* **43**, 358–368. https://doi.org/10.1037/lhb0000334 (2019).
56. Freire, A., Lee, K., Williamson, K. S., Stuart, S. J. & Lindsay, R. Lineup identification by children: Effects of clothing bias. *Law Hum. Behav.* **28**, 339–354. https://doi.org/10.1023/B:LAHU.0000029142.00834.e3 (2004).
57. Saraiva, R. B. *et al.* Eyewitness metamemory predicts identification performance in biased and unbiased line-ups. *Leg. Criminol. Psychol.* **25**, 111–132. https://doi.org/10.1111/lcrp.12166 (2020).
58. Steblay, N. K. & Wells, G. L. Assessment of bias in police lineups. *Psychol. Public Policy Law* **26**, 393–412. https://doi.org/10.1037/law0000287 (2020).
59. Wetmore, S. A. *et al.* Effect of retention interval on showup and lineup performance. *J. Appl. Res. Mem. Cogn.* **4**, 8–14. https://doi.org/10.1016/j.jarmac.2014.07.003 (2015).
60. Gronlund, S. D., Carlson, C. A., Dailey, S. B. & Goodsell, C. A. Robustness of the sequential lineup advantage. *J. Exp. Psychol. Appl.* **15**, 140–152. https://doi.org/10.1037/a0015082 (2009).
61. Fitzgerald, R. J., Oriet, C. & Price, H. L. Suspect filler similarity in eyewitness lineups: A literature review and a novel methodology. *Law Hum. Behav.* **39**, 62–74. https://doi.org/10.1037/lhb0000095 (2015).
62. Smith, A. M., Wells, G. L., Smalarz, L. & Lampinen, J. M. 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). *Psychol. Sci.* **29**, 1548–1551. https://doi.org/10.1177/0956797617698528 (2018).
63. Horry, R., Memon, A., Wright, D. B. & Milne, R. Predictors of eyewitness identification decisions from video lineups in England: A field study. *Law Hum. Behav.* **36**, 257–265. https://doi.org/10.1037/h0093959 (2012).
64. Clark, S. E. A re-examination of the effects of biased lineup instructions in eyewitness identification. *Law Hum. Behav.* **29**, 575–604. https://doi.org/10.1007/s10979-005-7121-1 (2005).
65. Brewer, N. & Wells, G. L. The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *J. Exp. Psychol. Appl.* **12**, 11–30. https://doi.org/10.1037/1076-898X.12.1.11 (2006).
66. Keast, A., Brewer, N. & Wells, G. L. Children's metacognitive judgments in an eyewitness identification task. *J. Exp. Child Psychol.* **97**, 286–314. https://doi.org/10.1016/j.jecp.2007.01.007 (2007).
67. Malpass, R. S. & Devine, P. G. Eyewitness identification: Lineup instructions and the absence of the offender. *J. Appl. Psychol.* **66**, 482–489. https://doi.org/10.1037/0021-9010.66.4.482 (1981).
68. Lampinen, J. M. *et al.* Comparing detailed and less detailed pre-lineup instructions. *Appl. Cogn. Psychol.* **34**, 409–424. https://doi.org/10.1002/acp.3627 (2020).
69. Technical Working Group for Eyewitness Evidence. *Eyewitness Evidence: A Guide for Law Enforcement* (National Institute of Justice, 1999).
70. Karageorge, A. & Zajac, R. Exploring the effects of age and delay on children's person identifications: Verbal descriptions, lineup performance, and the influence of wildcards. *Br. J. Psychol.* **102**, 161–183. https://doi.org/10.1348/000712610X507902 (2011).
71. Zajac, R. & Karageorge, A. The wildcard: A simple technique for improving children's target-absent lineup performance. *Appl. Cogn. Psychol.* **23**, 358–368. https://doi.org/10.1002/acp.1511 (2009).
72. Havard, C. & Memon, A. The mystery man can help reduce false identification for child witnesses: Evidence from video line-ups. *Appl. Cogn. Psychol.* **27**, 50–59. https://doi.org/10.1002/acp.2870 (2013).
73. Havard, C., Laybourn, P. & Klecha, B. The mystery man can increase the reliability of eyewitness identifications for older adult witnesses. *J. Police Crim. Psychol.* **32**, 214–224. https://doi.org/10.1007/s11896-016-9214-9 (2017).
74. Wilcock, R. & Bull, R. Novel lineup methods for improving the performance of older eyewitnesses. *Appl. Cogn. Psychol.* **24**, 718–736. https://doi.org/10.1002/acp.1582 (2010).

75. Havard, C. Are children less reliable at making visual identifications than adults? A review. *Psychol. Crime Law* **20**, 372–388. https://doi.org/10.1080/1068316X.2013.793334 (2014).
76. Erickson, W. B., Lampinen, J. M. & Moore, K. N. Eyewitness identifications by older and younger adults: A meta-analysis and discussion. *J. Police Crim. Psychol.* **31**, 108–121. https://doi.org/10.1007/s11896-015-9176-3 (2016).
77. Doob, A. N. & Kirshenbaum, H. M. Bias in police lineups – Partial remembering. *J. Policy Sci. Admin.* **1**, 287–293 (1973).
78. Malpass, R. S. & Lindsay, R. C. L. Measuring lineup fairness. *Appl. Cogn. Psychol.* **13**, S1–S7. https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+%3cS1::AIDACP678%3e3.0.CO;2-9 (1999).
79. Malpass, R. S., Tredoux, C. G. & McQuiston-Surrett, D. E. Lineup construction and lineup fairness. In *The Handbook of Eyewitness Psychology, Vol. 2. Memory for People* (eds Lindsay, R. C. L. *et al.*) 155–178 (Lawrence Erlbaum Associates, Mahwah, 2007).
80. Mansour, J. K., Beaudry, J. L., Kalmet, N., Bertrand, M. I. & Lindsay, R. C. L. Evaluating lineup fairness: Variations across methods and measures. *Law Hum. Behav.* **41**, 103–115. https://doi.org/10.1037/lhb0000203 (2017).
81. Corey, D., Malpass, R. S. & McQuiston, D. E. Parallelism in eyewitness and mock witness identifications. *Appl. Cogn. Psychol.* **13**, S41–S58. https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+%3cS41::AID-ACP632%3e3.0.CO;2-A (1999).
82. Wixted, J. T. Dual-process theory and signal-detection theory of recognition memory. *Psychol. Rev.* **114**, 152–176. https://doi.org/10.1037/0033-295X.114.1.152 (2007).
83. Yonelinas, A. P. & Parks, C. M. Receiver operating characteristics (ROCs) in recognition memory: a review. *Psychol. Bull.* **133**, 800–832. https://doi.org/10.1037/0033-2909.133.5.800 (2007).
84. Malmberg, K. J. On the form of ROCs constructed from confidence ratings. *J. Exp. Psychol. Learn. Mem. Cogn.* **28**, 380–387. https://doi.org/10.1037/0278-7393.28.2.380 (2002).
85. Bröder, A. & Schütz, J. Recognition ROCs are curvilinear—or are they? On premature arguments against the two-high-threshold model of recognition. *J. Exp. Psychol. Learn. Mem. Cogn.* **35**, 587–606. https://doi.org/10.1037/a0015279 (2009).

Author contributions

N.M.M., K.W., R.B., and A.B. contributed to the idea and conception of the paper. N.M.M. reanalyzed the data and wrote the manuscript with subsequent input from all co-authors. All authors contributed through discussion and interpretation of the results and gave final approval for publication.

Funding

Open Access funding enabled and organized by Projekt DEAL. The work reported herein was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – BU 945/10-1, project number 456214986.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to N.M.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

scientific reports

Check for updates

OPEN

Measuring lineup fairness from eyewitness identification data using a multinomial processing tree model

Nicola Marie Menne✉, Kristina Winter, Raoul Bell & Axel Buchner

The mock-witness task is typically used to evaluate the fairness of lineups. However, the validity of this task has been questioned because there are substantial differences between the tasks for mock witnesses and eyewitnesses. Unlike eyewitnesses, mock witnesses must select a person from the lineup and are alerted to the fact that one lineup member might stand out from the others. It therefore seems desirable to base conclusions about lineup fairness directly on eyewitness data rather than on mock-witness data. To test the importance of direct measurements of biased suspect selection in eyewitness identification decisions, we assessed the fairness of lineups containing either morphed or non-morphed fillers using both mock witnesses and eyewitnesses. We used Tredoux's *E* and the proportion of suspect selections to measure lineup fairness from mock-witness choices and the two-high threshold eyewitness identification model to measure the biased selection of the suspects directly from eyewitness identification decisions. Results obtained in the mock-witness task and the model-based analysis of data obtained in the eyewitness task converged in showing that simultaneous lineups with morphed fillers were significantly more unfair than simultaneous lineups with non-morphed fillers. However, mock-witness and eyewitness data converged only when the eyewitness task mimicked the mock-witness task by including pre-lineup instructions that (1) discouraged eyewitnesses to reject the lineups and (2) alerted eyewitnesses that a photograph might stand out from the other photographs in the lineup. When a typical eyewitness task was created by removing these two features from the pre-lineup instructions, the morphed fillers no longer lead to unfair lineups. These findings highlight the differences in the cognitive processes of mock witnesses and eyewitnesses and they demonstrate the importance of measuring lineup fairness directly from eyewitness identification decisions rather than indirectly using the mock-witness task.

Mistaken eyewitness identification is a consistent and leading cause of wrongful convictions. In the United States, eyewitness misidentifications have contributed to 70 % of the more than 375 wrongful convictions uncovered by DNA-based exonerations[1]. One reason for wrongful convictions is that unfair lineups increase the likelihood of misidentifications of innocent suspects[2,3]. A lineup is considered fair when all fillers (distractors who are known to be innocent) serve as plausible alternatives to the suspect in the lineup such that there is no way to distinguish the suspect from the other lineup members without relying on memory for the culprit. Fair lineups provide protection of the innocent suspect because good fillers siphon misidentifications away from the innocent suspect[4,5]. This protective mechanism is absent in unfair lineups in which the suspect stands out from the other lineup members based on physical appearance or other distinct characteristics of the suspect's photograph[4,6]. It is clear from prior studies that unfair lineups dramatically increase the risk of mistakenly identifying the suspect in comparison to fair lineups[2,3,7,8]. For this reason, it is important to understand the numerous factors that can influence lineup fairness. However, progress will only be made when the fairness of a lineup is measured in a valid way.

Eyewitness researchers have typically used the mock-witness task[9] to assess lineup fairness[10]. In this task, persons who did not witness the crime—so-called mock witnesses—are asked to view the lineup and to choose the lineup member they believe to be the police suspect. One possibility is that mock witnesses are provided with the witness's description of the culprit as the basis for their choices [e.g.,[11,12]]. Alternatively, mock witnesses are not provided with any additional information other than the indication that the suspect might stand

Department of Experimental Psychology, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. ✉email: nicola.marie.menne@hhu.de

out from the other lineup members; armed with this information, mock witnesses are simply asked to indicate who they think the suspect is [e.g.,[13,14]]. This most basic evaluation of lineup fairness can be used to investigate whether there are cues that make the suspect stand out from the other lineup members that are unrelated to the facial appearance of the culprit[15]. For example, in photo lineups the facial photograph of the suspect may stem from a different source (e.g., social media) than the photographs of the fillers that may be taken from special databases[16]. Sometimes, the photographs of the fillers may even be digitally manipulated[8,17]. Therefore, a simple inspection of the characteristics of the photographs such as brightness, contrast, color balance or softness could reveal who the suspect is (more on this below). Given that mock witnesses have not seen the face of the culprit, they cannot make an identification that is based on memory. Instead, they have to rely on inferences that are either informed by a description of the culprit or based on other clues available in the lineup. A lineup is fair if the mock-witness choices are evenly distributed among the lineup members (in a six-person lineup, each lineup member, including the suspect, should be selected by $1 \div 6$ of the mock witnesses). A lineup is unfair if disproportionately many mock witnesses select the suspect[9]. Based on the choices of mock witnesses, several formal measures of lineup fairness can be computed. These measures reflect either the effective lineup size (as opposed to the nominal lineup size) or the bias with which the suspect is selected[15,18]. Effective lineup-size measures indicate the number of lineup members that could plausibly be considered as the culprit. One of the most popular effective lineup-size measures is Tredoux's E[19]. The proportion of suspect selections[9] is a popular measure of the bias with which the suspect is selected. This measure reflects the extent to which the suspect stands out from the other lineup members.

The mock-witness task was originally developed to measure lineup fairness in real criminal cases in which the suspect's guilt is unknown to the police, not for measuring lineup fairness in laboratory experiments[20]. Nevertheless, it has become increasingly common in experimental research to rely on the mock-witness task[10] as it provides a seemingly straightforward solution to the problem of how to assess the fairness of lineups. However, the validity of the mock-witness task has been criticized on the grounds that there are substantial differences between the tasks of mock witnesses and eyewitnesses[10,15,21].

First, mock witnesses are typically encouraged or even forced to choose one of the lineup members while lineup rejections are discouraged or even prevented, respectively. If participants are discouraged from rejecting the lineup but ignore or defy these instructions, their data are excluded from analysis [e.g.,[22]]. However, in order to avoid this loss of data, mock witnesses are typically denied the option to reject the lineup. Instead, mock witnesses are usually forced to guess who the suspect is [e.g.,[9,14]]. In contrast, eyewitnesses are encouraged to reject the lineup if they are unsure as to whether or not the culprit is in the lineup. More specifically, eyewitnesses are typically given two-sided pre-lineup instructions that emphasize the fact that it is equally important to select the culprit in culprit-present lineups and to reject culprit-absent lineups [e.g.,[23–25]]. This is also the procedure recommended by several guidelines for how lineups should be conducted[26–28]. Two-sided instructions decrease the probability of selecting one of the lineup members based on guessing, which is highly desirable in eyewitness tasks because the reduction of guessing-based selections reduces false identifications of innocent suspects that could lead to wrongful convictions[29–32].

Second, the task of a mock witness differs necessarily from that of an eyewitness. Given that mock witnesses have not seen the face of the culprit, they cannot make a memory-based decision but have to perform a non-memory-based comparison of the faces in the lineup. Unlike eyewitnesses, mock witnesses are thus alerted to the fact that one person might stand out from the other lineup members. When using a description-based mock-witness task, participants are typically asked to choose the person who best fits the culprit's description which implies that the description fits one person better than the others [e.g.,[11]]. When no description is presented, mock witnesses are explicitly told to choose the person who looks most distinctive or stands out from the other lineup members [e.g.,[13]]. Both types of instructions can be expected to encourage non-memory-based comparisons among the lineup members which may make participants sensitive to unfairness cues, possibly to the degree to which participants notice cues they would not have noticed otherwise. When participants are not provided with a description of the culprit's face, it is impossible to search for the culprit in the lineup and the only remaining strategy is to carefully compare the photographs in the lineup to identify the face that stands out. This is markedly different from the memory-based identification task of eyewitnesses who have to match each lineup member to their memory representation of the culprit in order to decide whether or not one person represents the culprit[33]. Any features that are unrelated to the identity of the culprit such as brightness, contrast, color balance and softness of the photographs are irrelevant to this task and may be thus ignored by the eyewitnesses. Given these striking differences between the mock-witness task and the eyewitness task, the processes underlying the observed behavior may well differ between mock witnesses and eyewitnesses. It is thus unclear whether the mock-witness task can be used to draw valid conclusions about eyewitness identification decisions.

Fortunately, it is not necessary to rely on the mock-witness task to arrive at measures of lineup fairness. This is so because a valid measurement model is available for estimating biased suspect selection in unfair lineups directly from eyewitness data: the two-high threshold (2-HT) eyewitness identification model[32,34]. This model belongs to the class of multinomial processing tree (MPT) models, a family of models for estimating the probability of latent processes from categorical data[35,36]. For an overview of the MPT modeling approach, we recommend the very useful tutorial by Schmidt et al.[37]. Based on the full range of data categories observed in the eyewitness task (that is, suspect identifications, filler identifications and lineup rejections in both culprit-present and culprit-absent lineups), the model provides measures of the latent processes underlying eyewitness identification decisions. Specifically, the set of processes measured by the 2-HT eyewitness identification model comprises the detection of culprit presence and absence, the selection of a lineup member based on guessing and, most importantly in the present context, the process of biased suspect selection. The process of biased suspect selection will play a central role here because it reflects the process of selecting a suspect that stands out from the fillers in unfair lineups, as validation studies have shown[32,34].

A graphical illustration of the 2-HT eyewitness identification model is shown in Fig. 1. The model tree in the upper half of Fig. 1 illustrates the latent processes underlying eyewitness identification decisions from lineups in which the culprit is present. A culprit is detected with probability *dP* (for detection of the presence of the culprit). If participants do not detect the culprit, which occurs with probability 1 – *dP,* then two types of non-detection-based processes can still lead to the correct identification of the culprit in lineups with the culprit present. First, and most importantly for the present purposes, participants may select the suspect without relying on memory if the suspect stands out from the fillers. This process of biased suspect selection in unfair lineups is represented by parameter *b*. Second, in case of no biased selection of the suspect, which occurs with probability 1 – *b,* participants can still select one of the lineup members based on guessing with probability *g* (for guessing-based selection). In this case, participants will either pick out the suspect with a probability equal to 1 ÷ lineup size (approximately 0.16667 in the present case of six lineup members) or they will select one of the fillers with the complementary probability 1 – (1 ÷ lineup size). Guessing-based selection of one of the lineup members does not occur with probability 1 – *g*, in which case participants reject the lineup by not making an identification.

The model tree in the lower half of Fig. 1 refers to lineups from which the culprit is absent. Participants may correctly detect the absence of the culprit with probability *dA* (for detection of the absence of the culprit), resulting in a correct lineup rejection. If culprit-absence detection fails, which occurs with probability 1 – *dA*, the same non-detection-based biased and guessing-based selection processes occur as in culprit-present lineups: With probability *b,* the innocent suspect may stand out from the other lineup members and prompt participants to incorrectly select the innocent suspect. No biased selection occurs with probability 1 – *b*. In this case participants may still select a lineup member based on guessing with probability *g*. In culprit-absent lineups, this leads participants either to incorrectly pick out the innocent suspect (with probability 1 ÷ lineup size) or to select one of the fillers (with probability 1 – [1 ÷ lineup size]). Alternatively, participants may not select a lineup member based on guessing with probability 1 – *g,* which results in a correct rejection of the lineup in culprit-absent lineups.

The 2-HT eyewitness identification model has been extensively validated using novel experiments designed specifically for the purpose of testing the model's validity[32] and by fitting the model to published data obtained

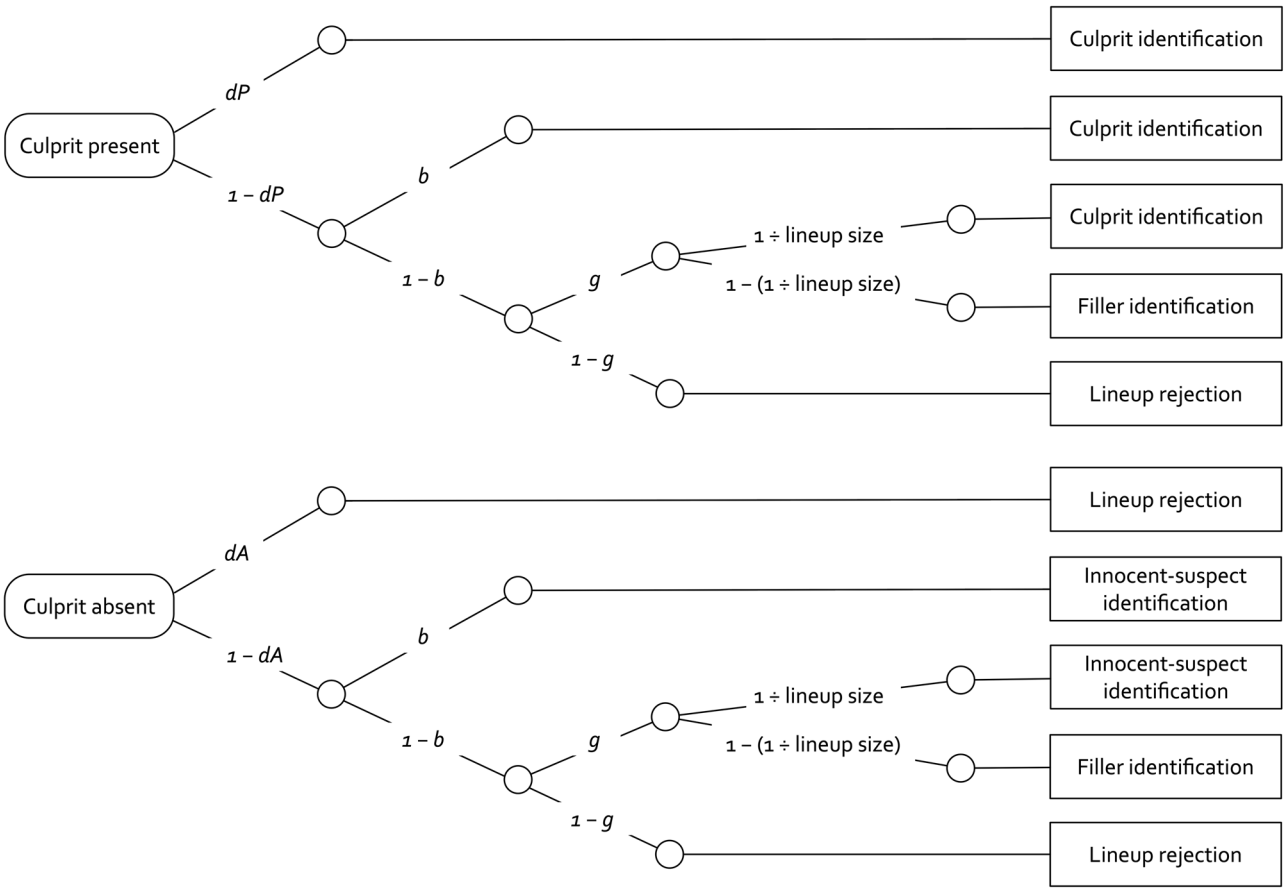

Figure 1. The 2-HT eyewitness identification model[32,34]. The rounded rectangles on the left represent the lineup types presented to the participants: culprit-present and culprit-absent lineups. The parameters attached to the branches of the trees denote transition probabilities of the latent cognitive processes postulated by the model (*dP:* probability of detecting the presence of the culprit; *b:* probability of biased selection of the suspect; *g:* probability of guessing-based selection among the lineup members; *dA:* probability of detecting the absence of the culprit). *Lineup size* represents the number of persons in the lineup. The rectangles on the right side show the categories of the observable responses.

in various laboratories[34]. Both approaches support the validity of the model by demonstrating that all parameters predictably reflect experimental manipulations of the processes they were designed to measure. A brief overview of the validation results for the biased-suspect-selection parameter *b* seems in order because this parameter is of central importance to the present study. Parameter *b* has been shown to sensitively reflect the unfairness of a lineup in which the suspect's face stood out from the fillers' faces because it was the only face without large birthmarks[32]. In addition, the biased-suspect-selection parameter *b* has been shown to be larger in unfair lineups with low suspect-filler similarity than in fair lineups with high suspect-filler similarity; parameter *b* has also been shown to be larger when the suspect's face stood out from the fillers due to distinctive facial features such as scars, bruisings, nose piercings and tattoos than when the suspect's face did not stand out[34].

In the experiments reported here, we measured the fairness of lineups containing either morphed or non-morphed photographs of fillers (hereinafter referred to as morphed and non-morphed lineups). This morphing manipulation is of applied relevance. Assembling lineups is often a challenging task because pertinent databases often do not provide enough facial photographs that match the description of the culprit[3,38]. To solve this problem, face-morphing software can be used to increase the selection of faces that can be used in the lineup[39,40]. What is more, the morphing process protects the identity of the fillers which is legally required, for instance, in Germany: Photographs must be digitally manipulated so that the persons originally depicted in the photographs are no longer recognizable before these photographs may legally be used as filler photographs in lineups[41]. The downside of this practice is that it often produces morphing artifacts such as shadows, double edges, ghosting effects or blurring and lets the image appear softer[42,43]. In morphed lineups, the photograph of the suspect might therefore stand out from the fillers because it is the only photograph in the lineup that has not been digitally manipulated. Witnesses could thus use the absence of morphing artifacts as the cue to the identity of the suspect which might lead to a biased selection of the suspect.

In the present series of experiments, we examined the effect of the morphing manipulation in the mock-witness and eyewitness tasks. Whether morphed lineups are unfair was tested in Experiment 1 using the traditional mock-witness task, thereby relying on two classical measures of lineup fairness based on mock-witness choices, Tredoux's *E* and the proportion of suspect selections. To anticipate, the results of the mock-witness task indicate that morphed simultaneous lineups are more unfair than non-morphed simultaneous lineups. In Experiments 2 to 4, we examined the effect of the morphing manipulation on eyewitness identification decisions using the 2-HT eyewitness identification model to measure biased suspect selection. In Experiment 2, we began by adding to the eyewitness task two features that are typical of the mock-witness task but highly unusual for the eyewitness task with the result that this version of the eyewitness task closely resembled the mock-witness task. These two features were then removed successively in Experiments 3 and 4 with the goal to identify the factors that may underlie the differences in the conclusions drawn based on data from the mock-witness task and the eyewitness task. Specifically, in Experiment 2, it was tested whether the biased-suspect-selection parameter *b* of the 2-HT eyewitness identification model reflects the unfairness of morphed lineups when participants (1) are discouraged from rejecting the lineups and (2) are alerted that a photograph might stand out from the other photographs in the lineup. When the eyewitness task thus closely resembled the mock-witness task, the eyewitness task led to the same conclusions as the mock-witness task: Biased suspect selection was enhanced in morphed simultaneous lineups in comparison to non-morphed simultaneous lineups. In the subsequent experiments, the procedure was brought closer to the standard procedure of typical eyewitness tasks. In Experiment 3, we removed the discouragement of lineup rejections. In Experiment 4, we removed both the discouragement of lineup rejections and the instruction to look for the photograph that stands out from the rest of the photographs. To anticipate, the results indicate that those who criticized the validity of the mock-witness task [e.g.,[21]] are correct: When the procedure was brought closer to the standard procedure of the eyewitness task, the effects of the morphing manipulation on biased suspect selection vanished. Specifically, the effect of the morphing manipulation on biased suspect selection was only descriptively present but not statistically significant in Experiment 3 and completely absent in Experiment 4. The results thus suggest that the mock-witness task has limited validity for drawing conclusions about eyewitness identification decisions. Instead, it is preferable to derive conclusions about lineup fairness directly from eyewitness identification decisions.

Experiment 1

In comparison to the eyewitness task, the mock-witness task provides an impoverished data structure because mock witnesses are hindered from rejecting the lineup and have actually not seen the culprit so that mock-witness lineups are essentially culprit-absent lineups. With only two of the six data categories of the eyewitness task left, it is not possible to use the 2-HT eyewitness identification model introduced above to analyze the data of the mock-witness task. Therefore, we relied on traditional mock-witness measures—Tredoux's *E* and the proportion of suspect selections—to measure the fairness of morphed and non-morphed simultaneous lineups in Experiment 1. However, in Experiment 2, the 2-HT eyewitness identification model was used to measure biased suspect selection in an eyewitness task that was modified to resemble the mock-witness task. To anticipate, the results obtained in the mock-witness task in Experiment 1 and the model-based analysis of eyewitness identification decisions in Experiment 2 converged in showing that morphed simultaneous lineups were significantly more unfair than non-morphed simultaneous lineups.

Method. All experiments reported here were conducted online. They were implemented using *SoSci Survey*[44] and were made available via https://www.soscisurvey.de. Participation was possible with a laptop or desktop computer, but not with a smartphone. All participants were recruited from the online research panel of Gapfish, Berlin, Germany (https://gapfish.com). Participants received a small monetary compensation for their participation.

Participants. Of the 851 participants who completed the socio-demographic questionnaire at the beginning of the experiment, 98 participants had to be excluded from the analysis because they did not complete the experiment or withdrew their consent to use their data ($n = 91$) or saw the lineups more than once due to repeated participation ($n = 7$). The final data set contained data from 753 participants (367 female, 384 male, 2 diverse) aged between 18 and 69 years ($M = 45$, $SD = 14$). The sample was characterized by a diversified level of education. We had aimed for a sample size of at least 750 valid datasets and ended data collection at the end of the day on which this criterion was met. Participants were randomly assigned to either the morphed lineup condition ($n = 385$) or the non-morphed lineup condition ($n = 368$).

Ethics statement. In each study, informed consent was obtained from all participants prior to participation. Ethical approval was received from the ethics committee of the Faculty of Mathematics and Natural Sciences at Heinrich Heine University Düsseldorf for a series of experiments of which the present experiments are a subset. All reported studies were carried out in accordance with the Declaration of Helsinki. In Experiments 2, 3 and 4, participants were warned that they would see a short video that included verbal and physical abuse. They were asked not to proceed if they felt uncomfortable expecting to watch such a video. At the end of the experiments, participants were informed that the crime they had witnessed had been staged.

Materials and procedure. Participants were told that a surveillance camera had recorded a crime scene in which four hooligans of a soccer club, FC Bayern München, attacked a soccer fan of a rivaling soccer club, Borussia Dortmund. Participants were informed that the police had constructed four lineups to test whether or not the suspects were the actual culprits. Participants received the instruction: "Each lineup consists of six pictures, one recent photo of a suspect and five photos from face databases" (all quotations in this article are translations of text that was originally presented in German). Participants were asked to indicate which lineup member was most likely to be the suspect in each lineup to help evaluate the fairness of the lineups. The instructions read: "We want to verify that the suspect's photograph does not stand out from the other lineup members. If the photograph stands out, then you can recognize the suspect even if you are a person who had not seen the recording. Therefore, please look at all photographs carefully. Please select the person that you think is the suspect by clicking on the 'Yes, is suspected' button that belongs to the particular face".

Participants subsequently saw four separate lineups, each consisting of one suspect and five morphed or non-morphed fillers (for an example, see Fig. 2). In total, eight male white students were used as suspects who also served as culprits or innocent suspects in Experiments 2 to 4. The set of eight suspects consisted of four pairs of

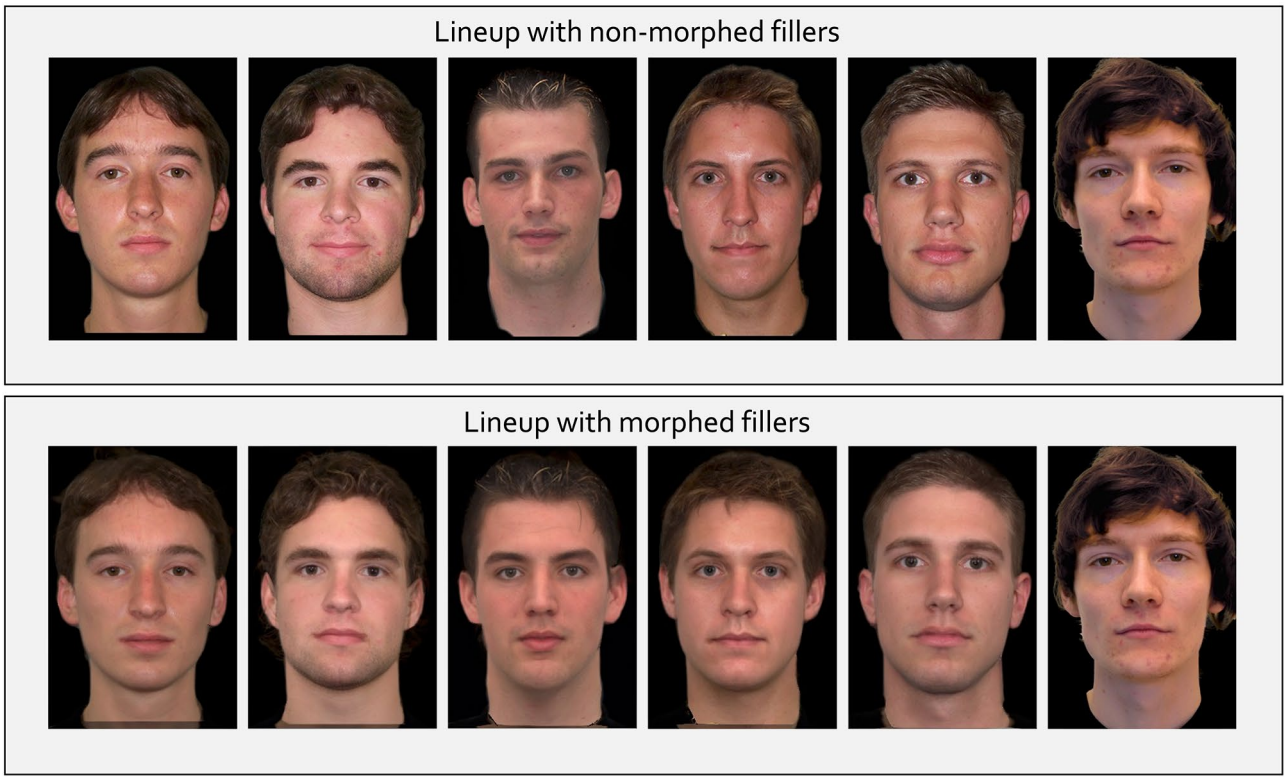

Figure 2. An illustration of a lineup with non-morphed and morphed fillers. The rightmost person represents the suspect but note that suspect and filler positions were always randomized in the experiments proper. We have written consent of the person representing the suspect to show the footage generated for the experiments. The photos of the fillers were taken from the Center for Vital Longevity Face Database[45] and the Radboud Faces Database[46], which are freely available for academic researchers.

suspects who resembled each other in terms of basic physical characteristics (e.g., hair color, hairstyle, stature). For each lineup, one suspect from each pair of suspects was randomly selected to be presented in the lineup. This is parallel to how the lineups were constructed in Experiments 2 to 4.

For the non-morphed lineup condition, five white male filler faces of persons aged between 18 and 29 years (hereinafter Set A) were chosen from the Center for Vital Longevity Face Database[45] for each pair of suspects. To create the fillers for the morphed lineup condition, five additional white male filler faces of similar age (hereinafter Set B) were selected for each suspect pair. These faces were obtained from three face databases: The Center for Vital Longevity Face Database [[45], https://agingmind.utdallas.edu/download-stimuli/face-database/], the FEI Face Database [[47], https://fei.edu.br/~cet/facedatabase.html] and the Radboud Faces Database [[46], http://www.rafd.nl]. All fillers were selected based on their similarity (as determined by the authors) to the corresponding suspects in terms of hair color, hairstyle and stature as well as their suitability for morphing (e.g., no glasses or piercings). Using *MorphAge* (Version 5.1, Creaceed, at https://creaceed.com/morphage), each filler from Set A was morphed with one filler from Set B by marking landmarks on one face (nose, eyes, eyebrows, mouth, ears, hairline and jaw-line) and matching each landmark to the corresponding point on the other face. Both faces of fillers from Set A and Set B were blended in a 50:50 ratio (i.e., a morph consisted of 50 % of each face). This procedure generated five morphed fillers for each suspect pair (for an example, see Fig. 3). All faces (i.e., those of the suspects and those of the fillers) were shown in frontal view against a black background with no clothes visible. All faces had a neutral facial expression. All photographs were edited to equate brightness, lighting and the position of the face among the photographs of the fillers and those of the suspects. The photographs were displayed at a resolution of 142×214 pixels.

The four lineups were presented one after another in a simultaneous format. In each lineup, all six faces were shown together in a single row with the option to respond "Yes, is suspected" appearing underneath each photograph. The position of the suspect and the five fillers was randomized. Implementing the typical mock-witness task[9], participants were not given the option to reject the lineup. Once the participants had selected a person, they could proceed to the next lineup by pressing the "Next" button. The order in which the lineups appeared was randomly determined for each participant. After completing the four lineup trials, participants were debriefed and thanked for their participation. The experiment took about 10 min.

Results. For each lineup, the distribution of mock-witness choices across the six lineup members was determined. Based on these mock-witness data, lineup fairness was computed in two ways. First, effective lineup size was assessed using Tredoux's *E*, which provides an estimate of the number of plausible lineup members[19]. Tredoux's *E* takes on a minimum value of 1 and a maximum value of *k,* the number of lineup members (in our lineups, six). Each lineup member who receives fewer choices than expected by chance will cause a reduction of the value of Tredoux's *E,* starting from *k* and approaching 1. Tredoux's *E* was calculated separately for each of the four morphed and non-morphed lineups before an average effective size was computed separately for the morphed and the non-morphed lineup condition that is reported below (details on the data underlying these effective sizes are reported in the Open Science Framework repository at https://osf.io/zaybc/). Second, the average proportion of suspect selections was calculated for both morphed and non-morphed lineups as a measure of biased selection of the suspect[9]. This measure is straightforward to interpret: If the mock-witness choices are equally distributed across the lineup members (i.e., one-sixth of the choices fall on the suspect), a lineup would be considered perfectly fair. If a disproportionate number of mock witnesses pick out the suspect, a lineup is considered unfair. Thus, a greater proportion of participants choosing the suspect from morphed lineups than

Figure 3. On the left side, an original filler face from Set A is shown. On the right side, an original filler face from Set B is shown. In the center, the face morph is shown (morph rate = 50:50). The photograph on the left was taken from the Center for Vital Longevity Face Database[45]. The photograph on the right was taken from the Radboud Faces Database[46].

from non-morphed lineups would indicate that the morphed lineups are more biased toward the suspect than the non-morphed lineups.

The average Tredoux's E was higher for the non-morphed lineup condition ($M = 4.51$) than for the morphed lineup condition ($M = 3.44$), indicating that the morphed lineups were more unfair than the non-morphed lineups. The same conclusion can be reached when calculating the proportion of suspect selections in both conditions. The average proportion of suspect selections was significantly higher in the morphed lineup condition ($M = 47.5$ %) than in the non-morphed lineup condition ($M = 25$ %), as determined by a z-test for proportions ($z = 12.80$, $p < 0.001$).

Discussion. The results obtained in the traditional mock-witness task indicate that the morphed lineups were more unfair than the non-morphed lineups. These results thus lead to the conclusion that the police should stop using this morphing technique as it leads to artifacts that make the suspect stand out from the other lineup members. However, it has yet to be shown whether or not these findings are limited to the mock-witness task. Therefore, the purpose of Experiments 2 to 4 was to examine the effects of the same morphing manipulation on eyewitness identification decisions in simultaneous and sequential lineups.

Experiment 2

It cannot be taken for granted that the mock-witness choices validly reflect the processes that determine eyewitness identification decisions. Therefore, it has to be tested whether the morphing manipulation affects eyewitness identification decisions to the same extent as it affects mock-witness choices. As noted above, the mock-witness task differs from a typical eyewitness task in at least two significant ways. Unlike eyewitnesses, mock witnesses (1) are required to choose one of the lineup members and (2) are alerted to the fact that one lineup member might stand out from the others. Therefore, the aim of the following series of experiments was to test, across experiments, whether evidence for the unfairness of the morphed lineups emerged depending on whether these two features were present in the eyewitness task.

As the next step, we aimed at testing whether the unfairness effects of the morphing manipulation could be demonstrated in eyewitness identification decisions when the eyewitness task was modified to mimic the mock-witness task—that is, when participants (1) were discouraged from rejecting the lineup and (2) were alerted to the fact that the suspect may stand out from the fillers. The eyewitness task provides a richer data structure than the mock-witness task because suspect identifications, filler identifications and lineup rejections in both culprit-present and culprit-absent lineups can be observed. It is thus important to rely on a measurement model that takes the full data structure of the eyewitness task into account. The 2-HT eyewitness identification model capitalizes on the full range of data categories that can be observed in the eyewitness task. It has been successfully demonstrated that the model's parameter b sensitively reflects the biased selection of suspects[32,34] and can thus be used to assess the unfairness of lineups. If the biased-suspect-selection parameter b is sensitive to the morphing manipulation used in Experiment 1, the estimate of parameter b should be higher for simultaneous morphed lineups than for simultaneous non-morphed lineups.

An additional aspect not mentioned so far is that the mock-witness technique has been proposed to evaluate the fairness of simultaneous lineups but it is of limited use in estimating the fairness of sequential lineups[21,48]. However, in some countries such as the UK and Germany, the sequential presentation has become the standard way of conducting police lineups[49,50]. The second aim of Experiment 2 was thus to test the effect of the morphing manipulation on biased suspect selection in sequential lineups. Here it is useful that the 2-HT model has been demonstrated to validly reflect biased selection in both simultaneous and sequential lineups[32]. Previous research has demonstrated that sequential lineups provide some protection against biased suspect selection in unfair lineups[3,51,52]. For example, in simultaneous lineups, a photograph that differs from the other photographs in brightness, contrast, color balance or softness may pop out from the others. In sequential lineups, witnesses cannot compare the photographs side-by-side. Therefore, it may be less salient that one photograph stands out from the others in sequential lineups. There is thus reason to expect that the morphing manipulation is less likely to affect eyewitness identification decisions in sequential lineups than in simultaneous lineups. As a consequence, biased selection of the suspect may only be enhanced in simultaneous morphed lineups in comparison to simultaneous non-morphed lineups but may not differ between morphed and non-morphed sequential lineups.

Method. *Participants.* Of the 934 participants who completed the socio-demographic questionnaire at the beginning of the experiment, 151 participants had to be excluded from the analysis because they did not complete the experiment or withdrew their consent to use their data ($n = 120$), incorrectly answered the attention-check question ($n = 11$) or watched the video more than once due to repeated participation ($n = 20$). The final data set contained data from 783 participants (336 female, 445 male, 2 diverse) aged between 18 and 69 years ($M = 45$, $SD = 14$). The sample was characterized by a diversified level of education. We had aimed for a sample size of at least 750 valid datasets and stopped data collection at the end of the day on which this criterion was met. A sensitivity analysis using G*Power[53] showed that with a sample size of $N = 783$, four eyewitness identification decisions and an alpha error probability of 0.05, it was possible to detect even small effects of the morphing manipulation on the biased-suspect-selection parameter b of effect size $w = 0.06$ with a statistical power (1 – beta error probability) of 0.95. Participants were randomly assigned to one of the four lineup conditions: the morphed simultaneous lineup condition ($n = 199$), the non-morphed simultaneous lineup condition ($n = 190$), the morphed sequential lineup condition ($n = 196$) and the non-morphed sequential lineup condition ($n = 198$).

Materials and procedure. Staged-crime videos. The same two parallel videos were used (henceforth referred to as Video A and Video B) as in the validation experiments of Winter et al.[32]. Both videos showed the same

staged crime, but with different sets of actors: Four hooligans of the German soccer club FC Bayern München (henceforth referred to as the culprits) verbally and physically attacked a fan of the rival soccer club Borussia Dortmund (henceforth referred to as the victim) at a bus stop. The hooligans and their victim wore typical fan clothing of the soccer clubs (caps, shirts and scarfs in typical club colors). The four culprits poked fun at the victim, insulted him and tossed his personal belongings around. At the end of the video, the culprits pushed the victim to the ground. The four culprits continued to verbally and physically abuse the victim. Once the culprits noticed another pedestrian approaching (not visible in the videos), they ran away shouting loudly. Note that in many other lineup studies, participants are only exposed to a single culprit at encoding and thus to a single lineup at test [e.g.,[8,54–56]]. However, to increase the efficiency of data collection, we followed the lead of other researchers [e.g.,[57–60]] and presented our participants with a video showing four culprits. This procedure allowed us to generate four data points per participant instead of just one. Also note that multiple-culprit crimes are in fact quite frequent[61,62]. For instance, in 2008, 25 % of all crimes committed in the UK involved four or more culprits [[63], p. 287].

The two parallel videos had the same content (i.e., the videos contained the same verbal and physical abuse in the same sequence and with the same timing), but the culprits and the victims differed between the two videos. Care was taken to select the actors in such a way that the victim of Video A resembled the victim of Video B as closely as possible and that each of the four culprits of Video A resembled one of the four culprits of Video B in hair color, hairstyle and stature (i.e., Culprit 1 in Video A matched Culprit 1 in Video B, Culprit 2 in Video A matched Culprit 2 in Video B and so on). Note that the same eight faces had served as suspects in Experiment 1 to ensure comparability between the experiments. It was randomly determined whether participants watched Video A or Video B. Both videos were 130 s long and showed a clear view of the culprits' faces. The videos were presented in a resolution of 885×500 pixels.

Participants could start the video by clicking on the "Start" button. They were unable to proceed to the next page until they had watched the whole video. After the video had finished, participants had to answer a 10-alternatives attention-check question ("What kind of people were shown in the video?"; correct response: "Soccer fans") to test whether participants had been paying attention to the video. The order of the response options was randomized.

Lineup procedures. Participants who had correctly answered the attention-check question were asked to identify the four culprits: "The video you just watched showed aggressive FC Bayern München hooligans. You will be asked to identify these hooligans. For this purpose, we are going to show you some lineups". As in Experiment 1, participants were informed that "Each lineup consists of six pictures, one recent photo of the suspect and five photos taken from face databases". They were also made aware of the possibility that the suspect might stand out from the other lineup members: "It is possible that the suspect stands out from the other lineup members. If the suspect stands out, then you can recognize the suspect even if you have not seen the video. Therefore, please look at all photos carefully". Lineup rejections were discouraged by instructing the participants: "It is very likely that every lineup contains a culprit. Therefore, even if you are uncertain about whether or not the culprit is in the lineup, choose the picture that stands out among the others. Then you will almost certainly identify the culprit. To do this, click on the 'Yes, was present' button that belongs to that face. Only if you are very certain that the persons do [simultaneous lineups]/person does [sequential lineups] not represent any of the culprits, click on the 'No, none of these persons was present' [simultaneous lineups]/'No, this person was not present' [sequential lineups] button". Participants were not made aware of the number of lineups that were about to follow.

Participants saw a total of four separate lineups, two were randomly selected to be culprit-present lineups and the other two were culprit-absent lineups. The lineups consisted of the same faces as in Experiment 1. Each lineup consisted of the facial photographs of six persons, one suspect face and either five morphed or five non-morphed filler faces (see Fig. 2). The crossed lineup procedure introduced by Winter et al.[32] was used to manipulate the suspect's guilt. Specifically, if participants had seen Video A, two culprits of Video A (e.g., Culprits 1 and 3) served as the culprits in the culprit-present lineups, while two culprits of Video B (e.g., Culprits 2 and 4) served as the innocent suspects in the culprit-absent lineups. If participants had seen Video B, two culprits of Video B served as the culprits in the culprit-present lineups while two culprits of Video A served as the innocent suspects in the culprit-absent lineups. This approach had two advantages: First, culprit-absent lineups contained a designated innocent suspect to whom the fillers had been matched. This situation represents a more ecologically valid lineup procedure than using only fillers in culprit-absent lineups. This is so because, in practice, the photographs of the suspects (whose guilt or innocence is unknown) are taken from other sources (e.g., social media) than the photographs of the fillers which are usually taken from face databases and may be digitally altered. Second, culprit-present and culprit-absent lineups included the identical filler faces; only the identity of the suspect was changed. Which of the two suspects served as the culprit or innocent suspect depended on the random assignment to one of the two videos (see above). In that way, it was ensured that, on average, the degree of fairness was the same in culprit-present and culprit-absent lineups. A similar approach, the single-lineup procedure, has been proposed by Oriet and Fitzgerald[64]. In contrast to the crossed lineup procedure used here, the single lineup procedure implies showing all participants the same lineup after having seen one of two videos, one that contains the suspect in the lineup while the other contains a person who is not presented in the lineup but matches the physical description of the suspect. As in Experiment 1, all photographs were presented at a resolution of 142×214 pixels.

Participants were randomly assigned to either the simultaneous or the sequential lineup conditions. In the simultaneous lineup conditions, the six faces were shown together in one row. Participants made a decision by either clicking on the "Yes, was present" button underneath a face to identify a person as a culprit or by clicking on the "No, none of these persons was present" button located to the right of each lineup to reject the lineup.

After having made a decision, participants were asked to express how confident they were in their judgments in order to approximate the procedure to that of a real police lineup. Then they could initiate the presentation of the next lineup by clicking on the "Next" button. In the sequential lineup conditions, the faces were presented one at a time. For each of the six faces, participants decided whether or not the depicted face belonged to one of the culprits by clicking on either the "Yes, was present" button underneath the face or the "No, this person was not present" button located to the right of the face. A decision was required before participants could proceed to the next lineup member. If participants identified more than one face within a single lineup, only the last identification decision was counted. This procedure is legally prescribed in several jurisdictions such as Germany or the United States[24,50,65]. It also corresponds to the identification procedure in the simultaneous lineups in which it was possible for participants to revise their decision before clicking the "Next" button. After each decision, participants were asked to indicate their level of confidence in their judgment in the same manner as in the simultaneous lineup conditions. A lineup was counted as rejected if participants identified none of the lineup members. The order with which the lineups were presented was randomized, as was the position of the lineup members in each lineup. After their response to the fourth lineup, participants were debriefed and thanked. The experiment took about 10 min.

Results. Four instances of the model illustrated in Fig. 1 were used for the model-based analysis, one for the simultaneous morphed lineups, one for the simultaneous non-morphed lineups, one for the sequential morphed lineups and one for the sequential non-morphed lineups. Goodness-of-fit tests and parameter estimates were calculated using *multiTree*[66]. The alpha error probability was set to 0.05. The observed response frequencies and proportions for Experiments 2, 3 and 4 are reported in Table 1. Our goal was to start with a base model that was as simple as possible. Therefore, we imposed restrictions onto the 2-HT eyewitness identification model that seemed justified, if not required on a priori grounds, to achieve this goal. First, there was no a priori reason why the ability to detect the absence of the culprit should differ as a function of the conditions in the present experiment [see[32,34] for conditions that can be expected to affect the probability of culprit-absence detection]. Therefore, the parameter representing culprit-absence detection (dA) was set to be equal across the four lineup conditions. Second, there was no a priori reason why the ability to detect the presence of the culprit should differ between the morphed and non-morphed lineups. However, previous results that were obtained with the same stimulus materials and procedure[32] suggest that culprit-presence detection is somewhat better in simultaneous than in sequential lineups. Therefore, the culprit-presence-detection parameter dP was set to be equal between the simultaneous morphed and non-morphed conditions and between the sequential morphed and non-morphed conditions. Third, there was no a priori reason why guessing-based selection should differ between the morphed and non-morphed lineups. However, previous results[32] suggest that guessing-based selection is enhanced in sequential in comparison to simultaneous lineups. Therefore, the guessing-based-selection parameter g was set to be equal between the simultaneous morphed and non-morphed conditions and between the sequential morphed and non-morphed conditions. The asymptotically chi-square distributed likelihood-ratio goodness-of-fit statistic (with degrees of freedom reported in parentheses) [for details, see[67]] indicated that the base model incorporating these restrictions fit the data, $G^2(7) = 2.55$, $p = 0.924$.

In an MPT model such as the 2-HT eyewitness identification model, hypotheses can be tested directly at the level of the parameters representing the postulated processes. For instance, the hypothesis that biased suspect selection is higher in the morphed lineup conditions than in the non-morphed lineup conditions can be tested

Lineup format	Type of lineup fillers	Culprit-present lineups			Culprit-absent lineups		
		Culprit identifications	Filler identifications	Lineup rejections	Innocent- suspect identifications	Filler identifications	Lineup rejections
Experiment 2							
Simultaneous	Morphed	177 (0.44)	120 (0.30)	101 (0.25)	93 (0.23)	161 (0.40)	144 (0.36)
	Non-morphed	155 (0.41)	117 (0.31)	108 (0.28)	66 (0.17)	160 (0.42)	154 (0.41)
Sequential	Morphed	112 (0.29)	220 (0.56)	60 (0.15)	58 (0.15)	247 (0.63)	87 (0.22)
	Non-morphed	130 (0.33)	201 (0.51)	65 (0.16)	67 (0.17)	234 (0.59)	95 (0.24)
Experiment 3							
Simultaneous	Morphed	154 (0.39)	117 (0.30)	123 (0.31)	70 (0.18)	131 (0.33)	193 (0.49)
	Non-morphed	132 (0.37)	95 (0.26)	133 (0.37)	47 (0.13)	135 (0.38)	178 (0.49)
Sequential	Morphed	116 (0.30)	208 (0.53)	66 (0.17)	53 (0.14)	235 (0.60)	102 (0.26)
	Non-morphed	126 (0.34)	184 (0.49)	66 (0.18)	58 (0.15)	220 (0.59)	98 (0.26)
Experiment 4							
Simultaneous	Morphed	135 (0.37)	86 (0.23)	145 (0.40)	49 (0.13)	119 (0.33)	198 (0.54)
	Non-morphed	144 (0.36)	119 (0.30)	137 (0.34)	55 (0.14)	139 (0.35)	206 (0.52)
Sequential	Morphed	125 (0.32)	189 (0.49)	74 (0.19)	69 (0.18)	209 (0.54)	110 (0.28)
	Non-morphed	126 (0.34)	163 (0.44)	79 (0.21)	62 (0.17)	199 (0.54)	107 (0.29)

Table 1. Observed response frequencies and proportions (in parentheses) as a function of lineup format and the type of lineup fillers observed in Experiments 2, 3 and 4. The proportions are rounded to two decimal places and therefore do not always add up exactly to 1.

by restricting parameter b to be equal between these conditions. If the model with this restriction fits significantly worse to the data than the base model (measured by the ΔG^2 difference statistic with degrees of freedom corresponding to the difference between degrees of freedom of the model with the additional restriction and the degrees of freedom of the base model), we would have to reject the equality assumption implied by the restriction and would conclude that parameter b differs between conditions.

Figure 4 shows the parameter estimates of the biased-suspect-selection parameter b for morphed and non-morphed lineups as a function of lineup format. The probability of biased suspect selection was higher for morphed lineups than for non-morphed lineups when simultaneous lineups were used, $\Delta G^2(1) = 5.31$, $p = 0.021$, $w = 0.04$, in accordance with the mock-witness results of Experiment 1. However, biased suspect selection did not differ between morphed and non-morphed lineups when sequential lineups were used, $\Delta G^2(1) = 2.04$, $p = 0.153$, $w = 0.03$. In addition, the probability of biased suspect selection was significantly higher for morphed simultaneous than for morphed sequential lineups, $\Delta G^2(1) = 21.89$, $p < 0.001$, $w = 0.08$, but it did not differ between non-morphed simultaneous and sequential lineups, $\Delta G^2(1) = 1.60$, $p = 0.207$, $w = 0.02$.

The parameter estimates for culprit-presence detection (dP), guessing-based selection (g) and culprit-absence detection (dA) as a function of lineup format are shown in Table 2. Given that our hypotheses pertained only to biased suspect selection (b), we do not include an analysis or discussion of the other parameters here. However, we will provide a brief overview and interpretation of the results of Experiments 2 to 4 pertaining to culprit-presence detection and guessing-based selection in the General Discussion.

Discussion. In Experiment 2, it was tested whether the morphing manipulation that affected mock-witness-based measures of unfairness in Experiment 1 would affect the biased selection of the suspect that was directly estimated from the identification decisions of eyewitnesses in simultaneous and sequential lineups if the eyewitness task closely resembled the mock-witness task. Eyewitnesses were discouraged from rejecting the lineup

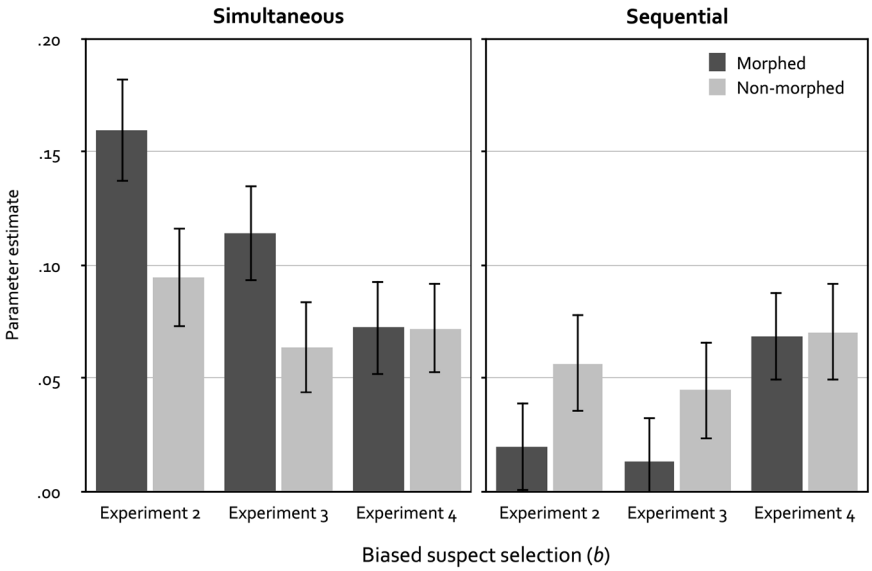

Figure 4. Parameter estimates of parameter b reflecting the probability of biased suspect selection as a function of lineup format (simultaneous vs. sequential) and the type of lineup fillers (morphed vs. non-morphed) in Experiments 2, 3 and 4. The error bars represent the standard errors.

Lineup format	Experiment 2			Experiment 3			Experiment 4		
	dP	*g*	*dA*	*dP*	*g*	*dA*	*dP*	*g*	*dA*
Simultaneous	0.27 (0.03)	0.59 (0.02)	0.04 (0.02)	0.26 (0.02)	0.50 (0.02)	0.06 (0.03)	0.26 (0.02)	0.46 (0.02)	0.06 (0.03)
Sequential	0.17 (0.02)	0.80 (0.02)		0.19 (0.02)	0.78 (0.02)		0.18 (0.03)	0.74 (0.02)	

Table 2. Parameter estimates for culprit-presence detection (dP), guessing-based selection (g) and culprit-absence detection (dA) as a function of lineup format (Experiments 2, 3 and 4). Within the base model, parameters dP and g were each set to be equal between morphed and non-morphed lineups, separately for simultaneous and sequential lineups. Parameter dA was set to be equal among the four lineup conditions. Values in parentheses represent the standard errors. See text for details.

and were made aware of the fact that the suspect may stand out from the fillers. Under these conditions, the model-based results converged with those based on the mock-witness measures obtained in Experiment 1. In simultaneous lineups, the morphing manipulation significantly increased the biased selection of the suspect, represented by parameter *b* of the 2-HT eyewitness identification model.

By using the 2-HT eyewitness identification model it was also possible to measure the effect of the morphing manipulation on biased suspect selection in sequential lineups. The probability of biased suspect selection did not differ significantly between morphed and non-morphed lineups when they were presented sequentially. In sequential lineups, it is not possible to compare the photographs side by side; participants can only look at each lineup member individually. Without the direct comparison of all of the photographs in the lineup, it may have been difficult or even impossible for the participants to identify the fact that it is the absence of morphing artifacts that distinguishes the suspect from the fillers. In consequence, biased suspect selection was less prevalent in sequential lineups than in simultaneous lineups. This is in line with previous research indicating that sequential lineups provide some protection against biased suspect selection in unfair lineups[3,51,52], presumably because participants cannot compare lineup faces side-by-side and thus are unable to detect the cues that distinguish the suspect from the fillers.

In Experiment 2, eyewitnesses were used instead of mock witnesses, but we discouraged participants from rejecting the lineup and alerted them to the fact that one lineup member might stand out from the others. This is typical for instructions that are used in the mock-witness task but deviates from the pre-lineup instructions that are recommended for the eyewitness task[26–28]. Experiments 3 and 4 serve to test whether the same or different results are obtained when the procedure is brought closer to the standard procedure of typical eyewitness tasks.

Experiment 3

Experiment 3 was identical to Experiment 2 with the exception that the pre-lineup instructions did not discourage participants from rejecting the lineup. Previous research has consistently demonstrated that one-sided instructions that emphasize only the need to identify the culprit but ignore the need to reject culprit-absent lineups enhance both correct culprit identifications and false innocent-suspect identifications[29,68]. The requirement to identify one of the lineup members as the culprit may increase the likelihood that participants make the non-memory-based identification decision to identify the person who stands out from the other persons in the lineup. It is unclear whether participants rely on the morphing artifacts when they receive two-sided instructions that emphasize both the need to identify the culprit and the need to reject culprit-absent lineups. It seemed thus interesting to test whether biased suspect selection is enhanced in simultaneous morphed lineups in comparison to simultaneous non-morphed lineups in Experiment 3 in which participants were alerted to the fact that the face of the suspect might stand out from the other faces but participants were not discouraged from rejecting the lineup.

Method. *Participants.* Of the 894 participants who completed the socio-demographic questionnaire at the beginning of the experiment, 134 had to be excluded from the analysis because they did not complete the experiment or withdrew their consent to use their data ($n = 112$), incorrectly answered the attention-check question ($n = 13$) or watched the video more than once due to repeated participation ($n = 9$). The final data set contained data from 760 participants (324 female, 434 male, 2 diverse) aged between 18 and 85 years ($M = 45$, $SD = 15$). The sample was characterized by a diversified level of education. We had aimed for a sample size of at least 750 valid datasets and stopped data collection at the end of the day on which this criterion was met. A sensitivity analysis showed that with a sample size of $N = 760$ participants, four eyewitness identification decisions and an alpha error probability of 0.05, it was possible to detect even small effects of the morphing manipulation on the biased-suspect-selection parameter *b* of effect size $w = 0.07$ with a statistical power (1 – beta error probability) of 0.95[53]. As in Experiment 2, participants were randomly assigned to one of the four lineup conditions: the morphed simultaneous lineup condition ($n = 197$), the non-morphed simultaneous lineup condition ($n = 180$), the morphed sequential lineup condition ($n = 195$) and the non-morphed sequential lineup condition ($n = 188$).

Materials and procedure. Materials and procedure were identical to those of Experiment 2 with the following exception. Instead of discouraging lineup rejections by implying that the culprit would be present in the lineup, the instructions emphasized the possibility that the culprit might not be present. As in Experiment 2, however, participants were alerted that a photograph might stand out from the other photographs in the lineup. The instructions read: "The video you just watched showed aggressive FC Bayern München hooligans. You will be asked to identify these hooligans. For this purpose, we are going to show you some lineups. Each lineup consists of six pictures, one recent photo of the suspect and five photos taken from face databases. It is possible that the suspect stands out from the other lineup members. If the suspect stands out, then you can recognize the suspect even if you have not seen the video. Therefore, please look at all photos carefully. You will be asked to indicate if one of the lineup members is one of the FC Bayern München hooligans shown in the video. It is also possible that none of the hooligans is in the lineup. If you recognize a face, then click on the 'Yes, was present' button that belongs to that face. Otherwise click on the 'No, none of these persons was present' [simultaneous lineups]/'No, this person was not present' [sequential lineups] button".

Results. The same assumptions as in Experiment 2 were used to arrive at the base model. This base model fit the data, $G^2(7) = 3.58$, $p = 0.827$.

Figure 4 shows the parameter estimates of the biased-suspect-selection parameter *b* for morphed and non-morphed lineups as a function of lineup format. Parallel to Experiment 2, the probability of biased suspect selection was descriptively higher for morphed simultaneous lineups compared to non-morphed simultaneous

lineups, but in contrast to Experiment 2 this difference was no longer statistically significant, $\Delta G^2(1) = 3.63$, $p = 0.057$, $w = 0.03$. As in Experiment 2, biased suspect selection did not differ between morphed and non-morphed lineups when sequential lineups were used, $\Delta G^2(1) = 1.49$, $p = 0.222$, $w = 0.02$. Also as in Experiment 2, the probability of biased suspect selection was significantly higher for morphed simultaneous than for morphed sequential lineups, $\Delta G^2(1) = 12.53$, $p < 0.001$, $w = 0.06$, but it did not differ between non-morphed simultaneous and sequential lineups, $\Delta G^2(1) = 0.42$, $p = 0.515$, $w = 0.01$.

The parameter estimates for culprit-presence detection (dP), guessing-based selection (g) and culprit-absence detection (dA) as a function of lineup format are shown in Table 2.

Discussion. In Experiment 3, the probability of biased suspect selection no longer differed significantly between morphed and non-morphed simultaneous lineups. Emphasizing that the culprit might or might not be in the lineup reduced the probability of biased suspect selection in simultaneous morphed lineups compared to Experiment 2 in which lineup rejections were discouraged. This was expected given the plausible assumption that instructions discouraging lineup rejections cause eyewitnesses to search harder than they usually do for cues that make the suspect stand out.

However, at a descriptive level the probability of biased suspect selection was still larger in morphed simultaneous lineups compared to non-morphed simultaneous lineups. Moreover, when the fillers were morphed, biased suspect selection was still less prevalent in sequential lineups than in simultaneous lineups, which is consistent with the results of Experiment 2 as well as previous research[3,51,52]. Thus, there was still some evidence of an effect of the morphing manipulation on the data obtained in Experiment 3.

Experiment 4

Experiment 4 was identical to Experiment 3 with the exception that participants were not alerted to the fact that the suspect's photograph might stand out from the other photographs in the lineup. Instead, participants were presented with instructions that are given in a typical eyewitness task in which participants are not discouraged from rejecting the lineup and are not alerted that a photograph might stand out from the other photographs in the lineup[26–28]. The main question was whether the effect of the morphing manipulation on biased suspect selection in simultaneous lineups would be abolished under these conditions that, within the present series of experiments, most closely mirror real police lineup procedures.

Method. *Participants.* Of the 958 participants who completed the socio-demographic questionnaire at the beginning of the experiment, 197 participants had to be excluded from the analysis because they did not complete the experiment or withdrew their consent to use their data ($n = 155$), incorrectly answered the attention-check question ($n = 13$), watched the video more than once due to repeated participation ($n = 22$) or a technical error occurred during the experiment ($n = 7$). The final data set contained data from 761 participants (335 female, 426 male) aged between 18 to 80 years ($M = 48$, $SD = 17$). The sample was characterized by a diversified level of education. We had aimed for a sample size of at least 750 participants and stopped data collection at the end of the day on which this criterion was met. A sensitivity analysis showed that with a sample size of $N = 761$ participants, four eyewitness identification decisions and an alpha error probability of 0.05, it was possible to detect even small effects of the morphing manipulation on the biased-suspect-selection parameter b of effect size $w = 0.07$ with a statistical power (1 – beta error probability) of 0.95[53]. Participants were randomly assigned to one of the four lineup conditions: the morphed simultaneous lineup condition ($n = 183$), the non-morphed simultaneous lineup condition ($n = 200$), the morphed sequential lineup condition ($n = 194$) and the non-morphed sequential lineup condition ($n = 184$).

Materials and procedure. Materials and procedure were identical to those of Experiment 3 with the exception that the instructions no longer alerted participants that a photograph might stand out from the other photographs in the lineup and thus corresponded to those used in typical eyewitness identification situations. The instructions read: "The video you just watched showed aggressive FC Bayern München hooligans. You will be asked to identify these hooligans. For this purpose, we are going to show you some lineups. In each lineup, you will see some faces. You will be asked to indicate if one of the lineup members is one of the FC Bayern München hooligans shown in the video. It is also possible that none of the hooligans is in the lineup. If you recognize a face, then click on the 'Yes, was present' button that belongs to that face. Otherwise, click on the 'No, none of these persons was present' [simultaneous lineups]/'No, this person was not present' [sequential lineups] button".

Results. The same assumptions as in Experiments 2 and 3 were used to arrive at the base model. This base model fit the data, $G^2(7) = 6.61$, $p = 0.471$.

Figure 4 shows the parameter estimates of the biased-suspect-selection parameter b for morphed and non-morphed lineups as a function of lineup format. Crucially, the descriptive difference between morphed and non-morphed simultaneous lineups that was still evident in Experiment 3 was absent in Experiment 4. This result is so clear from the sizes of the parameter estimates that it does not require a statistical test, but for completeness, we report here that the probability of biased suspect selection did not differ significantly between morphed simultaneous lineups and non-morphed simultaneous lineups, $\Delta G^2(1) < 0.01$, $p = 0.992$, $w < 0.01$. Further, as in Experiments 2 and 3, biased suspect selection did not differ between morphed and non-morphed lineups when sequential lineups were used, $\Delta G^2(1) < 0.01$, $p = 0.950$, $w < 0.01$. Finally, the probability of biased suspect selection differed neither between morphed simultaneous and sequential lineups, $\Delta G^2(1) = 0.02$, $p = 0.898$, $w < 0.01$, nor between non-morphed simultaneous and sequential lineups, $\Delta G^2(1) < 0.01$, $p = 0.952$, $w < 0.01$.

The parameter estimates for culprit-presence detection (*dP*), guessing-based selection (*g*) and culprit-absence detection (*dA*) as a function of lineup format are shown in Table 2.

Discussion. In Experiment 4, the effect of the morphing manipulation on biased suspect selection was completely absent. When lineup rejections were not discouraged and participants were not alerted that a photograph might stand out from the other photographs in the lineup, the probability of biased suspect selection did not differ between morphed and non-morphed lineups in both simultaneous and sequential formats. In contrast to Experiments 2 and 3, the probability of biased suspect selection was comparable between simultaneous and sequential lineups even in the morphed lineup condition. Thus, when the task characteristics closely mirrored the conditions of a real lineup procedure, there was absolutely no evidence of an effect of the morphing manipulation on biased suspect selection in any of the lineups.

General discussion

The well-validated 2-HT eyewitness identification model[32,34] allows measuring lineup fairness directly from eyewitness identification decisions without relying on the choices of mock witnesses. The problem with using mock witnesses is that their task differs substantially from the task of eyewitnesses. As a consequence, there are doubts as to whether lineup fairness measured in the mock-witness task can predict lineup fairness in a typical eyewitness task[10,15,20,21,48]. The present series of experiments demonstrates that these doubts are justified.

We measured the fairness of lineups containing either morphed or non-morphed fillers. This morphing manipulation is of applied relevance considering that the morphing technique can serve as a method both to create fillers when the pertinent databases do not contain enough photographs that are similar enough to descriptions of the suspect[39,40] and to morph photographs to protect the identities of the persons depicted in the filler photographs. The latter is required, for instance, in Germany[41]. These practical advantages notwithstanding, morphing also comes with potential disadvantages in that artifacts may arise during the morphing process[42,43]. Given that only the photographs of the fillers are digitally manipulated while the photograph of the suspect is not, the absence of morphing artifacts can serve as a cue to the identity of the suspect. In the worst case, these morphing artifacts could lead to unfair lineups from which witnesses may choose the suspect not because they recognize the suspect's face but because the suspect's facial photograph can be identified without relying on memory. We started by examining the fairness of morphed and non-morphed lineups using measures that were obtained from the traditional mock-witness task. From the mock-witness choices, we calculated Tredoux's *E* and the proportion of suspect selections as the most prominent measures of effective lineup size and biased suspect selection, respectively. Both measures provided evidence that morphed lineups were more unfair than non-morphed lineups. In Experiment 2, the 2-HT eyewitness identification model was used to estimate biased selection of the suspect directly from eyewitness identification decisions. As a first step, we deviated from the recommended standard procedure of the eyewitness task to make the eyewitness task as similar as possible to a mock-witness task. Specifically, lineup rejections were discouraged and participants were alerted that a photograph might stand out from the other photographs in the lineup. When these instructions were used—that are highly unusual for the eyewitness task but typical for the mock-witness task—, the model's biased-suspect-selection parameter *b* was enhanced in morphed simultaneous lineups in comparison to non-morphed simultaneous lineups, consistent with the measures of unfairness in the mock-witness task. Under these circumstances, using morphed fillers in simultaneous lineups thus lead to the biased selection of the suspects irrespective of their guilt. Based only on the results of Experiments 1 and 2 one may thus be tempted to conclude that the police must stop using morphing techniques to digitally manipulate filler photographs when the lineups are presented in the simultaneous format.

However, in Experiment 3, in which the pre-lineup instructions did not discourage participants from rejecting the lineup, the difference in the biased-suspect-selection parameter *b* between simultaneous morphed and non-morphed lineups was numerically reduced in comparison to Experiment 2 and no longer statistically significant. The difference in biased suspect selection between simultaneous morphed and non-morphed lineups was even completely absent in Experiment 4 in which the pre-lineup instructions did not discourage lineup rejections and did not alert participants that a photograph might stand out from the other photographs in the lineup. This situation most closely corresponds to the standard eyewitness task. The fact that the morphing manipulation did not affect eyewitness identification decisions in the standard eyewitness task contradicts the conclusion that would have to be drawn from the mock-witness data (Experiment 1) and the data obtained in a variant of the eyewitness task that closely mimicked the mock-witness task (Experiment 2). Similar contradictions between mock-witness data and eyewitness results have been reported in other studies[3,10,48]. Together, these results support the assumption that mock-witness choices may not be a good basis for drawing conclusions about eyewitness identification decisions[21].

Given that the model-based analysis did not yield signs of a morphing unfairness when lineup rejections were not discouraged and participants were not alerted that a photograph might stand out from the other photographs in the lineup, it is possible to assume that these two procedural differences between the mock-witness task and the typical eyewitness task are two major reasons as to why mock-witness choices fail to align with eyewitness identification decisions in the standard eyewitness task: First, whereas mock witnesses are typically required to choose a lineup member, eyewitnesses may choose to reject the lineup. Second, mock witnesses are made aware of the possibility that one lineup member might stand out from the others. Eyewitnesses, in contrast, must make a memory-based identification decision by matching each individual face to their memory representation of the culprit in order to be able to decide whether the culprit is in the lineup. Therefore, the unfairness of the lineup is overestimated in the mock-witness task in comparison to the standard eyewitness task.

Of course, the mock-witness task remains a valuable tool in actual criminal cases, that is, in the situation for which the task has been developed originally, as has been pointed out by Quigley-McBride and Wells[20].

This is so simply because there is currently no better alternative for assessing lineup fairness in practice where the goal is to ensure that a lineup is fair before it is presented to real witnesses. However, as the data presented here have shown, results obtained with mock witnesses may well differ from those obtained with eyewitnesses and thus should be used with caution. In lineup research, in contrast, a measurement model should be used which allows determining whether or not a lineup is unfair in the eyewitness identification situation proper. Otherwise, researchers may draw incorrect conclusions based on invalid fairness assessment procedures, which could lead practitioners to discard appropriate techniques for lineup construction. For instance, here we have shown that morphing artifacts affect mock-witness choices in simultaneous lineups. Such results may well lead policy makers to ban the morphing of photographs for lineup presentation and to eliminate this technique from the set of techniques the police is allowed to use in order to construct lineups. However, as we have shown, the same morphing artifacts that affect choices in situations in which participants have received instructions that are typical of the mock-witness task (Experiments 1 and 2) need not affect eyewitness identification decisions in a typical eyewitness task (Experiment 4). Given these results, there seems to be no reason to ban the morphing of photographs when constructing photo lineups, provided it can be ensured that witnesses receive standard lineup instructions and do not feel pressured to make an identification.

We included both simultaneous and sequential lineups in the present series of experiments because the 2-HT eyewitness identification model is a tool for measuring lineup fairness in both types of lineup formats[32]. This is a distinguishing feature of the present model given that previous research has shown that the mock-witness task is of limited use in estimating lineup fairness in sequential lineups[21,48]. It has previously been shown that unfair simultaneous lineups led to more identifications of innocent suspects than unfair sequential lineups, suggesting that sequential lineups provide more protection for the innocent suspect when the lineup is unfair[3,51,52]. This conclusion is supported by the findings reported here. The results of Experiments 2, 3 and 4 consistently showed no effect of the morphing manipulation on biased suspect selection in sequential lineups, even when the instructions closely resembled those of a mock-witness task. When lineup identifications are made under conditions that do not qualify as best practices—that is, when lineups are unfair and instructions encourage non-memory-based decisions—, sequential lineups provide some protection against unfairness in comparison to simultaneous lineups.

For quite some time, another advantage of sequential lineups seemed to be that, compared with simultaneous lineups, sequential lineups have often been found to be associated with a higher diagnosticity ratio[69]—that is, a higher ratio of the proportion of correct culprit identifications to the proportion of false innocent-suspect identifications[70,71]. This result seemed to indicate that sequential lineups perform better than simultaneous lineups when the goal is to separate culprits from innocent suspects. However, it has been argued that the diagnosticity ratio is an inadequate measure of lineup performance because it confounds the ability to distinguish between a culprit and an innocent suspect with response bias [e.g.,[72]]. Receiver Operating Characteristic (ROC) analyses do not have this problem and have shown either that simultaneous lineups perform better than sequential lineups[23,49,59,73,74] or that sequential and simultaneous lineups perform equally well[55,75–77]. ROC analyses are said to have the advantage of delivering a performance measure that is not confounded by response bias[23]. However, ROC analyses focussing on the partial area under the curve—that have become commonplace in lineup research—are based on the proportion of correct culprit identifications and false innocent-suspect identifications, as a consequence of which they do not exploit the information contained in filler identifications and lineup rejections separately; these data categories are combined based on the reasoning that both filler identifications and lineup rejections have no legal consequences [[4,78], but see[79] for an interesting suggestion on how to create a full ROC based on the full range of response categories]. However, there is information to be gained when these two response categories are analyzed separately. For instance, a filler identification in a culprit-absent lineup is an error. A lineup rejection in a culprit-absent lineup is a correct response. Obviously, many such filler identifications and few lineup rejections indicate bad performance, whereas few such filler identifications and many lineup rejections indicate good performance. The 2-HT eyewitness identification model used here exploits this information in that it takes into account the full range of data categories available from lineup procedures (see Fig. 1). In doing so, the model provides measures for four types of cognitive processes of which we have so far focused on the process of biased suspect selection (represented by parameter *b*) exclusively. We would now like to focus on the process of culprit-presence detection represented by parameter *dP*. An advantage of the 2-HT eyewitness identification model is that parameter *dP* is not confounded with lineup fairness[32,34], that is, parameter *dP* is a pure measure of culprit-presence detection even in unfair lineups. In the model-based analyses reported here, the estimates of parameter *dP* were consistently higher in the simultaneous lineup conditions than in the sequential lineup conditions (see Table 2). This difference was significant in Experiment 2, $\Delta G^2(1) = 8.25$, $p = 0.004$, $w = 0.05$, and Experiment 4, $\Delta G^2(1) = 4.80$, $p = 0.028$, $w = 0.04$, but not in Experiment 3, $\Delta G^2(1) = 3.78$, $p = 0.052$, $w = 0.04$. A small superiority of simultaneous over sequential lineups was also found by Winter et al.[32] when applying the 2-HT eyewitness identification model to both simultaneous and sequential lineups. This pattern in the results based on the 2-HT eyewitness identification model is in good agreement with the results of ROC-based analyses in which a superiority of simultaneous over sequential lineups was sometimes found [e.g.,[23,49,59,73,74]] but not always [e.g.,[55,75–77]].

Parallel to the results of Winter et al.[32], we also found a consistently higher probability of guessing-based selection (captured by parameter *g*) in sequential lineups in comparison to simultaneous lineups in Experiments 2, 3 and 4 (see Table 2). Note that this general pattern is already evident from surface-level data: The rate of identifications was consistently higher in the sequential lineup conditions than in the simultaneous lineup conditions (0.81 vs. 0.67 in Experiment 2, 0.78 vs. 0.58 in Experiment 3 and 0.76 vs. 0.55 in Experiment 4). At first glance, this may seem unexpected given that previous research has indicated that sequential lineups are

associated with more conservative responding than simultaneous lineups [e.g.,[75]]. However, in contrast to many previous studies, we did not inform our participants in the sequential lineup conditions that only their first "yes" response counts. Instead, we explicitly followed standard police protocols[24,50,65] and the original protocol outlined by Lindsay and Wells[80] and continued the presentation of the sequential lineup even after an early positive response; only the participant's final decision was coded as their identification decision. This differs from the first-yes-counts protocol that is typically used with sequential lineups in eyewitness research. Horry et al.[65] have shown that this first-yes-counts protocol systematically reduces suspect identifications and increases lineup rejections by discouraging participants from guessing. These results are easily explained: When only the first "yes" response counts, eyewitnesses may shy away from 'using up' their only identification response too early in the sequence because they do not know whether there will be a better alternative later in sequence. This will necessarily lead to conservative responding. In contrast, the (more realistic) lineup protocol that has been used here can be expected to produce relatively liberal responding and thus a higher prevalence of guessing-based selections among lineup members in the sequential lineup. However, this rather interesting aspect of the present study has to be further addressed in future experiments.

Recently, Quigley-McBride and Wells[20] have proposed an alternative method to measure lineup fairness directly from actual eyewitness data. Specifically, they have recommend calculating the *resultant* lineup fairness based on the innocent-suspect identifications and filler identifications in culprit-absent lineups. Given that it seems interesting to compare these resultant lineup-fairness measures with the biased-suspect-selection parameter b of the 2-HT eyewitness identification model, we calculated the average resultant proportion of suspect selections (i.e., innocent-suspect identifications ÷ [innocent-suspect identifications + filler identifications in culprit-absent lineups]) and the average resultant Tredoux's E for Experiments 2, 3 and 4. Note that these calculations are based only on the identifications in culprit-absent lineups whereas the 2-HT eyewitness identification model takes into account all data of both culprit-present lineups and culprit-absent lineups. As a result of being based on a reduced data set, we may expect more variability in the values calculated for the resultant-lineup fairness measures. Still, the resultant proportions of suspect selections reflect the unfairness of morphed opposed to non-morphed lineups in a way that is largely parallel to that of the biased-suspect-selection parameter b in the present series of experiments (0.37 vs. 0.29, 0.35 vs. 0.26, 0.29 vs. 0.28 for morphed vs. non-morphed simultaneous lineups in Experiments 2, 3 and 4, respectively; 0.19 vs. 0.22, 0.18 vs. 0.21, 0.25 vs. 0.24 for morphed vs. non-morphed sequential lineups in Experiments 2, 3 and 4, respectively). In addition, the resultant proportions of suspect selections reflect the higher unfairness in simultaneous lineups than in sequential lineups. The resultant Tredoux's E was calculated separately for each of the four simultaneous and sequential morphed and non-morphed lineups before an average resultant Tredoux's E was computed for the simultaneous and sequential morphed and non-morphed lineup conditions, as in Experiment 1. The average resultant Tredoux's E was descriptively smaller for morphed lineups than for non-morphed lineups in Experiment 2, but the data pattern is more variable in Experiments 3 and 4 (4.36 vs. 4.56, 4.41 vs. 4.26, 4.75 vs. 4.16 for morphed vs. non-morphed simultaneous lineups in Experiments 2, 3 and 4, respectively; 5.07 vs. 5.39, 5.37 vs. 5.16, 5.22 vs. 5.29 for morphed vs. non-morphed sequential lineups in Experiments 2, 3 and 4, respectively). In all experiments, the resultant Tredoux's E was descriptively smaller in simultaneous lineups than in sequential lineups (more details on the analyses of the resultant lineup-fairness measures and the distribution of eyewitness identification decisions across lineup members are provided in the Open Science Framework repository at https://osf.io/zaybc/).

A limitation of the present research is that a cross-experiment comparison was used to demonstrate that morphing artifacts cause unfairness in an anomalous identification situation—comparable to that of mock witnesses—but do not enhance biased suspect selection in a standard eyewitness task. Future research could extend the present research by performing a within-experiment comparison to examine more directly how the morphing effect on biased suspect selection interacts with the different lineup conditions. Another limitation of the present study is that only one of the two possible types of mock-witness tasks was used here. Participants were asked to choose the person who stands out from the other lineup members. In another variant of the mock-witness task, participants are provided with a description of the culprit as the basis for their choice [e.g.,[11,12]]. Given that it cannot be taken for granted that the search processes are the same for these two different types of mock-witness tasks, future research should focus on whether the same conclusions can be obtained with the description-based mock-witness task.

Conclusion

Lineup fairness is a critical factor affecting the likelihood of misidentifications, yet there is surprisingly little research on how to determine the fairness of lineups. Traditionally, researchers have relied on the mock-witness task to evaluate lineup fairness[10] although this method has been criticized based on the fact that the task of mock witnesses differs from that of eyewitnesses [e.g.,[21]]. The present series of experiments not only demonstrates that those who had questioned the usefulness of the mock-witness task[10,15,20,21,48] were correct but also sheds light on the crucial differences between the mock-witness task and the eyewitness task that are responsible for the divergent effects. While the mock-witness task showed that morphed lineups were more unfair than non-morphed lineups, the morphing manipulation did not affect eyewitness identification decisions in a typical lineup procedure. This discrepancy was due to two task differences: First, unlike eyewitnesses, mock witnesses are not allowed to reject lineups. Second, mock witnesses are made aware of the possibility that one lineup member might stand out from the others. In contrast, eyewitnesses must match each lineup member to their memory representation of the culprit. In lineup research, it therefore seems desirable to measure lineup fairness directly from eyewitness data using a measurement model such as the 2-HT eyewitness identification model rather than to rely on mock-witness-based measures.

Data availability

All raw data analyzed during this study are available in the manuscript or in the Open Science Framework repository (https://osf.io/zaybc/).

Received: 24 August 2022; Accepted: 7 April 2023
Published online: 18 April 2023

References

1. Innocence Project. *Exonerate the Innocent*. https://innocenceproject.org/exonerate/ (2023). Accessed 14 May 2023.
2. Fitzgerald, R. J., Price, H. L., Oriet, C. & Charman, S. D. The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis. *Psychol. Public Policy Law* **19**, 151–164. https://doi.org/10.1037/a0030618 (2013).
3. Steblay, N. K. & Wells, G. L. Assessment of bias in police lineups. *Psychol. Public Policy Law* **26**, 393–412. https://doi.org/10.1037/law0000287 (2020).
4. Wells, G. L., Smalarz, L. & Smith, A. M. ROC analysis of lineups does not measure underlying discriminability and has limited value. *J. Appl. Res. Mem. Cogn.* **4**, 313–317. https://doi.org/10.1016/j.jarmac.2015.08.008 (2015).
5. Smith, A. M., Wells, G. L., Lindsay, R. C. L. & Penrod, S. D. Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law Hum. Behav.* **41**, 127–145. https://doi.org/10.1037/lhb0000219 (2017).
6. Smith, A. M., Wells, G. L., Smalarz, L. & Lampinen, J. M. 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). *Psychol. Sci.* **29**, 1548–1551. https://doi.org/10.1177/0956797617698528 (2018).
7. Wetmore, S. A. *et al.* Effect of retention interval on showup and lineup performance. *J. Appl. Res. Mem. Cogn.* **4**, 8–14. https://doi.org/10.1016/j.jarmac.2014.07.003 (2015).
8. Colloff, M. F., Wade, K. A. & Strange, D. Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychol. Sci.* **27**, 1227–1239. https://doi.org/10.1177/0956797616655789 (2016).
9. Doob, A. N. & Kirshenbaum, H. M. Bias in police lineups—Partial remembering. *J. Policy Sci. Admin.* **1**, 287–293 (1973).
10. Lee, J., Mansour, J. K. & Penrod, S. D. Validity of mock-witness measures for assessing lineup fairness. *Psychol. Crime Law* **28**, 215–245. https://doi.org/10.1080/1068316X.2021.1905811 (2022).
11. Humphries, J. E., Holliday, R. E. & Flowe, H. D. Faces in motion: Age-related changes in eyewitness identification performance in simultaneous, sequential, and elimination video lineups. *Appl. Cogn. Psychol.* **26**, 149–158. https://doi.org/10.1002/acp.1808 (2012).
12. Mansour, J. K., Beaudry, J. L., Kalmet, N., Bertrand, M. I. & Lindsay, R. C. L. Evaluating lineup fairness: Variations across methods and measures. *Law Hum. Behav.* **41**, 103–115. https://doi.org/10.1037/lhb0000203 (2017).
13. Brigham, J. C., Meissner, C. A. & Wasserman, A. W. Applied issues in the construction and expert assessment of photo lineups. *Appl. Cogn. Psychol.* **13**, S73–S92. https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+%3cS73::AID-ACP631%3e3.3.CO;2-W (1999).
14. Flowe, H. D. & Humphries, J. E. An examination of criminal face bias in a random sample of police lineups. *Appl. Cogn. Psychol.* **25**, 265–273. https://doi.org/10.1002/acp.1673 (2011).
15. Malpass, R. S., Tredoux, C. G. & McQuiston-Surrett, D. E. Lineup construction and lineup fairness. In *The Handbook of Eyewitness Psychology, Memory for People* Vol. 2 (eds Lindsay, R. C. L. *et al.*) 155–178 (Lawrence Erlbaum Associates, 2007).
16. Bergold, A. N. & Heaton, P. Does filler database size influence identification accuracy? *Law Hum. Behav.* **42**, 227–243. https://doi.org/10.1037/lhb0000289 (2018).
17. Zarkadi, T., Wade, K. A. & Stewart, N. Creating fair lineups for suspects with distinctive features. *Psychol. Sci.* **20**, 1448–1453. https://doi.org/10.1111/j.1467-9280.2009.02463.x (2009).
18. Malpass, R. S. & Lindsay, R. C. Measuring lineup fairness. *Appl. Cogn. Psychol.* **13**, S1–S7. https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+%3cS1::AID-ACP678%3e3.0.CO;2-9 (1999).
19. Tredoux, C. G. Statistical inference on measures of lineup fairness. *Law Hum. Behav.* **22**, 217–237. https://doi.org/10.1023/A:1025746220886 (1998).
20. Quigley-McBride, A. & Wells, G. L. Methodological considerations in eyewitness identification experiments. In *Methods, Measures, and Theories in Eyewitness Identification Tasks* (eds Smith, A. M. *et al.*) 85–112 (Taylor and Francis, 2021). https://doi.org/10.4324/9781003138105.
21. Corey, D., Malpass, R. S. & McQuiston, D. E. Parallelism in eyewitness and mock witness identifications. *Appl. Cogn. Psychol.* **13**, S41–S58. https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+%3cS41::AID-ACP632%3e3.0.CO;2-A (1999).
22. Wells, G. L., Leippe, M. R. & Ostrom, T. M. Guidelines for empirically assessing the fairness of a lineup. *Law Hum. Behav.* **3**, 285–293. https://doi.org/10.1007/BF01039807 (1979).
23. Mickes, L., Flowe, H. D. & Wixted, J. T. Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *J. Exp. Psychol. Appl.* **18**, 361–376. https://doi.org/10.1037/a0030609 (2012).
24. Wells, G. L., Steblay, N. K. & Dysart, J. E. Double-blind photo lineups using actual eyewitnesses: An experimental test of a sequential versus simultaneous lineup procedure. *Law Hum. Behav.* **39**, 1–14. https://doi.org/10.1037/lhb0000096 (2015).
25. Smith, A. M., Wilford, M. M., Quigley-McBride, A. & Wells, G. L. Mistaken eyewitness identification rates increase when either witnessing or testing conditions get worse. *Law Hum. Behav.* **43**, 358–368. https://doi.org/10.1037/lhb0000334 (2019).
26. Technical Working Group for Eyewitness Evidence. *Eyewitness Evidence: A Guide for Law Enforcement* (National Institute of Justice, 1999).
27. Wells, G. L. & Quigley-McBride, A. Applying eyewitness identification research to the legal system: A glance at where we have been and where we could go. *J. Appl. Res. Mem. Cogn.* **5**, 290–294. https://doi.org/10.1016/j.jarmac.2016.07.007 (2016).
28. Wells, G. L. *et al.* Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law Hum. Behav.* **44**, 3–36. https://doi.org/10.1037/lhb0000359 (2020).
29. Malpass, R. S. & Devine, P. G. Eyewitness identification: Lineup instructions and the absence of the offender. *J. Appl. Psychol.* **66**, 482–489. https://doi.org/10.1037/0021-9010.66.4.482 (1981).
30. Clark, S. E. A re-examination of the effects of biased lineup instructions in eyewitness identification. *Law Hum. Behav.* **29**, 575–604. https://doi.org/10.1007/s10979-005-7121-1 (2005).
31. Lampinen, J. M. *et al.* Comparing detailed and less detailed pre-lineup instructions. *Appl. Cogn. Psychol.* **34**, 409–424. https://doi.org/10.1002/acp.3627 (2020).
32. Winter, K., Menne, N. M., Bell, R. & Buchner, A. Experimental validation of a multinomial processing tree model for analyzing eyewitness identification decisions. *Sci. Rep.* **12**, 15571. https://doi.org/10.1038/s41598-022-19513-w (2022).
33. Wixted, J. T. & Mickes, L. A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychol. Rev.* **121**, 262–276. https://doi.org/10.1037/a0035940 (2014).
34. Menne, N. M., Winter, K., Bell, R. & Buchner, A. A validation of the two-high threshold eyewitness identification model by reanalyzing published data. *Sci. Rep.* **12**, 13379. https://doi.org/10.1038/s41598-022-17400-y (2022).

35. Batchelder, W. H. & Riefer, D. M. Theoretical and empirical review of multinomial process tree modeling. *Psychon. Bull. Rev.* **6**, 57–86. https://doi.org/10.3758/BF03210812 (1999).
36. Erdfelder, E. *et al.* Multinomial processing tree models: A review of the literature. *Z. Psychol./J. Psychol.* **217**, 108–124. https://doi.org/10.1027/0044-3409.217.3.108 (2009).
37. Schmidt, O., Erdfelder, E. & Heck, D. W. Tutorial on multinomial processing tree modeling: How to develop, test, and extend MPT models. *Psychol. Methods*. https://doi.org/10.1037/met0000561 (in press)
38. Peska, L. & Trojanova, H. Towards recommender systems for police photo lineup. In *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems* 19–23 (Association for Computing Machinery, 2017). https://doi.org/10.1145/3125486.3125490.
39. Fitzgerald, R. J., Oriet, C. & Price, H. L. Suspect filler similarity in eyewitness lineups: A literature review and a novel methodology. *Law Hum. Behav.* **39**, 62–74. https://doi.org/10.1037/lhb0000095 (2015).
40. Lucas, C. A., Brewer, N. & Palmer, M. A. Eyewitness identification: The complex issue of suspect-filler similarity. *Psychol. Public Policy Law* **27**, 151–169. https://doi.org/10.1037/law0000243 (2021).
41. Federal Ministry of the Interior of the State of North-Rhine-Westphalia. *Wahllichtbildvorlage im Strafverfahren [Photo lineups in criminal proceedings]*. https://recht.nrw.de/lmi/owa/br_bes_text?anw_nr=1&gld_nr=2&ugl_nr=2056&bes_id=9147&val=9147&ver=7&sg=&aufgehoben=N&menu=1 (2023). Accessed 14 May 2023.
42. Alley, T. R. & Cunningham, M. R. Article commentary: Averaged faces are attractive, but very attractive faces are not average. *Psychol. Sci.* **2**, 123–125. https://doi.org/10.1111/j.1467-9280.1991.tb00113.x (1991).
43. Borghi, G., Franco, A., Graffieti, G. & Maltoni, D. Automated artifact retouching in morphed images with attention maps. *IEEE Access* **9**, 136561–136579. https://doi.org/10.1109/ACCESS.2021.3117718 (2021).
44. Leiner, D. J. *SoSci Survey* [computer software]. https://www.soscisurvey.de (2021).
45. Minear, M. & Park, D. C. A lifespan database of adult facial stimuli. *Behav. Res. Methods Instrum. Comput.* **36**, 630–633. https://doi.org/10.3758/BF03206543 (2004).
46. Langner, O. *et al.* Presentation and validation of the Radboud faces database. *Cogn. Emot.* **24**, 1377–1388. https://doi.org/10.1080/02699930903485076 (2010).
47. Thomaz, C. E. & Giraldi, G. A. A new ranking method for principal components analysis and its application to face image analysis. *Image Vis. Comput.* **28**, 902–913. https://doi.org/10.1016/j.imavis.2009.11.005 (2010).
48. Lindsay, R. C., Smith, S. M. & Pryke, S. Measures of lineup fairness: Do they postdict identification accuracy? *Appl. Cogn. Psychol.* **13**, S93–S107. https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+%3cS93::AID-ACP633%3e3.0.CO;2-X (1999).
49. Seale-Carlisle, T. M. & Mickes, L. US line-ups outperform UK line-ups. *R. Soc. Open Sci.* **3**, 160300. https://doi.org/10.1098/rsos.160300 (2016).
50. German Federal Ministry of the Interior and Community. *Richtlinien für das Strafverfahren und das Bußgeldverfahren (RiStBV) [Guidelines for criminal proceedings and fine proceedings]*. https://www.verwaltungsvorschriften-im-internet.de/bsvwvbund_01011977_420821R5902002.htm (2021). Accessed 14 May 2023.
51. Lindsay, R. C. L. *et al.* Biased lineups: Sequential presentation reduces the problem. *J. Appl. Psychol.* **76**, 796–802. https://doi.org/10.1037/0021-9010.76.6.796 (1991).
52. Carlson, C. A., Gronlund, S. D. & Clark, S. E. Lineup composition, suspect position, and the sequential lineup advantage. *J. Exp. Psychol. Appl.* **14**, 118–128. https://doi.org/10.1037/1076-898X.14.2.118 (2008).
53. Faul, F., Erdfelder, E., Lang, A. & Buchner, A. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191. https://doi.org/10.3758/bf03193146 (2007).
54. Karageorge, A. & Zajac, R. Exploring the effects of age and delay on children's person identifications: Verbal descriptions, lineup performance, and the influence of wildcards. *Br. J. Psychol.* **102**, 161–183. https://doi.org/10.1348/000712610X507902 (2011).
55. Gronlund, S. D. *et al.* Showups versus lineups: An evaluation using ROC analysis. *J. Appl. Res. Mem. Cogn.* **1**, 221–228. https://doi.org/10.1016/j.jarmac.2012.09.003 (2012).
56. Smith, A. M. Why do mistaken identification rates increase when either witnessing or testing conditions get worse? *J. Appl. Res. Mem. Cogn.* **9**, 495–507. https://doi.org/10.1016/j.jarmac.2020.08.002 (2020).
57. Brigham, J. C. Target person distinctiveness and attractiveness as moderator variables in the confidence-accuracy relationship in eyewitness identifications. *Basic Appl. Soc. Psychol.* **11**, 101–115. https://doi.org/10.1207/s15324834basp1101_7 (1990).
58. Wilcock, R. & Bull, R. Novel lineup methods for improving the performance of older eyewitnesses. *Appl. Cogn. Psychol.* **24**, 718–736. https://doi.org/10.1002/acp.1582 (2010).
59. Dobolyi, D. G. & Dodson, C. S. Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. *J. Exp. Psychol. Appl.* **19**, 345–357. https://doi.org/10.1037/a0034596 (2013).
60. Palmer, M. A., Brewer, N., Weber, N. & Nagesh, A. The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *J. Exp. Psychol. Appl.* **19**, 55–71. https://doi.org/10.1037/a0031602 (2013).
61. Hobson, Z., Wilcock, R. & Valentine, T. Multiple suspect showing: A survey of police identification officers. *Policing* **7**, 79–87. https://doi.org/10.1093/police/pas021 (2012).
62. Tupper, N., Sauerland, M., Sauer, J. D. & Hope, L. Eyewitness identification procedures for multiple perpetrator crimes: A survey of police in Sweden, Belgium, and the Netherlands. *Psychol. Crime Law* **25**, 992–1007. https://doi.org/10.1080/1068316X.2019.1611828 (2019).
63. Hobson, Z. J. & Wilcock, R. Eyewitness identification of multiple perpetrators. *Int. J. Police Sci. Manag.* **13**, 286–296. https://doi.org/10.1350/ijps.2011.13.4.253 (2011).
64. Oriet, C. & Fitzgerald, R. J. The single lineup paradigm: A new way to manipulate target presence in eyewitness identification experiments. *Law Hum. Behav.* **42**, 1–12. https://doi.org/10.1037/lhb0000272 (2018).
65. Horry, R., Fitzgerald, R. J. & Mansour, J. K. "Only your first yes will count": The impact of prelineup instructions on sequential lineup decisions. *J. Exp. Psychol. Appl.* **27**, 170–186. https://doi.org/10.1037/xap0000337 (2021).
66. Moshagen, M. multiTree: A computer program for the analysis of multinomial processing tree models. *Behav. Res. Methods* **42**, 42–54. https://doi.org/10.3758/BRM.42.1.42 (2010).
67. Hu, X. & Batchelder, W. H. The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika* **59**, 21–47. https://doi.org/10.1007/BF02294263 (1994).
68. Brewer, N. & Wells, G. L. The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *J. Exp. Psychol. Appl.* **12**, 11–30. https://doi.org/10.1037/1076-898X.12.1.11 (2006).
69. Wells, G. L. & Lindsay, R. C. L. On estimating the diagnosticity of eyewitness nonidentifications. *Psychol. Bull.* **88**, 776–784. https://doi.org/10.1037/0033-2909.88.3.776 (1980).
70. Steblay, N., Dysart, J., Fulero, S. & Lindsay, R. C. Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law Hum. Behav.* **25**, 459–473. https://doi.org/10.1023/A:1012888715007 (2001).
71. Steblay, N. K., Dysart, J. E. & Wells, G. L. Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychol. Public Policy Law* **17**, 99–139. https://doi.org/10.1037/a0021650 (2011).
72. Gronlund, S. D., Wixted, J. T. & Mickes, L. Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Curr. Dir. Psychol. Sci.* **23**, 3–10. https://doi.org/10.1177/0963721413498891 (2014).

73. Carlson, C. A. & Carlson, M. A. An evaluation of lineup presentation, weapon presence, and a distinctive feature using ROC. *J. Appl. Res. Mem. Cogn.* **3**, 45–53. https://doi.org/10.1016/j.jarmac.2014.03.004 (2014).
74. Seale-Carlisle, T. M., Wetmore, S. A., Flowe, H. D. & Mickes, L. Designing police lineups to maximize memory performance. *J. Exp. Psychol. Appl.* **25**, 410–430. https://doi.org/10.1037/xap0000222 (2019).
75. Clark, S. E. Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspect. Psychol. Sci.* **7**, 238–259. https://doi.org/10.1177/1745691612439584 (2012).
76. Andersen, S. M., Carlson, C. A., Carlson, M. A. & Gronlund, S. D. Individual differences predict eyewitness identification performance. *Pers. Individ. Differ.* **60**, 36–40. https://doi.org/10.1016/j.paid.2013.12.011 (2014).
77. Meisters, J., Diedenhofen, B. & Musch, J. Eyewitness identification in simultaneous and sequential lineups: An investigation of position effects using receiver operating characteristics. *Memory* **26**, 1297–1309. https://doi.org/10.1080/09658211.2018.1464581 (2018).
78. Wells, G. L., Smith, A. M. & Smalarz, L. ROC analysis of lineups obscures information that is critical for both theoretical understanding and applied purposes. *J. Appl. Res. Mem. Cogn.* **4**, 324–328. https://doi.org/10.1016/j.jarmac.2015.08.010 (2015).
79. Smith, A. M., Yang, Y. & Wells, G. L. Distinguishing between investigator discriminability and eyewitness discriminability: A method for creating full receiver operating characteristic curves of lineup identification performance. *Perspect. Psychol. Sci.* **15**, 589–607. https://doi.org/10.1177/1745691620902426 (2020).
80. Lindsay, R. & Wells, G. L. Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *J. Appl. Psychol.* **70**, 556–564. https://doi.org/10.1037/0021-9010.70.3.556 (1985).

Author contributions

N.M.M., K.W., R.B. and A.B. contributed to the study conception, design, material preparation and data analysis. N.M.M. collected the data and wrote the first draft of the manuscript. K.W., R.B. and A.B. critically revised the manuscript. All authors gave final approval for publication.

Funding

Open Access funding enabled and organized by Projekt DEAL. The work reported herein was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—BU 945/10-1, project number 456214986.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to N.M.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

scientific reports

Check for updates

OPEN

The effects of lineup size on the processes underlying eyewitness decisions

Nicola Marie Menne✉, Kristina Winter, Raoul Bell & Axel Buchner

Here we apply the two-high threshold eyewitness identification model to identify the effects of lineup size on the detection-based and non-detection-based processes underlying eyewitness decisions. In Experiment 1, lineup size was manipulated by showing participants simultaneous or sequential lineups that contained either three or six persons. In Experiment 2, the lineups contained either two or five persons. In both experiments, the culprit was better detected in smaller than in larger lineups. Furthermore, participants made fewer guessing-based selections in smaller than in larger lineups. However, guessing-based selection in larger lineups was not increased to a level sufficient to offset the effect of increased protection of suspects in larger lineups due to the fact that the guessing-based selections that occur are distributed across more persons. The results show that increasing the lineup size causes several changes in the detection-based and non-detection-based processes underlying eyewitness decisions.

Eyewitness identification via lineup procedures is a major source of evidence in criminal investigations. In a lineup procedure, an eyewitness is presented with a suspect (who is guilty or innocent) along with a number of fillers (who are known to be innocent) and is asked to make an identification or to reject the lineup. However, human memory is unreliable and highly prone to error. According to the Innocence Project[1], mistaken eyewitness identifications have played a role in 70 % of the more than 375 wrongful convictions that have been revealed by DNA analyses. While many structural and procedural aspects of lineups have been studied, including the lineup presentation format (e.g.,[2]), the instructions given to eyewitnesses (e.g.,[3]) or the characteristics of fillers (e.g.,[4]), there has been relatively little research addressing the question of how the number of fillers in a lineup affects eyewitness decisions. In addition, studies to date have mainly focused on the question as to which lineup size should be preferred by relying on overall measures of lineup quality (e.g.,[5–7]). The present study aims to complement the existing literature by providing a more detailed dissection of how lineup size affects the processes underlying eyewitness decisions in lineups. Improving our understanding of the processes underlying lineup size effects seems desirable considering the large variability in lineup size policies across jurisdictions. For instance, in the United States, a lineup most often contains five fillers[8], the police in the United Kingdom typically use eight fillers[9] and in Germany, the recommended lineup includes at least seven fillers[10]. Before considering the potential effects of lineup size on the latent processes underlying eyewitness decisions, it must be pointed out that it is plausible that increasing the lineup size diminishes the probability of suspect identifications simply due to the mathematical consequences of selecting among a larger group of persons. If participants randomly guess among the available options, the sampling probability of selecting the suspect among the fillers is inversely proportional to the size of the lineup. However, it is quite possible that variations in the number of persons that have to be considered in a lineup also affect the latent detection-based and non-detection-based processes underlying eyewitness decisions in addition to this sampling probability. Here we rely on the well-validated two-high threshold (2-HT) eyewitness identification model[11,12] to disentangle the effects of lineup size on the detection-based and non-detection-based processes involved in eyewitness decisions.

Multinomial processing tree models, to which class the 2-HT eyewitness identification model belongs, are statistical models for categorical data that have been widely used to infer latent cognitive processes from observable behavior (for reviews, see[13,14]). These models form an easily accessible class of measurement models for which both excellent tutorials[15] and easy-to-use free software for parameter estimation and statistical hypothesis tests exist[16]. Multinomial processing tree models are based on the assumption that an observable response originates from a sequence of latent processes. These sequences are visualized as branches in a processing tree. The

Department of Experimental Psychology, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. ✉email: nicola.marie.menne@hhu.de

branches consist of links and nodes. The links terminating in nodes represent probabilities with which certain cognitive processes occur. These are the parameters of the model that can be measured based on observable data.

Based on the full range of data categories in a typical lineup task (i.e., suspect identifications, filler identifications and lineup rejections in both culprit-present and culprit-absent lineups), the 2-HT eyewitness identification model provides measures of the latent processes from which eyewitness decisions originate (Fig. 1). Specifically, the model serves to separately measure the detection of culprit presence (parameter *dP*), the detection of culprit absence (parameter *dA*), guessing-based selection among the lineup members (parameter *g*) and biased suspect selection in unfair lineups (parameter *b*). The model has the further advantage that it incorporates the inverse relationship between the random-sampling probability and lineup size in terms of a fixed constant that is independent of the parameters representing the detection-based and non-detection-based processes that have to be estimated from the data. In validation studies, it has been empirically shown that the parameters of the 2-HT eyewitness identification model sensitively reflect manipulations of the processes they were intended to measure in both simultaneous and sequential lineups[11,12]. The model has already been successfully applied to examine the effects of first-yes-counts instructions on guessing-based selection[17] and to measure the effects of lineup fairness[18].

The upper tree in Fig. 1 represents the processes that may occur in response to culprit-present lineups. The presence of the culprit is detected with probability *dP*, resulting in a correct identification of the culprit. This process is enhanced under conditions that facilitate memory formation and retrieval such as the better encoding of the culprit's face[11,12]. If the presence of the culprit is not detected, which occurs with probability 1 – *dP*, two types of non-detection-based processes may still lead to a correct culprit identification in culprit-present lineups. With probability *b*, biased selection of the culprit may occur because the culprit stands out from the other persons in an unfair lineup. For instance, parameter *b* increases if the suspect has distinctive facial features (e.g., birthmarks, tattoos, nose piercings) that make the suspect stand out from the fillers[11,12]. No biased suspect selection occurs with the complementary probability 1 – *b*. In this case, participants may still select one of the persons

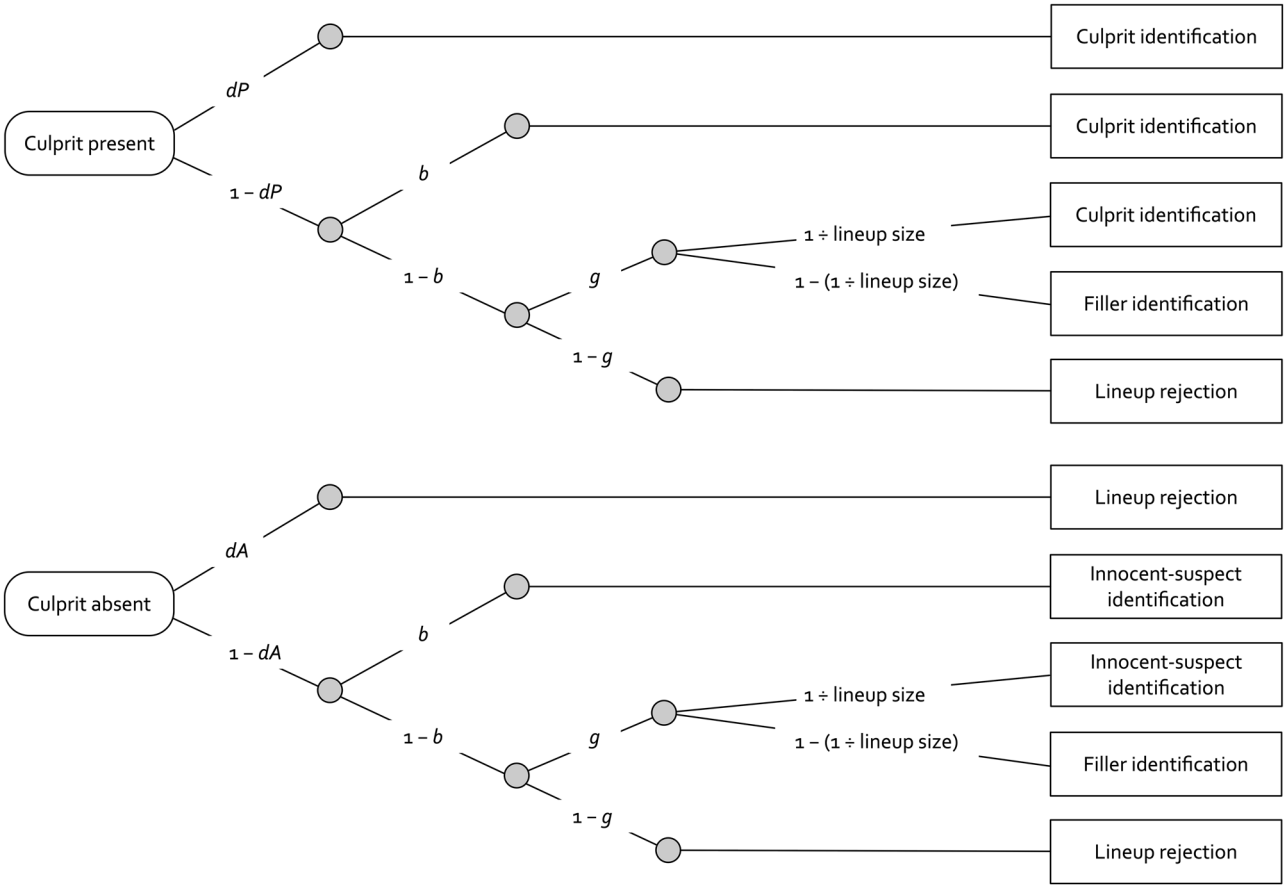

Figure 1. Illustration of the 2-HT eyewitness identification model. Rounded rectangles on the left side represent the two different lineup types presented at test (culprit-present and culprit-absent lineups) and the rectangles on the right side represent the observable response categories. The letters along the branches represent the probabilities of the latent processes postulated by the model (*dP*: probability of detecting the presence of the culprit; *b*: probability of biased selection of the suspect; *g*: probability of guessing-based selection among the lineup members; *dA*: probability of detecting the absence of the culprit). The random-sampling probability that is given by 1 ÷ lineup size is not a to-be-estimated parameter but a direct function of the number of persons in the lineup.

in the lineup based on guessing with probability *g*. Parameter *g* increases, for instance, if the lineup instructions insinuate that the culprit is in the lineup[11,12]. Guessing-based selection leads either to the identification of the culprit with the random-sampling probability that is equal to 1 ÷ lineup size or to the identification of one of the fillers with the complementary probability 1 – (1 ÷ lineup size). Alternatively, participants may not select a lineup member based on guessing with probability 1 – *g*, in which case the lineup is falsely rejected.

The lower tree of the 2-HT eyewitness identification model represents the processes that may occur in response to culprit-absent lineups. The absence of the culprit is detected with probability *dA*, leading to a correct lineup rejection. This may occur, for instance, if all lineup members in a culprit-absent lineup can be ruled out as the culprit because their faces share a common feature (e.g., birthmarks) that the culprit did not have[12]. Culprit-absence detection fails with probability 1 – *dA*, in which case the non-detection-based processes occur in the same way as in culprit-present lineups. With probability *b*, biased selection of the innocent suspect occurs because the innocent suspect stands out from the other persons in the lineup. No biased suspect selection occurs with probability 1 – *b*. In this case, participants may still make a guessing-based selection with probability *g*, resulting in either the identification of the innocent suspect with probability 1 ÷ lineup size or the identification of a filler with probability 1 – (1 ÷ lineup size). If participants do not make a guessing-based selection, which occurs with probability 1 – *g*, then the lineup is correctly rejected. More details can be found elsewhere[11,12].

Focusing on the role of lineup size in the model immediately clarifies the mathematical consequences of larger lineups on suspect identifications: The random-sampling probability that expresses the probability with which guessing-based selection leads to the identification of the suspect decreases with increasing lineup size. For instance, in a three-person lineup, a guessing-based selection will lead to a suspect identification with a random-sampling probability of 1/3, whereas in a six-person lineup, the random-sampling probability is reduced to 1/6.

Here we report two well-powered experiments to examine the effects of lineup size on the detection-based and non-detection-based processes underlying eyewitness decisions. Participants viewed a video of a staged crime and were subsequently presented with four lineups. In Experiment 1, lineup size was manipulated between participants by presenting lineups consisting of either three or six persons. Experiment 2 served as a conceptual replication of Experiment 1 with the only difference being that lineup size was manipulated by presenting lineups consisting of either two or five persons. In addition to simultaneous lineups (in which the lineup members are presented together), we also presented sequential lineups (in which the lineup members are presented one at a time) because sequential lineups have become the standard format for presenting lineups in countries such as the United Kingdom or Germany[10,19].

The present study avoids two methodological problems in lineup size research that have been pointed out by Juncu and Fitzgerald[20]. First, in most studies on lineup size, no designated innocent suspect was included in culprit-absent lineups. As a consequence, the rate of innocent-suspect identifications had to be estimated by dividing the overall number of false identifications in culprit-absent lineups by the number of persons in the lineup (e.g.,[6,7]). However, this method places an upper limit on the rate of innocent-suspect identifications—the inverse of the size of the lineup—which may underestimate the real-world risk to innocent suspects because it is based on the assumption that the lineups are perfectly fair and biased selections cannot occur[20,21]. In the present experiments, culprit-absent lineups with a designated innocent suspect were used. This type of lineup composition is more ecologically valid than using culprit-absent lineups containing only fillers. In practice, the photograph of the suspect (whose guilt is unknown to the police) is often taken from a different source (e.g., from a mugshot or social media) than the photographs of the fillers (typically taken from databases) and may therefore differ to some extent from that of the fillers. Second, in some studies on lineup size, the number of persons in the lineup was confounded with the identity of the fillers such that certain fillers were only used in the larger lineups but not in the smaller lineups (for an overview, see[20]). In the present experiments, a random-selection procedure (see the Lineup procedures section below) guaranteed that the identity of the lineup fillers was not confounded with the lineup-size variable.

Among the few studies that have investigated the effects of lineup size on eyewitness decisions, some failed to find a significant difference in either the proportions of suspect identifications or the proportions of lineup rejections[22–26]. However, none of these studies had a sufficiently large sample size to detect an effect of lineup size on eyewitness decisions with a high sensitivity[27]. The results of studies using larger sample sizes indicate that there may be a trade-off between the aims of protecting the innocent suspects on the one side and of prosecuting the culprits on the other. Adding lineup fillers decreases the proportion of false innocent-suspect identifications but also decreases the proportion of culprit identifications[5–7]. However, these response rates may hide theoretically important effects as they confound different underlying processes. The increase in culprit identifications and innocent-suspect identifications in smaller compared to larger lineups may simply be a consequence of the fact that the sampling probability with which random guessing-based selection among the lineup members leads to the identification of the suspect is inversely related to lineup size. However, it is also possible to hypothesize that the increase in culprit identifications in smaller lineups may additionally be the result of enhanced culprit-presence detection. Previous research either found that lineup size did not affect the ability to discriminate between culprits and innocent suspects[6,7] or provided evidence for an increased ability to discriminate between culprits and innocent suspects in smaller compared to larger lineups[5]. Increased culprit-presence detection in smaller lineups is expected based on the assumption that "larger lineups may prove to be more difficult due to the increase in cognitive demands required by a larger lineup, that in turn may hinder identification accuracy" ([26], pp. 25–26). In fact, the mere processing of visual information has already been demonstrated to interfere with the accuracy of memories (e.g.,[28–30]). In the case of an increased number of filler faces in larger lineups, the additional faces represent more visual information that may require additional processing resources which may then be diverted from the to-be-detected culprit's face, resulting in a diminished ability to detect the culprit (cf.[28–30]). Here we use the 2-HT eyewitness identification model that represents the inverse relationship between the random-sampling probability and lineup size by a fixed constant—given by 1 ÷ lineup size—which is independent of the parameters

representing the detection-based and non-detection-based processes that have to be estimated from the data. Based on the literature available to date, it remains an open question as to whether culprit-presence detection, represented by the model's parameter *dP*, is increased in smaller compared to larger lineups or whether *dP* is independent of lineup size.

The second test of interest refers to guessing-based selection. Even though both culprit identifications and innocent-suspect identifications decrease with increasing lineup size—which may at first glance be taken as evidence for a decreased reliance on guessing—, a closer look at the other data categories that are available from the lineups reveals that lineup rejections decrease and filler identifications increase with increasing lineup size[20]. This pattern of results may be taken to suggest that the protection of suspects that is granted by large lineup sizes is only caused by the dispersion of guessing-based selections among the lineup members due to the inverse relationship to the random-sampling probability while the probability with which eyewitnesses make guessing-based selections may actually increase in larger lineups compared to smaller lineups. To test this hypothesis more directly, the model's guessing-based-selection parameter *g* was compared between the larger and the smaller lineups.

Experiment 1

Methods

Participants

Data were collected using the research panel of respondi AG based in Cologne, Germany (https://www.respondi.com) who compensated participants for their participation. Of the 1821 participants who initially filled out the socio-demographic questionnaire at the beginning of the experiment, 284 had to be excluded because they either failed to complete the experiment or withdrew the consent to use their data ($n = 234$), saw the staged-crime video more than once due to repeated participation ($n = 39$) or failed the attention check ($n = 11$; for an explanation, see the Materials and procedure section). The final sample, characterized by a diverse level of education, included the data of 1537 participants (668 female, 865 male, 4 diverse) aged between 18 and 88 years ($M = 46$, $SD = 16$). A sensitivity analysis with G*Power[31] showed that, given a sample size of $N = 1537$, four eyewitness decisions per participant and $\alpha = \beta = 0.05$, it was possible to detect even small effects of lineup size of size $w = 0.05$ on the model parameters across the lineup format conditions ($df = 2$). Participants were randomly assigned to one of the four lineup conditions: the three-person sequential lineup condition ($n = 382$), the six-person sequential lineup condition ($n = 393$), the three-person simultaneous lineup condition ($n = 380$) and the six-person simultaneous lineup condition ($n = 382$).

Ethics statement

In Experiments 1 and 2, participants gave informed consent prior to participation. The ethics committee of the Faculty of Mathematics and Natural Sciences at Heinrich Heine University granted approval for a series of experiments to which the present experiments belong. Both experiments were conducted in accordance with the Declaration of Helsinki. Participants were informed that they would see a video including elements of verbal and physical abuse. In case of any discomfort arising from watching such a video, participants were advised to withdraw from the study. At the end of the experiments, participants were debriefed that the crime had been staged for research purposes.

Materials and procedure

Materials and procedure were essentially as described by Winter et al.[12,17] and Menne et al.[18] with the exception of the lineup size manipulation described below. The experiment was conducted online. It was programmed with *Sosci Survey*[32] and was made available via https://www.soscisurvey.de. Participants were only allowed to participate with a desktop or laptop computer, not with a tablet or smartphone.

Staged-crime videos. Participants were shown one of two staged-crime videos (henceforth referred to as Video 1 and Video 2). In both videos, four hooligans of the German soccer club FC Bayern München (the culprits) attacked a fan of the rivaling soccer club Borussia Dortmund (the victim) at a bus station. The victim and the culprits wore fan clothing (caps, shirts and scarfs) showing their club's color and logo. The culprits insulted the victim, made fun of him and tossed his personal belongings around. As the video progressed, the culprits became more physically aggressive. They pushed the victim around until he got knocked to the ground. The four culprits carried on with their verbal and physical abuse of the victim until they suddenly became aware of the presence of another person approaching (not depicted in the videos). At that point, they swiftly ran away. Following the lead of other researchers (e.g.,[33–36]), our participants were shown a crime video including four culprits to increase the efficiency of data collection while maintaining ecological validity given that a substantial number of real-world crimes involve multiple culprits[37–39].

Both videos depicted the identical crime event, following the same sequence and timing, but the victims and the culprits were different actors. However, the actors were selected in such a way that the victim of Video 1 resembled the victim of Video 2 and that each of the four culprits of Video 1 had a high resemblance (as determined by the authors) to one of the four culprits of Video 2 in terms of body shape, hair color and hairstyle (i.e., Culprit A in Video 1 resembled Culprit A in Video 2, Culprit B in Video 1 resembled Culprit B in Video 2 and so on). Participants were randomly assigned to one of the two versions of the staged-crime video. The videos lasted about 130 s and were presented at a resolution of 885×500 pixels. The videos allowed participants to have a clear view on the culprits' faces.

Participants started the video by pressing the 'Start' button. They could not fast-forward, replay or pause the video and could only proceed to the next page after they had seen the whole video. To ensure that the video had

been attended, a 10-alternatives attention-check question was displayed afterwards, asking participants to select the option that correctly indicated what type of persons had been shown in the video (correct response: "soccer fans"). Participants who failed the attention check were excluded from further analyses.

Lineup procedures. Immediately after the attention-check question, participants were informed that they would see several lineups with a series of faces and that their task was to identify the FC Bayern München hooligans from the video. Participants were also told that the lineups may or may not contain a previously seen face to emphasize not only the need to identify the culprit when the culprit is present but also the need to reject the lineup when the culprit is absent.

Participants were then presented with four separate lineups in a randomized order, each corresponding to one of the four culprits in the video. Depending on the lineup size condition, the lineups consisted of the facial photographs of one suspect and of either two or five fillers. Two of the four lineups were culprit-present lineups, the other two were culprit-absent lineups. The crossed lineup procedure[12,18] was used to manipulate culprit presence. Two of the culprits of the video participants had seen were randomly selected to be presented in the two culprit-present lineups alongside the fillers. In the two culprit-absent lineups, the photographs of the suspects were photographs of the actors from the video participants had not seen. For instance, if participants had viewed Video 1, then two randomly selected culprits from Video 1 (e.g., Culprit B and Culprit D) served as the culprits in the culprit-present lineups and two culprits from Video 2 (in this example, Culprit A and Culprit C) served as the innocent suspects in the culprit-absent lineups. This lineup construction procedure helps to ensure that the photographs of the culprits and of the innocent suspects differs to the same degree from the photographs of the fillers. This represents an ecological valid procedure. In the real world, the photograph of the suspect (whose status of being innocent or guilty is unknown to the police) is often taken from a different source (e.g., from a mugshot or social media) than the photographs of the fillers, which are typically obtained from face databases.

For each lineup, five male filler faces aged between 18 and 29 years were taken from the Center for Vital Longevity Face Database[40]. Fillers were chosen based on their resemblance (as determined by the authors) to the culprits in terms of body shape, hair color and hairstyle. In the three-person lineup conditions, two fillers were randomly selected for each participant from the pool of five fillers of each lineup. The randomization ensured that the critical difference between the lineup size conditions was the number of fillers and not the filler's identity. The positions of the suspect and filler photographs were randomized in each lineup. The photographs of the suspects and the fillers showed a front view of face and neck against a black background. All faces had neutral facial expressions. All photographs were matched for lighting and face size and were displayed at a resolution of 142×214 pixels. Examples of the lineups used in Experiments 1 and 2 can be found at https://osf.io/ckdbr/.

Depending on the lineup format condition, the lineups were presented either simultaneously or sequentially. In simultaneous lineups, the photographs of the suspect and the fillers were shown together in one row. Participants could identify a person as the culprit by pressing the "Yes, was present" button underneath the photograph of this person or they could reject the lineup by pressing the "No, none of these persons was present" button located to the right of each lineup. After their decision, participants were asked to indicate their confidence in their response to make the procedure similar to that of a real police lineup. The participants could then initiate the presentation of the subsequent lineup by pressing the "Next" button. In sequential lineups, the photographs of the suspect and the fillers were presented, one at a time, in random succession. For each lineup member, participants decided whether or not the depicted person was one of the culprits of the video they had seen by pressing the "Yes, was present" button underneath the photograph or the "No, this person was not present" button located at the right side of the photograph. Participants were required to make a decision before they could proceed to the next photograph. After making each decision, participants were asked to indicate their confidence in their response. If participants identified more than one face in a lineup, we followed the standard police procedure in Germany and the United States[10,41,42], considering the last identification as a revision of any prior identification. Therefore, the last identification was used in the present analysis. If participants identified none of the persons in the lineup as one of the culprits, the lineup was counted as rejected. After participants had made their decision in the fourth lineup, they were debriefed and thanked for their time.

Results

For all analyses reported in this article, parameter estimates and goodness-of-fit tests were calculated using *multiTree*[16]. The α level was set to 0.05. The observed response frequencies and proportions for Experiments 1 and 2 are shown in Table 1.

To analyze the results of Experiment 1, we needed four instances of the model illustrated in Fig. 1, one for the three-person sequential lineups, one for the six-person sequential lineups, one for the three-person simultaneous lineups and one for the six-person simultaneous lineups. In the three-person lineup conditions, the constant $1 \div$ lineup size was set to 0.33333 to approximate the random-sampling probability of 1/3. In the six-person lineup conditions, $1 \div$ lineup size was set to 0.16667 to approximate the random-sampling probability of 1/6. Our aim was to use a base model that was as simple as possible. Therefore, we imposed parameter restrictions onto the 2-HT eyewitness identification model that were the same as those used in prior applications of this model[11,12,17]. First, given that the same sets of suspects and fillers were used in all conditions, lineup fairness must necessarily be the same across the four conditions. Therefore, the biased-suspect-selection parameter b was set to be equal across the four lineup conditions. Second, there was no a priori reason to assume that the ability to detect the absence of the culprit should differ among the conditions (for manipulations that can be expected to influence the ability to detect the absence of the culprit, see[11,12]). Therefore, the culprit-absence-detection parameter dA was set to be equal across the four conditions. The base model incorporating these restrictions fit

Lineup format	Lineup size	Culprit-present lineups			Culprit-absent lineups		
		Culprit identifications	Filler identifications	Lineup rejections	Innocent-suspect identifications	Filler identifications	Lineup rejections
Experiment 1							
Sequential	Three persons	386 (0.51)	205 (0.27)	173 (0.23)	206 (0.27)	241 (0.32)	317 (0.41)
	Six persons	273 (0.35)	339 (0.43)	174 (0.22)	127 (0.16)	374 (0.48)	285 (0.36)
Simultaneous	Three persons	371 (0.49)	120 (0.16)	269 (0.35)	139 (0.18)	184 (0.24)	437 (0.58)
	Six persons	283 (0.37)	191 (0.25)	290 (0.38)	87 (0.11)	249 (0.33)	428 (0.56)
Experiment 2							
Sequential	Two persons	499 (0.61)	124 (0.15)	193 (0.24)	259 (0.32)	161 (0.20)	396 (0.49)
	Five persons	320 (0.40)	316 (0.39)	166 (0.21)	143 (0.18)	378 (0.47)	281 (0.35)
Simultaneous	Two persons	433 (0.55)	72 (0.09)	281 (0.36)	204 (0.26)	98 (0.12)	484 (0.62)
	Five persons	320 (0.41)	172 (0.22)	296 (0.38)	115 (0.15)	232 (0.29)	441 (0.56)

Table 1. Response frequencies and proportions (in parentheses) as a function of lineup format and lineup size observed in Experiments 1 and 2. The proportions are rounded to two decimal places and thus do not always add up to 1.

the data, $G^2(6) = 8.39$, $p = 0.211$. The estimates of parameters b and dA were 0.07 ($SE = 0.01$) and 0.11 ($SE = 0.02$), respectively. The estimates of parameters dP and g are displayed in Fig. 2.

One advantage of multinomial processing tree models is that hypothesis tests can easily be performed directly at the level of the postulated processes (for details, see the tutorial by[15]). For instance, the hypothesis that culprit-presence detection differs between the three-person lineups and the six-person lineups can be tested by setting the culprit-presence-detection parameter dP to be equal between these conditions. If the fit of the model including this restriction is significantly worse than the fit of the base model not including this restriction, then it can be concluded that culprit-presence detection differs between conditions in the direction indicated by the parameter estimates. Participants in the three-person lineup conditions were significantly more likely to detect the culprit than participants in the six-person lineup conditions, $\Delta G^2(2) = 22.17$, $p < 0.001$. In addition, the probability of guessing-based selection was significantly higher when the lineups contained six persons than when the lineups contained three persons, $\Delta G^2(2) = 10.86$, $p = 0.004$.

Discussion

The results of Experiment 1 support the hypothesis that culprit-presence detection is better in smaller compared to larger lineups. Participants were more likely to identify the culprit in three-person lineups than in six-person lineups. In addition, the model-based analysis demonstrated that participants were more likely to select a lineup member based on guessing in six-person lineups than in three-person lineups, thus providing direct evidence that increasing the lineup size causes increased guessing-based selection. The consequences of these guessing-based selections in terms of whether the suspect or a filler is selected is then given by the random-sampling probability, which is a fixed constant in the analysis.

Given the debate about the need for replication[43], it seemed desirable to test the reliability and robustness of these findings before drawing firm conclusions. Therefore, Experiment 2 served as a conceptual replication of Experiment 1, the only difference to Experiment 1 being that participants were presented with lineups consisting of either two or five persons.

Experiment 2

Method

Participants

Participants were recruited and compensated as in Experiment 1. Of the 1851 participants who initially filled out the socio-demographic questionnaire at the beginning of the experiment, 255 had to be excluded because they either failed to complete the experiment or withdrew the consent to use their data ($n = 217$), saw the staged-crime video more than once due to repeated participation ($n = 19$) or failed the attention check ($n = 19$). The final sample, characterized by a diverse level of education, included the data of 1596 participants (717 female, 872 male, 7 diverse) aged between 18 and 74 years ($M = 46$, $SD = 16$). A sensitivity analysis with G*Power[31] showed that, given a sample size of $N = 1596$, four eyewitness decisions per participant and $\alpha = \beta = 0.05$, it was possible to detect even small effects of lineup size of size $w = 0.05$ on the model parameters across the lineup format conditions ($df = 2$). Participants were randomly assigned to one of the four lineup conditions: the two-person sequential lineup condition ($n = 408$), the five-person sequential lineup condition ($n = 401$), the two-person simultaneous lineup condition ($n = 393$) and the five-person simultaneous lineup condition ($n = 394$).

Materials and procedure

The materials and procedure were the same as those used in Experiment 1 with the following exception. Instead of presenting three-person or six-person lineups, the participants were randomly assigned to view two-person or

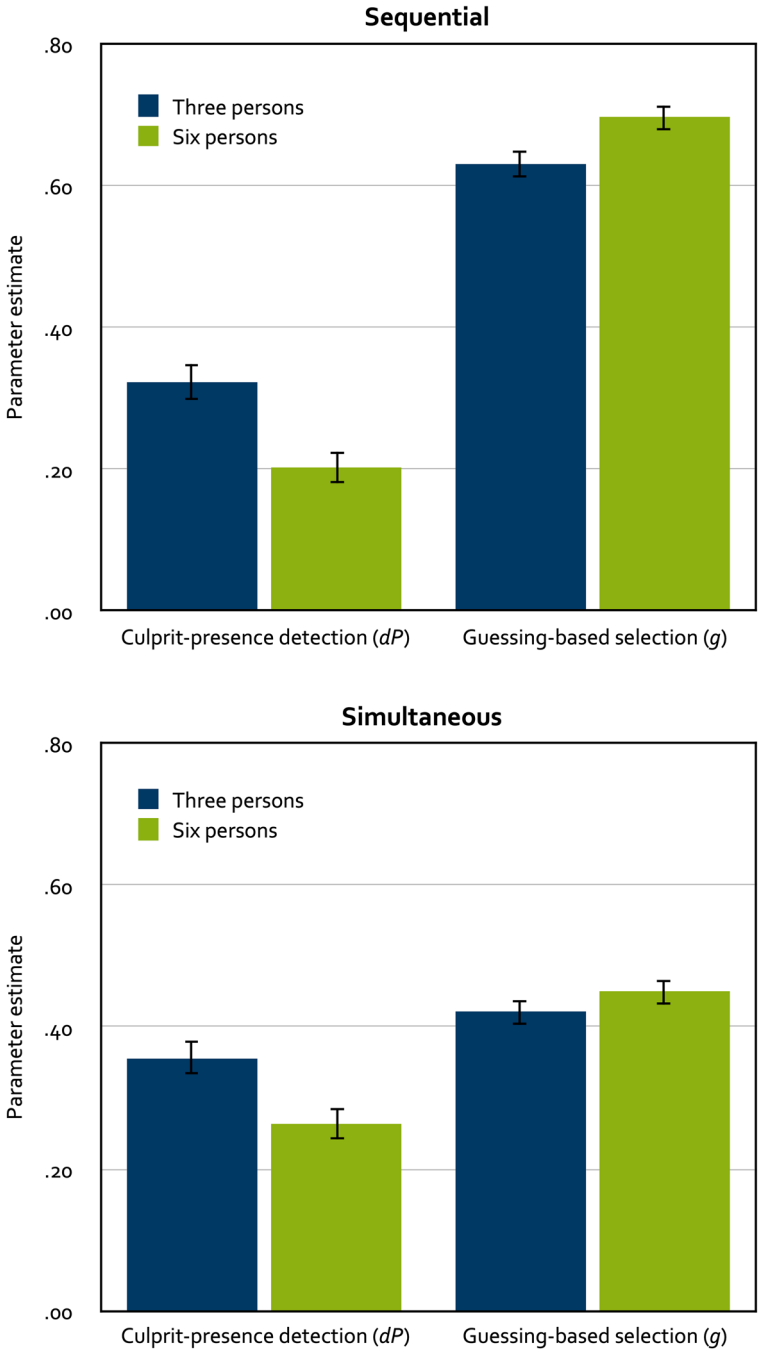

Figure 2. Parameter estimates of parameter *dP* reflecting the probability of detecting the presence of the culprit (left panels) and parameter *g* reflecting the probability of guessing-based selection among the lineup members (right panels) as a function of lineup size for sequential lineups (top panels) and simultaneous lineups (bottom panels) in Experiment 1. The error bars represent standard errors.

five-person lineups consisting of the facial photographs of the suspect and either one or four fillers, respectively. Parallel to Experiment 1, the fillers were randomly selected from the pool of five fillers of each lineup.

Results

To analyze the results of Experiment 2, we needed four instances of the model depicted in Fig. 1, one for the two-person sequential lineups, one for the five-person sequential lineups, one for the two-person simultaneous lineups and one for the five-person simultaneous lineups. In the two-person lineup conditions, the random-sampling probability is given by 1 ÷ lineup size and was thus set to 0.5. In the five-person lineup conditions, the random-sampling probability is given by 1 ÷ lineup size and was thus set to 0.2. The same assumptions as in Experiment 1 were used to arrive at the base model. Specifically, the biased-suspect-selection parameter *b* and the

culprit-absence-detection parameter *dA* were each set to be equal across the four lineup conditions. The model incorporating these restrictions fit the data, $G^2(6) = 12.14$, $p = 0.059$. The estimates of parameters *b* and *dA* were 0.10 ($SE = 0.01$) and 0.11 ($SE = 0.02$), respectively. The estimates of parameters *dP* and *g* are displayed in Fig. 3.

Participants in the two-person lineup conditions were significantly more likely to detect the culprit than participants in the five-person lineup conditions, $\Delta G^2(2) = 38.72$, $p < 0.001$. In addition, the probability of guessing-based selection was significantly higher when the lineups contained five persons than when the lineups contained two persons, $\Delta G^2(2) = 62.98$, $p < 0.001$.

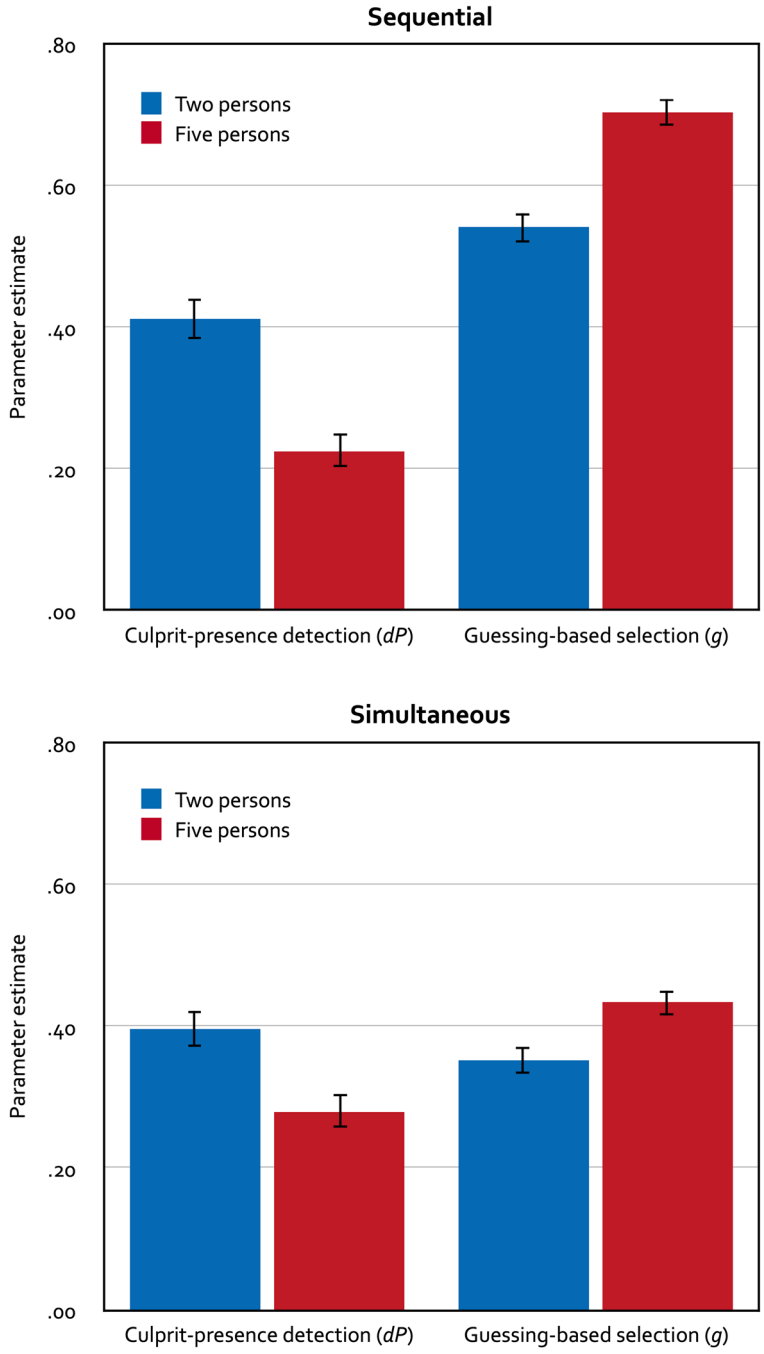

Figure 3. Parameter estimates of parameter *dP* reflecting the probability of detecting the presence of the culprit (left panels) and parameter *g* reflecting the probability of guessing-based selection among the lineup members (right panels) as a function of lineup size for sequential lineups (top panels) and simultaneous lineups (bottom panels) in Experiment 2. The error bars represent standard errors.

Discussion

The aim of Experiment 2 was to determine whether the main findings of Experiment 1 could be replicated using lineups consisting of either two or five persons. As in Experiment 1, culprit-presence detection was better in smaller compared to larger lineups and increasing the lineup size caused increased guessing-based selection among the lineup members, thus supporting the robustness of those findings.

General discussion

In many jurisdictions, larger minimum lineup sizes are recommended[20,44] presumably because large lineups seem to be better than small lineups at protecting the suspect from being selected based on guessing. This protective effect may already result from the decreased sampling probability when eyewitnesses randomly select a person from a larger lineup. However, it is unclear whether having to consider a larger number of faces in a lineup may have additional effects on detection-based and non-detection-based processes. The aim of the present study was to apply the 2-HT eyewitness identification model to identify the effects of lineup size on the latent processes underlying eyewitness decisions. In Experiment 1, participants were presented with simultaneous or sequential lineups consisting of either three or six persons. In Experiment 2, participants were presented with simultaneous or sequential lineups consisting of either two or five persons. Taken together, the data of Experiments 1 and 2 provide a coherent picture by showing a decrease of culprit-presence detection and an increase in guessing-based selections with increasing lineup size.

The model-based results thus demonstrate that the increased culprit-identification rates in the smaller compared to the larger lineups, as also previously found in other studies on lineup size[5–7], cannot solely be attributed to the increased sampling probability with which random guessing-based selection among the lineup members leads to suspect identifications in smaller lineups. It is also a consequence of improved culprit-presence detection in smaller compared to larger lineups. Here we can only speculate as to why this is the case. It is well established that the processing of visual information can interfere with memory of visual details (e.g.,[28–30]). With an increasing number of fillers in a lineup, a larger portion of resources may be devoted to the processing of the visual details of more and more fillers' faces, leaving fewer and fewer resources for processing the visual details of the suspect's face, resulting in a diminished ability to detect the culprit. Whatever the reason, the higher probability of culprit-presence detection is an interesting feature of smaller compared to larger lineups. As a side note, in the present experiments participants saw not one but four lineups and were thus exposed to identification conditions which may perhaps have been even more resource-demanding than the identification conditions in studies with only one lineup. However, the effect of lineup size on culprit-presence detection is clearly present even when only the data of the first of the four lineups are analyzed. This is the case for the data from both Experiments 1 and 2 (the data and equation files of these analyses are available at https://osf.io/ckdbr/). This suggests that the effect of lineup size is robust regardless of whether one or more lineups are used.

An interesting implication of the present analysis is that the effects of lineup size on the non-detection-based processes can be dissected into different components. Larger lineups in comparison to smaller lineups drive down the probability with which the suspect is selected given that guessing-based selection occurs (i.e., $1 \div$ lineup size, see Fig. 1). This property of larger lineups is quite obvious. Analogously, when a lecturer prepares an exam with multiple-alternative forced-choice items, more instead of fewer response options seem desirable so that students have a lower probability of selecting the correct option when making guessing-based selections. To illustrate this point using the present Experiment 1, if eyewitnesses select a person from the lineup based on guessing, which occurs with the conditional probability g, then the conditional probability with which the suspect is sampled is 1/3 in a three-person lineup, but only 1/6 in a six-person lineup. The 2-HT eyewitness identification model represents this simple mathematical consequence of sampling from more options explicitly in the sampling probability given by the constant $1 \div$ lineup size. The probability with which the suspect is selected based on guessing is $g \cdot (1 \div \text{lineup size})$. To illustrate, the probability with which the suspect is selected based on guessing is $g \cdot 1/3$ in a three-person lineup and $g \cdot 1/6$ in a six-person lineup. However, including more fillers also affects the probability of guessing-based selection, represented by the model parameter g. In a lineup, each face provides an opportunity for a guessing-based selection. Larger lineups provide more opportunities for guessing-based selections than smaller lineups. For instance, in a three-person lineup, the eyewitness has three opportunities to make a guessing-based selection, whereas in a six-person lineup, the eyewitness has six opportunities to make a guessing-based selection. It is thus plausible that the overall probability of making a guessing-based selection in a lineup (parameter g) increases with increasing lineup size, which is indeed what was found in the present model-based analysis. Just as in the eyewitness situation, including a larger number of plausible response options in multiple-alternative forced-choice questions provides more opportunities for guessing a false option and, as a result, impairs performance by increasing the number of false answers and by decreasing the number of correct answers (e.g.,[45,46]). Similarly, at the level of raw response rates, larger lineups in the present experiments were associated with an increased rate of filler identifications and a decreased rate of rejections of culprit-absent lineups (see Table 1). This result is in line with the findings of Juncu and Fitzgerald[20] who meta-analytically reviewed the body of lineup size research. They found an increase in filler identifications and a decrease in lineup rejections when lineup size increased, which already suggest at the level of raw response rates that increasing the lineup size enhances guessing-based selection.

To summarize, larger lineups provide more opportunities for guessing-based selection than smaller lineups, as a consequence of which guessing-based selections become more likely (parameter g is larger for larger lineups than for smaller lineups). However, if guessing-based selection occurs, then the selections are dispersed over a larger number of persons, including fillers that are known to be innocent (the sampling probability given by the constant $1 \div$ lineup size is smaller for larger lineups than for smaller lineups). Given that the probability with which the suspect is selected based on guessing is given by $g \cdot (1 \div \text{lineup size})$, the question is whether the

increase in guessing-based selection, as represented by parameter *g*, is compensated or even overcompensated by a decrease of the sampling probability given by the constant 1 ÷ lineup size. In both Experiments 1 and 2, larger lineups were associated with increased guessing-based selection compared to smaller lineups, but the decrease in the sampling probability in larger lineups overcompensated this effect. As a result, larger lineups were associated with a decreased rate of suspect identifications at the level of the raw response rates in both experiments (see Table 1). The increased protection of the suspect that is seen as the main advantage of larger lineups compared to smaller lineups was thus replicated in the present experiments. This result is consistent with findings of previous research on lineup size (e.g.,[5–7,20]).

In contrast to most prior work, we included both simultaneous and sequential lineups in the present experiments. Given that in some countries such as the United Kingdom or Germany, the sequential format is the standard way for presenting police lineups[10,19], it is important to investigate the effects of lineup size in both lineup formats. The present results suggest that increasing the size of the lineups decreases culprit-presence detection and increases guessing-based selection in both simultaneous and sequential lineups. Moving beyond the main research question, it seems noteworthy that culprit-presence detection did not differ significantly between the simultaneous and the sequential lineup conditions in both experiments (Experiment 1: $\Delta G^2(2) = 5.47$, $p = 0.065$; Experiment 2, $\Delta G^2(2) = 3.24$, $p = 0.198$) although culprit-presence detection was descriptively better for simultaneous than for sequential lineups in all conditions except in the two-person lineup condition in Experiment 2 (see Figs. 2 and 3). It may seem more surprising, at least at first glance, that guessing-based selection was significantly more likely in the sequential lineup conditions than in the simultaneous lineup conditions in both Experiment 1, $\Delta G^2(2) = 209.42$, $p < 0.001$, and Experiment 2, $\Delta G^2(2) = 210.16$, $p < 0.001$. This finding is in line with previous model-based analyses[12,18] but contradicts the common assumption that sequential lineups induce more conservative responding than simultaneous lineups[47]. Here it seems relevant that, in contrast to most previous research, participants in the sequential lineup conditions in the present experiments were not informed that only their first positive response would count. Instead, the presentation of the sequential lineups continued after a positive identification had been made; only the participant's last identification was counted as an identification decision (if there was no identification decision, then the lineup was counted as rejected). This sequential lineup protocol was used here because it corresponds to standard police protocols[9,10,41,48] and to the original sequential lineup protocol outlined by Lindsay and Wells[49]. Horry et al.[42] have demonstrated that a lineup is more likely to be rejected and the suspect is less likely to be selected when a first-yes-counts protocol is used, strongly suggesting that a first-yes-counts protocol discourages participants from guessing. This finding is to be expected because participants will be reluctant to use their only positive response too early in the sequence, not knowing whether better alternatives might be presented later in the sequence. The more ecologically valid sequential lineup protocol applied here gives participants the opportunity to change their mind by allowing them to select a lineup member later in the sequence even though another lineup member has already been selected earlier. In fact, there is direct evidence that the probability of guessing-based selection is significantly less likely when first-yes-counts instructions are used than when the standard police protocol is used[17].

Conclusion

The purpose of the present study is to contribute to a better understanding of lineup size effects on eyewitness decisions using the 2-HT eyewitness identification model. By taking into account the full range of data categories observed in typical lineup procedures (i.e., suspect identifications, filler identifications and lineup rejections in both culprit-present and culprit-absent lineups), the model provides new insights into how lineup size affects the latent detection-based and non-detection-based processes underlying eyewitness decisions. The results demonstrate that, compared to smaller lineups, larger lineups are associated with a decreased ability to detect the culprit in the lineup and an increased probability of selecting a lineup member based on guessing. However, the increase in guessing-based selection in larger lineups is overcompensated by a reduced probability of randomly selecting the suspect among the lineup members if guessing-based selection occurs. As a consequence, the rate of innocent-suspect identifications decreases with increasing lineup size, albeit to a lesser degree than the rate of culprit identifications. In line with previous research[20], the present results indicate that there is a trade-off between the aims of providing optimal conditions for detecting the culprit on the one side and protecting the suspect from the consequences of guessing-based selections on the other. Specifically, under the circumstances realized in the present experiments, larger lineups were associated with a disadvantage in culprit-presence detection and guessing-based selection while their main advantage may be seen in a broader dispersion of those guessing-based selections among the lineup members.

Data availability

The data and equation files of all analyses have been made publicly available at the Open Science Framework and can be accessed at https://osf.io/ckdbr/.

Received: 17 May 2023; Accepted: 3 October 2023
Published online: 11 October 2023

References

1. Innocence Project. *Exonerate the Innocent.* https://innocenceproject.org/exonerate/ (2023). Accessed 1 Aug 2023.
2. Mickes, L., Flowe, H. D. & Wixted, J. T. Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *J. Exp. Psychol. Appl.* **18**, 361–376. https://doi.org/10.1037/a0030609 (2012).
3. Malpass, R. S. & Devine, P. G. Eyewitness identification: Lineup instructions and the absence of the offender. *J. Appl. Psychol.* **66**, 482–489. https://doi.org/10.1037/0021-9010.66.4.482 (1981).

4. Wells, G. L., Rydell, S. M. & Seelau, E. P. The selection of distractors for eyewitness lineups. *J. Appl. Psychol.* **78**, 835–844. https://doi.org/10.1037/0021-9010.78.5.835 (1993).
5. Meissner, C. A., Tredoux, C. G., Parker, J. F. & MacLin, O. H. Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Mem. Cognit.* **33**, 783–792. https://doi.org/10.3758/BF03193074 (2005).
6. Akan, M., Robinson, M. M., Mickes, L., Wixted, J. T. & Benjamin, A. S. The effect of lineup size on eyewitness identification. *J. Exp. Psychol. Appl.* **27**, 369–392. https://doi.org/10.1037/xap0000340 (2020).
7. Wooten, A. R. *et al.* The number of fillers may not matter as long as they all match the description: The effect of simultaneous lineup size on eyewitness identification. *Appl. Cogn. Psychol.* **34**, 590–604. https://doi.org/10.1002/acp.3644 (2020).
8. Police Executive Research Forum. *A National Survey of Eyewitness Identification Procedures in Law Enforcement Agencies.* https://www.policeforum.org/assets/docs/Free_Online_Documents/Eyewitness_Identification/a%20national%20survey%20of%20eyewitness%20identification%20procedures%20in%20law%20enforcement%20agencies%202013.pdf (2013). Accessed 1 Aug 2023.
9. Home Office. *Police and Criminal Evidence Act 1984 (PACE) Code D.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/903812/pace-code-d-2017.pdf (2017). Accessed 31 July 2023.
10. German Federal Ministry of the Interior and Community. *Richtlinien für das Strafverfahren und das Bußgeldverfahren (RiStBV) [Guidelines for criminal proceedings and fine proceedings].* https://www.verwaltungsvorschriften-im-internet.de/bsvwvbund_28032023_BMJRB3313104000060001.htm (2023). Accessed 31 July 2023.
11. Menne, N. M., Winter, K., Bell, R. & Buchner, A. A validation of the two-high threshold eyewitness identification model by reanalyzing published data. *Sci. Rep.* **12**, 13379. https://doi.org/10.1038/s41598-022-17400-y (2022).
12. Winter, K., Menne, N. M., Bell, R. & Buchner, A. Experimental validation of a multinomial processing tree model for analyzing eyewitness identification decisions. *Sci. Rep.* **12**, 15571. https://doi.org/10.1038/s41598-022-19513-w (2022).
13. Batchelder, W. H. & Riefer, D. M. Theoretical and empirical review of multinomial process tree modeling. *Psychon. Bull. Rev.* **6**, 57–86. https://doi.org/10.3758/BF03210812 (1999).
14. Erdfelder, E. *et al.* Multinomial processing tree models: A review of the literature. *Z. Psychol./J. Psychol.* **217**, 108–124. https://doi.org/10.1027/0044-3409.217.3.108 (2009).
15. Schmidt, O., Erdfelder, E. & Heck, D. W. Tutorial on multinomial processing tree modeling: How to develop, test, and extend MPT models. *Psychol. Methods.* https://doi.org/10.1037/met0000561 (2023).
16. Moshagen, M. multiTree: A computer program for the analysis of multinomial processing tree models. *Behav. Res. Methods* **42**, 42–54. https://doi.org/10.3758/BRM.42.1.42 (2010).
17. Winter, K., Menne, N. M., Bell, R. & Buchner, A. Evaluating the impact of first-yes-counts instructions on eyewitness performance using the two-high threshold eyewitness identification model. *Sci. Rep.* **13**, 6572. https://doi.org/10.1038/s41598-023-33424-4 (2023).
18. Menne, N. M., Winter, K., Bell, R. & Buchner, A. Measuring lineup fairness from eyewitness identification data using a multinomial processing tree model. *Sci. Rep.* **13**, 6290. https://doi.org/10.1038/s41598-023-33101-6 (2023).
19. Seale-Carlisle, T. M. & Mickes, L. US line-ups outperform UK line-ups. *R. Soc. Open Sci.* **3**, 160300. https://doi.org/10.1098/rsos.160300 (2016).
20. Juncu, S. & Fitzgerald, R. J. A meta-analysis of lineup size effects on eyewitness identification. *Psychol. Public Policy Law* **27**, 295–315. https://doi.org/10.1037/law0000311 (2021).
21. Clark, S. E., Howell, R. T. & Davey, S. L. Regularities in eyewitness identification. *Law Hum. Behav.* **32**, 187–218. https://doi.org/10.1007/s10979-006-9082-4 (2008).
22. Cutler, B. L., Penrod, S. D. & Martens, T. K. Improving the reliability of eyewitness identification: Putting context into context. *J. Appl. Psychol.* **72**, 629–637. https://doi.org/10.1037/0021-9010.72.4.629 (1987).
23. Nosworthy, G. J. & Lindsay, R. Does nominal lineup size matter? *J. Appl. Psychol.* **75**, 358–361. https://doi.org/10.1037/0021-9010.75.3.358 (1990).
24. Levi, A. M. An analysis of multiple choices in MSL lineups, and a comparison with simultaneous and sequential ones. *Psychol. Crime Law* **12**, 273–285. https://doi.org/10.1080/10683160500238782 (2006).
25. Levi, A. M. Research note: Evidence for moving to an 84-person photo lineup. *J. Exp. Criminol.* **3**, 377–391. https://doi.org/10.1007/s11292-007-9042-0 (2007).
26. Pozzulo, J. D., Dempsey, J. L. & Wells, K. Does lineup size matter with child witnesses. *J. Police Crim. Psychol.* **25**, 22–26. https://doi.org/10.1007/s11896-009-9055-x (2010).
27. Seale-Carlisle, T. M., Wetmore, S. A., Flowe, H. D. & Mickes, L. Designing police lineups to maximize memory performance. *J. Exp. Psychol. Appl.* **25**, 410–430. https://doi.org/10.1037/xap0000222 (2019).
28. Wais, P. E., Rubens, M. T., Boccanfuso, J. & Gazzaley, A. Neural mechanisms underlying the impact of visual distraction on retrieval of long-term memory. *J. Neurosci.* **30**, 8541–8550. https://doi.org/10.1523/JNEUROSCI.1478-10.2010 (2010).
29. Vredeveldt, A., Hitch, G. J. & Baddeley, A. D. Eyeclosure helps memory by reducing cognitive load and enhancing visualisation. *Mem. Cognit.* **39**, 1253–1263. https://doi.org/10.3758/s13421-011-0098-8 (2011).
30. Vredeveldt, A. *et al.* A field evaluation of the Eye-Closure Interview with witnesses of serious crimes. *Law Hum. Behav.* **39**, 189–197. https://doi.org/10.1037/lhb0000113 (2015).
31. Faul, F., Erdfelder, E., Lang, A. & Buchner, A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191. https://doi.org/10.3758/bf03193146 (2007).
32. Leiner, D. J. *SoSci Survey [Computer software].* https://www.soscisurvey.de (2021).
33. Brigham, J. C. Target person distinctiveness and attractiveness as moderator variables in the confidence-accuracy relationship in eyewitness identifications. *Basic Appl. Soc. Psych.* **11**, 101–115. https://doi.org/10.1207/s15324834basp1101_7 (1990).
34. Wilcock, R. & Bull, R. Novel lineup methods for improving the performance of older eyewitnesses. *Appl. Cogn. Psychol.* **24**, 718–736. https://doi.org/10.1002/acp.1582 (2010).
35. Dobolyi, D. G. & Dodson, C. S. Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. *J. Exp. Psychol. Appl.* **19**, 345–357. https://doi.org/10.1037/a0034596 (2013).
36. Palmer, M. A., Brewer, N., Weber, N. & Nagesh, A. The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *J. Exp. Psychol. Appl.* **19**, 55–71. https://doi.org/10.1037/a0031602 (2013).
37. Hobson, Z. J. & Wilcock, R. Eyewitness identification of multiple perpetrators. *Int. J. Police Sci. Manag.* **13**, 286–296. https://doi.org/10.1350/ijps.2011.13.4.253 (2011).
38. Hobson, Z., Wilcock, R. & Valentine, T. Multiple suspect showing: A survey of police identification officers. *Policing* **7**, 79–87. https://doi.org/10.1093/police/pas021 (2012).
39. Tupper, N., Sauerland, M., Sauer, J. D. & Hope, L. Eyewitness identification procedures for multiple perpetrator crimes: A survey of police in Sweden, Belgium, and the Netherlands. *Psychol. Crime Law* **25**, 992–1007. https://doi.org/10.1080/1068316X.2019.1611828 (2019).
40. Minear, M. & Park, D. C. A lifespan database of adult facial stimuli. *Behav. Res. Methods Instrum. Comput.* **36**, 630–633. https://doi.org/10.3758/BF03206543 (2004).
41. Wells, G. L., Steblay, N. K. & Dysart, J. E. Double-blind photo lineups using actual eyewitnesses: An experimental test of a sequential versus simultaneous lineup procedure. *Law Hum. Behav.* **39**, 1–14. https://doi.org/10.1037/lhb0000096 (2015).

42. Horry, R., Fitzgerald, R. J. & Mansour, J. K. "Only your first yes will count": The impact of prelineup instructions on sequential lineup decisions. *J. Exp. Psychol. Appl.* **27**, 170–186. https://doi.org/10.1037/xap0000337 (2021).
43. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716. https://doi.org/10.1126/science.aac4716 (2015).
44. Fitzgerald, R. J., Rubínová, E. & Juncu, S. Eyewitness identification around the world. In *Methods, Measures, and Theories in Eyewitness Identification Tasks* (eds Smith, A. M. *et al.*) 294–322. https://doi.org/10.4324/9781003138105 (Taylor and Francis, 2021).
45. Roediger, H. L. & Marsh, E. J. The positive and negative consequences of multiple-choice testing. *J. Exp. Psychol. Learn. Mem. Cogn.* **31**, 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155 (2005).
46. Butler, A. C. & Roediger, H. L. Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Mem. Cognit.* **36**, 604–616. https://doi.org/10.3758/MC.36.3.604 (2008).
47. Clark, S. E. Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspect. Psychol. Sci.* **7**, 238–259. https://doi.org/10.1177/1745691612439584 (2012).
48. State Bar of Michigan: Eyewitness Identification Task Force. *Law Enforcement and Eyewitness Identifications: A policy writing guide.* https://www.michigan.gov/-/media/Project/Websites/mcoles/TD-Webpage/state_bar_eyewitness_law_enforcement_policy_writing_guide.pdf?rev=6e19807bd78447d2b54e6f9cc7ab186b (2015). Accessed 31 July 2023.
49. Lindsay, R. & Wells, G. L. Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *J. Appl. Psychol.* **70**, 556–564. https://doi.org/10.1037/0021-9010.70.3.556 (1985).

Author contributions

N.M.M., K.W., R.B. and A.B. contributed to the study conception, design, material preparation and data analysis. N.M.M collected the data and wrote the first draft of the manuscript with subsequent input from all co-authors. All authors gave final approval for publication.

Funding

Open Access funding enabled and organized by Projekt DEAL. The work reported herein was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—BU 945/10-1, project number 456214986.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to N.M.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Erklärung über den Eigenanteil an den in der Dissertation enthaltenen Einzelarbeiten

Meine Dissertationsschrift umfasst drei Fachartikel mit insgesamt acht Reanalysen und sechs Experimenten. Im Folgenden ist für jeden der Fachartikel aufgeführt, welche Autor:innen bei der Planung der Experimente und Reanalysen, bei der Umsetzung der Experimente, bei der Datenauswertung und beim Verfassen der Manuskripte mitgearbeitet haben. Der überwiegende Teil der Arbeit lag jeweils bei der Erstautorin des Artikels.

Menne, N. M., Winter, K., Bell, R., & Buchner, A. (2022). A validation of the two-high threshold eyewitness identification model by reanalyzing published data. *Scientific Reports*, *12*, 13379. https://doi.org/10.1038/s41598-022-17400-y

Planung: Menne, N. M, Winter, K., Bell, R., Buchner, A.
Auswertung: Menne, N. M, Winter, K., Bell, R., Buchner, A.
Manuskript: Menne, N. M, Winter, K., Bell, R., Buchner, A.

Menne, N. M., Winter, K., Bell, R., & Buchner, A. (2023). Measuring lineup fairness from eyewitness identification data using a multinomial processing tree model. *Scientific Reports*, *13*, 6290. https://doi.org/10.1038/s41598-023-33101-6

Planung: Menne, N. M, Winter, K., Bell, R., Buchner, A.
Umsetzung: Menne, N. M, Winter, K., Bell, R., Buchner, A.
Auswertung: Menne, N. M, Winter, K., Bell, R., Buchner, A.
Manuskript: Menne, N. M, Winter, K., Bell, R., Buchner, A.

Menne, N. M., Winter, K., Bell, R., & Buchner, A. (2023). The effects of lineup size on the processes underlying eyewitness decisions. *Scientific Reports*, *13*, 17190. https://doi.org/10.1038/s41598-023-44003-y

Planung: Menne, N. M, Winter, K., Bell, R., Buchner, A.
Umsetzung: Menne, N. M, Winter, K., Bell, R., Buchner, A.
Auswertung: Menne, N. M, Winter, K., Bell, R., Buchner, A.
Manuskript: Menne, N. M, Winter, K., Bell, R., Buchner, A.

Erklärung an Eides statt

Hiermit versichere ich an Eides statt, dass ich die Dissertation mit dem Titel „Entscheidungen von Augenzeug:innen in Gegenüberstellungen: Validierung und Anwendung eines multinomialen Modells zur Erfassung der zugrundeliegenden Prozesse“ selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt habe.

Ich versichere insbesondere:

(1) Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt.

(2) Alle wörtlich oder dem Sinn nach aus anderen Texten entnommenen Stellen habe ich als solche kenntlich gemacht; dies gilt für gedruckte Texte ebenso wie für elektronische Ressourcen.

(3) Die Arbeit habe ich in der vorliegenden oder einer modifizierten Form noch nicht als Dissertation vorgelegt – sei es an der Heinrich-Heine-Universität oder an einer anderen Universität.

Datum: 07. Februar 2024

Name: Nicola Marie Menne

Unterschrift: