



# Disentangling Different Aspects of Between-Item Similarity Unveils Evidence Against the Ensemble Model of Lineup Memory

Constantin G. Meyer-Grant<sup>1</sup> · Karl Christoph Klauer<sup>1</sup>

Accepted: 23 March 2022  
© The Author(s) 2022

## Abstract

For modeling recognition decisions in a typical eyewitness identification lineup task with multiple simultaneously presented test stimuli (also known as simultaneous detection and identification), essentially two different models based on signal detection theory are currently under consideration. These two models mainly differ with respect to their assumptions regarding the interplay between the memory signals of different stimuli presented in the same lineup. The independent observations model (IOM), on the one hand, assumes that the memory signal of each simultaneously presented test stimulus is separately assessed by the decision-maker, whereas the ensemble model (EM), on the other hand, assumes that each of these memory signals is first compared with and then assessed relative to its respective context (i.e., the memory signals of the other stimuli within the same lineup). Here, we discuss some reasons why comparing confidence ratings between trials with and without a dud (i.e., a lure with no systematic resemblance to the target) in an otherwise fair lineup—results of which have been interpreted as evidence in favor of the EM—is in fact inconclusive for differentiating between the EM and the IOM. However, the lack of diagnostic value hinges on the fact that in these experiments two aspects of between-item similarity (viz. old–new and within-lineup similarity) are perfectly confounded. Indeed, if separately manipulating old–new similarity, we demonstrate that EM and IOM make distinct predictions. Following this, we show that previously published data are inconsistent with the predictions made by the EM.

**Keywords** Recognition memory · Similarity · Eyewitness identification · Ensemble models · Signal detection theory · Simultaneous detection and identification

## Introduction

In the past, basic recognition memory research largely relied on single-item recognition paradigms (e.g., Bröder & Schütz, 2009) in which participants are asked to decide for each individually presented test stimulus whether it is a previously studied *old* item (i.e., a *target*) or a nonstudied *new* item (i.e., a *lure*). Over the last few decades, memory researchers have proposed a plethora of different cognitive models to explore the latent processes underlying decisions in these situations (see, e.g., Malmberg, 2008; Parks & Yonelinas, 2008; Rotello, 2017; Jang et al., 2009).

A prominent assumption made by many such models is that a *continuous* memory-strength signal (the so-called

*familiarity*; Morrell et al., 2002; cf. Glanzer et al., 2009) is elicited by each test stimulus presented in a recognition memory task. These familiarity values are assumed to be realizations of a random variable (RV) and are thus stochastic in nature. The decision-maker can then compare each familiarity value to a *response criterion* ( $\lambda$ ) on the memory-strength dimension (Kellen & Klauer, 2018). Whenever a familiarity value exceeds the response criterion, the corresponding test stimulus will be categorized as being old (i.e., previously encountered during study). Otherwise, the test stimulus will be categorized as being new (i.e., not previously encountered). Importantly, old-item familiarities ( $\mathbf{X}_o$ ) are assumed to be distributed differently from new-item familiarities ( $\mathbf{X}_n$ ), enabling above-chance discrimination between these two types of stimuli (Macmillan & Creelman, 2005). Models adhering to these functional principles are usually referred to as *signal detection theory* (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002; Kellen & Klauer, 2018) models of recognition memory (Rotello, 2017; Kellen et al.,

✉ Constantin G. Meyer-Grant  
constantin.meyer-grant@psychologie.uni-freiburg.de

<sup>1</sup> Department of Psychology, University of Freiburg, 79085, Freiburg, Germany

2021). Often, both  $\mathbf{X}_o$  and  $\mathbf{X}_n$  are assumed to follow a normal distribution (i.e.,  $\mathbf{X}_o \sim \mathcal{N}(\mu_{\mathbf{X}_o}, \sigma_{\mathbf{X}_o})$  and  $\mathbf{X}_n \sim \mathcal{N}(\mu_{\mathbf{X}_n}, \sigma_{\mathbf{X}_n})$ ), leading to the so-called *Gaussian SDT model sub-class*.<sup>1</sup>

In recent years, however, more complex experimental paradigms have come into focus as they allow for novel and more refined tests of specific model assumptions (e.g., Kellen et al., 2021; Kellen & Klauer, 2014; Meyer-Grant & Klauer, 2021; Voormann et al., 2021). Consider, for example, the situation in which participants are presented with a set of  $m > 1$  stimuli at the same time in each test trial. Furthermore, assume that each of these sets can either contain an old stimulus (a target trial) or not (a non-target trial). Participants can now be asked to indicate whether they believe a target to be present or not (*1-out-of- $m$  detection sub-task*), and they can also be asked to identify the stimulus they think is most likely to be old ( *$m$ -alternative forced-choice sub-task*). Together, these two sub-tasks are referred to as *simultaneous detection and identification (SDAI)* (Macmillan & Creelman, 2005).

Notably, this paradigm is of great importance for applied memory research, especially for the investigation of eyewitness identification. This is due to the fact that SDAI is mostly equivalent to the widely used simultaneous lineup procedure (see, e.g., Mickes & Gronlund, 2017). In such a task, witnesses to a crime are presented with the suspect along with some innocent *fillers*. The suspect can either be guilty (i.e., a target-present lineup) or innocent (i.e., a target-absent lineup). Witnesses are then asked to indicate whether they believe the perpetrator is present in the current lineup, and if so, to identify the perpetrator.<sup>2</sup> However, this research tradition is only just beginning to explore the potential of mathematical modeling, with Gaussian SDT models in particular rapidly gaining popularity (e.g., Cohen et al., 2021; Wixted & Mickes, 2014; Wixted et al., 2018;

Lee & Penrod, 2019; Wixted & Mickes, 2018; Colloff et al., 2017).

Interestingly, SDAI has until recently been largely ignored by more basic memory researchers concerned with the evaluation and validation of such formal recognition memory models (but see Meyer-Grant & Klauer, 2021). However, a critical assessment of the core conceptual assumptions of said models seems advisable because applied researchers might otherwise draw erroneous conclusions based on improper models. This problem is further aggravated by the fact that the classical SDT framework can be extended in various ways to account for the responses in the 1-out-of- $m$  detection sub-task (Wixted et al., 2018). Two of these competing approaches—namely, the *independent observations model (IOM)* and the *ensemble model (EM)*—are of particular interest because both of them are considered to be reasonably well supported by previous empirical studies, despite proposing substantially different cognitive mechanisms (Wixted et al., 2018; Akan et al., 2021).<sup>3</sup> As we will outline below, this is mostly due to the fact that the experimental designs of past empirical investigations are inconclusive with respect to differentiating between these two models. For the purpose of addressing these issues, we will first briefly recapitulate and formalize the characteristics of both modeling accounts.

## Independent Observations Model

Arguably the simplest SDT model of SDAI is the IOM, which has been known for quite some time in the literature on SDT (Starr et al., 1975; Green et al., 1977; Green & Birdsall, 1978; Macmillan & Creelman, 2005). This model assumes that each item's familiarity value will be separately compared to the response criterion  $\lambda \in \mathbb{R}$ . If at least one of the  $m$  familiarity values exceeds  $\lambda$ , participants will give a “target present” response. Therefore, the model prediction for the probability of a *hit* (H) in the 1-out-of- $m$  detection sub-task (i.e., a “target present” response if an old item is indeed present<sup>4</sup>) is given by the probability that the maximum of all familiarity values exceeds  $\lambda$ , that is,

$$P_{\text{IOM}}(\text{H}) = P\left(\max\left\{\mathbf{X}_o, \max_{i \in \{1, \dots, r\}}\{(\mathbf{X}_n)_i\}\right\} > \lambda\right),$$

where  $r$  is the number of simultaneously presented lures.

<sup>1</sup>The normal distribution is chosen to model familiarity more for practical than substantive reasons (but see Footnote 8) and is therefore usually considered to be merely an auxiliary assumption (see, e.g., Kellen & Klauer, 2018; Kellen et al., 2021; Rouder et al., 2014). However, due to its almost ubiquitous use in modeling recognition memory, we will mainly focus on Gaussian SDT models in the following. Moreover, we will later demonstrate that—under quite general conditions—the main argument of the present work is not necessarily jeopardized by using alternative distributional assumptions.

<sup>2</sup>Note that for forensic purposes, witnesses should not be asked to always identify the person they believe is most likely to be the perpetrator, as it is not intended to suggest that the perpetrator must always be present in any given lineup. Instead, it is recommended that witnesses only give an identification response (i.e., a response to the  $m$ -alternative forced-choice sub-task) if they believe they recognize the perpetrator (Wells et al., 2020); that is, for lineups in which they believe the perpetrator to be present. In other words, whether or not witnesses are asked to provide an identification response depends on their decision in the 1-out-of- $m$  detection sub-task.

<sup>3</sup>Wixted et al. (2018) also discuss a third variant (viz., the *integration model*). However, this model could not be reconciled with their data and was therefore decisively rejected (Wixted et al., 2018), which is why we omit it here.

<sup>4</sup>Note that the meaning of the term “hit” in the context of SDAI refers to a target trial being correctly detected as such, regardless of whether that decision is followed by the selection of a target or a lure in the  $m$ -alternative forced-choice sub-task. In other words, the term “hit” refers exclusively to the responses in the 1-out-of- $m$  detection sub-task.

The IOM further assumes that the identification response (i.e., the response in the  $m$ -alternative forced-choice sub-task) will be made according to which item elicited the highest familiarity value (i.e., a maximum decision rule; see, e.g., Norman & Wickelgren, 1969). Thus, the model prediction for the probability of correctly identifying the old item (I) and a hit is given by

$$P_{\text{IOM}}(\text{I}, \text{H}) = P\left(\mathbf{X}_o > \lambda, \mathbf{X}_o > \max_{i \in \{1, \dots, r\}} \{(\mathbf{X}_n)_i\}\right).$$

Figure 1 shows examples of probability density functions (PDFs) of an equal-variance Gaussian SDT model (i.e.,  $\sigma_{\mathbf{X}_o} = \sigma_{\mathbf{X}_n} = 1$ ) for both target and lure familiarities as well as the corresponding PDFs of the IOM’s decision variables for a target (i.e.,  $\max\{\mathbf{X}_o, \mathbf{X}_n\}$ ) and a non-target (i.e.,  $\max\{(\mathbf{X}_n)_1, (\mathbf{X}_n)_2\}$ ) trial, both with  $m = 2$ . Moreover, Fig. 1 also depicts the respective model predictions for the so-called *receiver operating characteristic* (ROC) as well as the *identification operating characteristic* (IOC; Macmillan & Creelman, 2005). By plotting the model predictions for  $P(\text{H})$  and  $P(\text{I}, \text{H})$ , respectively, against the model predictions for the probability of a so-called *false alarm* (FA; i.e., a “target present” response in the absence of an old item<sup>5</sup>), these curves show how changes in the response criterion affect the relationship between the predicted response behavior in target and non-target trials (Macmillan & Creelman, 2005).

### Ensemble Model

The EM likewise assumes that each simultaneously presented test stimulus elicits an individual familiarity value. However, other than the IOM, it assumes that for each set of test items, the decision-maker first computes the mean familiarity of all simultaneously elicited item familiarity values

$$\mathbf{M} := \begin{cases} \frac{\mathbf{X}_o + \sum_{i=1}^m (\mathbf{X}_n)_i}{m}, & \text{given a target trial} \\ \frac{\sum_{i=1}^m (\mathbf{X}_n)_i}{m}, & \text{given a non-target trial} \end{cases}.$$

Then, the difference between  $\mathbf{M}$  and each individual item familiarity is computed. This difference is itself a RV with a distribution that depends on the true status of the respective stimulus (i.e., whether it is old or new) and on the present context of stimuli (i.e., whether it is a target or a non-target trial). We denote this RV as  $\mathbf{D}_o := \mathbf{X}_o - \mathbf{M}$  for old items and as  $(\mathbf{D}_n)_i := (\mathbf{X}_n)_i - \mathbf{M}$  for new items. If at least one of these decision variables exceeds the response criterion  $\lambda \geq 0$ , a “target present” response will be given. The model

predictions for the probability of a hit is consequently given by

$$P_{\text{EM}}(\text{H}) = P\left(\max\left\{\mathbf{D}_o, \max_{i \in \{1, \dots, r\}} \{(\mathbf{D}_n)_i\}\right\} > \lambda\right).$$

In other words, the EM assumes that the decision in the 1-out-of- $m$  detection sub-tasks depends on how much an individual item “stands out” from the rest of the lineup (i.e., a “target present” response is given if a single item appears sufficiently more familiar than the rest of the lineup). In contrast, the IOM assumes that each stimulus within a lineup is assessed separately by the decision-maker and a “target present” response is given if at least one item appears sufficiently familiar. From a theoretical perspective, these two mechanisms therefore correspond to what are known as *relative* and *absolute* decision strategies (e.g., Dunning & Stern, 1994; Charman & Wells, 2007; Clark et al., 2011), respectively.

The identification response, on the other hand, is— analogously to the IOM—determined by which item elicited the maximal familiarity value. Thus, the model prediction for the probability of a correct identification and a hit is given by

$$\begin{aligned} P_{\text{EM}}(\text{I}, \text{H}) &= P\left(\mathbf{D}_o > \lambda, \mathbf{X}_o > \max_{i \in \{1, \dots, r\}} \{(\mathbf{X}_n)_i\}\right) \\ &= P\left(\mathbf{D}_o > \lambda, \mathbf{D}_o > \max_{i \in \{1, \dots, r\}} \{(\mathbf{D}_n)_i\}\right). \end{aligned}$$

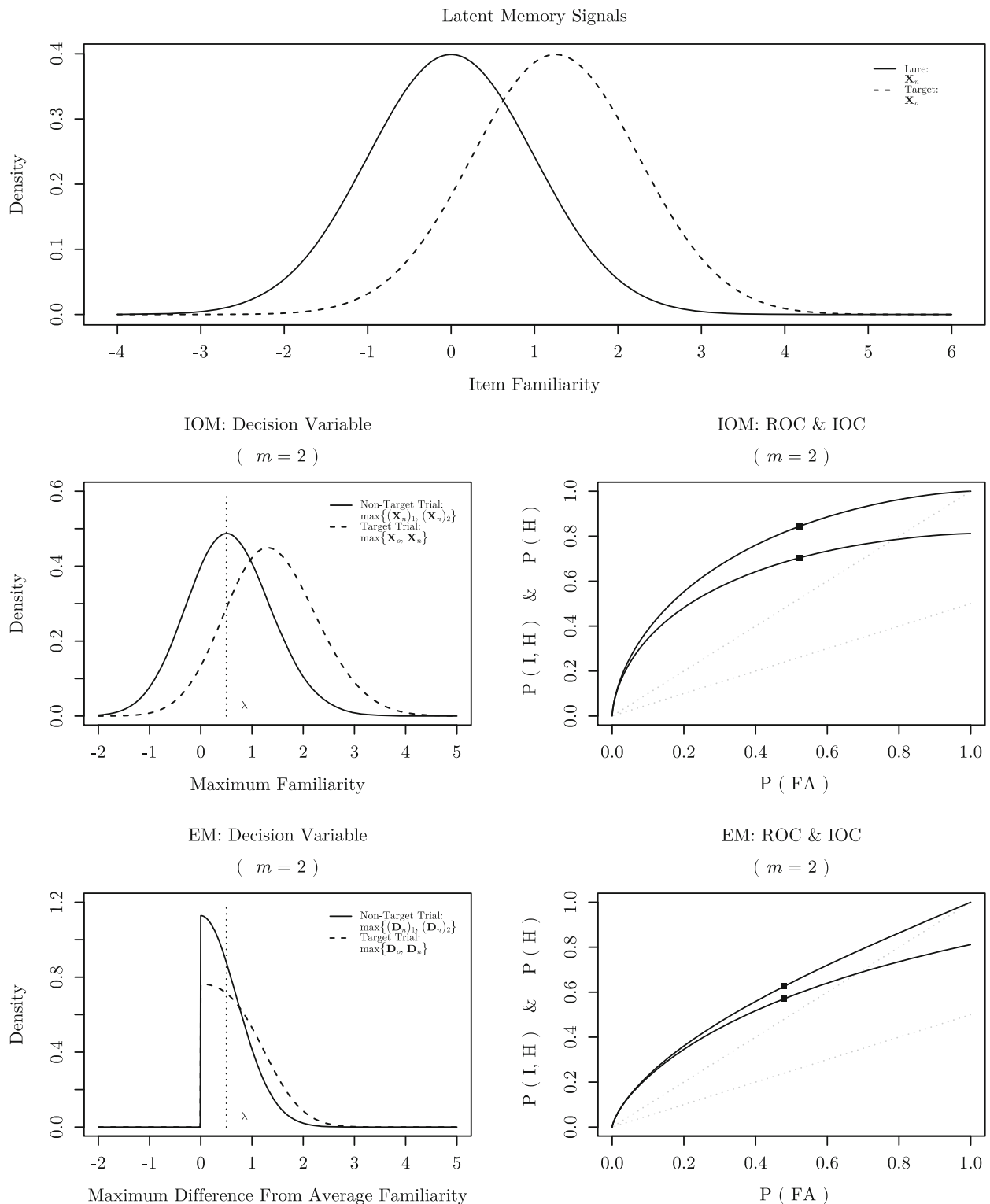
Figure 1 also shows the PDFs of the EM’s decision variables for a target (i.e.,  $\max\{\mathbf{D}_o, \mathbf{D}_n\}$ ) and a non-target (i.e.,  $\max\{(\mathbf{D}_n)_1, (\mathbf{D}_n)_2\}$ ) trial, both with  $m = 2$ , as well as the respective model predictions for the ROC and IOC.

### Similar Lures and the Dud-Alternative Effect

One piece of evidence that—at first glance—seems to favor the EM over the IOM is what Windschitl and Chambers (2004; see also Wixted et al., 2018; Charman et al., 2011) referred to as the *dud-alternative effect*. This effect describes the phenomenon that reported confidence (in the target being present<sup>6</sup>) tends to increase when an

<sup>5</sup>Like the term “hit,” the term “false alarm” refers exclusively to the responses in the 1-out-of- $m$  detection sub-task. That is, in the context of SDAI, the term “false alarm” denotes incorrectly reporting the presence of a target in a non-target trial.

<sup>6</sup>In the past, most experiments investigating the dud-alternative effect have measured confidence ratings associated with the subsequent identification response (i.e., the response regarding the  $m$ -alternative forced-choice sub-task; Horry & Brewer, 2016; Charman et al., 2011; Wixted et al., 2018). However, the confidence in the identification response is—strictly speaking—not what is actually modeled, but rather the confidence in the 1-out-of- $m$  detection response. This is the case because the decision variable of the 1-out-of- $m$  detection sub-task determines the confidence level in both models. From a practical perspective, it seems appropriate to assume that both kinds of confidence ratings are based on the same underlying decision variable and hence are identical. We therefore adopt this view in the following. Nevertheless, testing whether this assumption in fact holds might be a worthwhile endeavor in its own right, which is, however, beyond the scope of the present work.



**Fig. 1** Graphical representation of different equal-variance Gaussian SDT models for an SDAI task with  $m = 2$ . Upper panel: PDFs of lure ( $\mu_{X_n} = 0$ ; solid line) and target ( $\mu_{X_o} = 1.25$ ; dashed line) familiarities, respectively. Note that  $X_n$  and  $X_o$  are uncorrelated. Middle left panel: PDFs of the decision variables of the IOM in target (dashed line) and non-target (solid line) trials, respectively. The dotted line indicates the position of the response criterion  $\lambda = 0.5$ . Middle right panel: The predicted ROC (upper line) and IOF (lower line) of the

IOM. Dotted gray lines represent guessing level. Black squares indicate the predicted probabilities for the response criterion  $\lambda$ . Lower left panel: PDFs of the decision variable of the EM in target (dashed line) and non-target (solid line) trials, respectively. The dotted line indicates the position of the response criterion  $\lambda$ . Lower right panel: The predicted ROC (upper line) and IOF (lower line) of the EM. Dotted gray lines represent guessing level. Black squares indicate the predicted probabilities for the response criterion  $\lambda$

implausible alternative (a so-called *dud*, i.e., a regular lure which exhibits no systematic similarities to the target) is included in an otherwise *fair* lineup (i.e., a lineup comprised of stimuli systematically resembling each other; Charman et al., 2011; Horry & Brewer, 2016; Fitzgerald et al., 2013). By the same token, adding a lure that resembles the target (i.e., a similar lure) to an *unfair* lineup consisting of randomly assembled stimuli (i.e., the stimuli of the lineup do not systematically resemble each other) would tend to reduce confidence.

In the context of eyewitness identification, for example, similar lures are fillers whose visual characteristics match the verbal descriptions of the perpetrator and/or the general appearance of the suspect. The use of similar lures in real-life lineups (i.e., constructing fair lineups) is generally recommended since it has been shown to reduce the number of falsely identified innocent suspects (Wells et al., 2020; Fitzgerald et al., 2013; Smith et al., 2017). This is the case because the rate of “target present” responses was found to increase for unfair lineups compared to fair lineups, regardless of whether the suspect is guilty or innocent (Fitzgerald et al., 2013). Indeed, the *dud*-alternative effect is closely related to this observation as it describes an increase in confidence when the lineup becomes less fair (Horry & Brewer, 2016; Fitzgerald et al., 2013).

Importantly, confidence is intricately tied to the magnitude of the decision variable in SDT models of recognition memory; that is, the larger the decision variable, the more confident decision-makers are that a target was present (Kellen & Klauer, 2018; Dubé & Rotello, 2012; cf. Mickes et al., 2017). Modeling a confidence rating (i.e., a graded response) is then simply accomplished by assuming multiple staggered response criteria that partition the memory-strength dimension into multiple confidence regions. Furthermore, this entails that an increase in confidence is tantamount to an increase in hit (or false alarm) rate over different confidence levels. To explain the *dud*-alternative and related effects, it has thus been proposed that the more salient the target familiarity is compared to its context (facilitated by the inclusion of a *dud* in place of a similar lure), the larger the decision variable and—as a consequence—the higher the associated confidence. The EM is in essence an instantiation of this idea (Wixted et al., 2018), which is why the *dud*-alternative effect lends some credence to this account.

At this point, however, it is necessary to draw attention to one way in which between-item similarity has previously been modeled: by introducing a positive within-lineup correlation of memory signals while neither changing the mean of the lure ( $\mu_{X_n}$ ) nor the target ( $\mu_{X_o}$ ) familiarity distribution (see, e.g., Wixted et al., 2018). A positive within-lineup correlation ( $\rho_{\{X_o, X_n\}} > 0$ ) between target and lure familiarity implies that if the target appears to be

relatively familiar (i.e., more familiar than an average target) to the decision-maker, the lure tends to appear relatively familiar (i.e., more familiar than an average lure) as well, and vice versa. This modeling decision does address one key aspect of between-item similarity, but overlooks another crucial aspect that must not be ignored in this context.

In order to see why this is the case, consider a situation in which target and lure are de facto indistinguishable by a human observer (e.g., through only minimally changing the color of a single pixel in a picture that was presented during study, as discussed by Meyer-Grant & Klauer, 2021). According to the modeling approach described above, the correlation between the target and lure familiarity values  $\rho_{\{X_o, X_n\}}$  would approach one in such a case. Thus, the memory-strength “signals generated by the target and the lure on any given trial fall at precisely the same point on their respective distributions” (Wixted et al., 2018, p. 84; see also Hintzman, 2001). If, for example, the familiarity value elicited by the target falls one standard deviation below  $\mu_{X_o}$ , then the familiarity value elicited by a lure will fall one standard deviation below  $\mu_{X_n}$ . However, as long as the means of both the target ( $\mu_{X_o}$ ) and the lure ( $\mu_{X_n}$ ) familiarity distributions remain unaffected by this manipulation (and also  $\mu_{X_o} > \mu_{X_n}$ ), this implies that the more similar the stimuli are, the better the identification performance becomes, until eventually the model will predict the identification responses to be always correct (Wixted et al., 2018).<sup>7</sup>

Common sense, on the other hand, strongly suggest that the opposite must be the case. That is, if a decision-maker cannot distinguish between the stimuli, then (assuming no additional biases) the response must be based solely on guessing (see also Luus & Wells, 1991; Wells et al., 2020). Thus, in addition to an increase in the within-lineup correlation between familiarities, an increase in similarity should be accompanied by a decrease in the mean difference between target and lure familiarity distributions (i.e.,  $\mu_{X_n} - \mu_{X_o}$ ). More precisely, as  $\rho_{\{X_o, X_n\}}$  approaches one,  $\mu_{X_n} - \mu_{X_o}$  must approach zero.<sup>8</sup>

<sup>7</sup>Note that this behavior does not depend on whether we adopt the IOM or the EM framework, as both share the same mechanism for modeling the identification response.

<sup>8</sup>Another argument that could be made here is that one rationale for using the normal distribution to model latent memory strength is to assume that these values are formed by summation over partial memory information (see, e.g., Hintzman, 1984). The normality assumption then follows from the central limit theorem under quite general regularity conditions (Kellen & Klauer, 2018). Given that a similar lure systematically shares a certain proportion of target characteristics, the corresponding partial memory information will likewise coincide. This not only results in both values being correlated, but also in bringing the expected value of the similar lure familiarities closer to the expected value of the target familiarities.



This illustrates that in the dud-alternative paradigm we are dealing with two different but perfectly confounded aspects of between-item similarity: On the one hand, the stimuli of a lineup can resemble each other (i.e., the *within-lineup similarity*, as modeled by the within-lineup correlation), and, on the other hand, a new stimulus can resemble an old stimulus (i.e., the *old–new similarity*, as modeled by the mean difference between target and lure familiarity distributions).

Consider, for example, a forensic lineup of size  $m = 2$  which contains the suspect and a filler that closely resembles the description of the perpetrator. Since the suspect naturally also matches this description, both within-lineup similarity and old–new similarity are high in such a scenario. For an illustration of how to disentangle the two kinds of similarities, suppose that the crime was committed by two perpetrators that are not systematically similar to each other. This allows for the construction of a lineup that contains a suspect that resembles one of the perpetrators and a filler that matches the description of the other perpetrator. In this case, within-lineup similarity will be relatively low, whereas old–new similarity will be high.

Fortunately, SDT models of SDAI can easily be modified to account for both aspects of similarity present in the dud-alternative paradigm. Let us therefore only denote the familiarity of a dud (i.e., a regular lure) as  $\mathbf{X}_n$  and—in distinction to this—the familiarity of a similar lure as  $\mathbf{X}_s \sim \mathcal{N}(\mu_{\mathbf{X}_s}, \sigma_{\mathbf{X}_s})$ . Now, we can specify that  $\mathbf{X}_o$  and  $\mathbf{X}_s$  are positively correlated (denoted by the parameter  $\rho_{\{\mathbf{X}_o, \mathbf{X}_s\}} > 0$ ) while at the same time  $\mathbf{X}_o$  and  $\mathbf{X}_n$  are uncorrelated (i.e.,  $\rho_{\{\mathbf{X}_o, \mathbf{X}_n\}} = 0$ ). Furthermore, we assume that  $\mu_{\mathbf{X}_o} > \mu_{\mathbf{X}_s} > \mu_{\mathbf{X}_n}$  and that the closer a similar lure resembles the target, the larger  $\rho_{\{\mathbf{X}_o, \mathbf{X}_s\}}$  and the closer  $\mu_{\mathbf{X}_s}$  approaches  $\mu_{\mathbf{X}_o}$ .

Expanding the SDT model framework in the above-mentioned way does not compromise the conceptual appeal of the EM in terms of predicting a dud-alternative effect. But the presence of the dud-alternative effect is at the same time not yet sufficient to rule out the IOM. The reasons for this are twofold:

First, the presence or absence of a similar lure in the dud-alternative paradigm is clearly apparent to the decision-maker and thus not opaque to conscious understanding.<sup>9</sup> Thus, as was noted by Hanczakowski et al. (2014; see also Wixted et al. ; 2018), participants could alter their response criterion between both conditions; that is, they might adopt a more conservative response criterion when a similar lure is added to the lineup.

Second, the IOM in fact also predicts the dud-alternative effect to occur for many reasonable parameter specifications, even without assuming a shift in the response criterion. This is due to the fact that the lower the correlation between target and lure familiarities within the same lineup, the higher the chance that a relatively high lure familiarity will be elicited together with a relatively low target familiarity.<sup>10</sup> One can imagine that if a target and a lure do not resemble each other, sometimes the decision-maker will erroneously believe to recognize the lure while not recognizing the target. If, on the other hand, target and lure are highly similar, the decision-maker will most likely not mistake the lure for being old if the target already appears very unfamiliar. In essence, the inclusion of a dud adds another independent attempt to surpass the response criterion, which is why this event tends to occur more often (i.e., a hit becomes more likely; see also Fig. 2, top row of panels).<sup>11</sup>

One might object to this that the dud-alternative effect is also present when only considering the confidence ratings associated with a correct identification (i.e., a target identification; e.g., Horry & Brewer, 2016). Initially, the previous argument seems to be invalidated by this observation; however, this is actually not the case: The predicted probability of instances in which the lure familiarity value surpasses the target familiarity value, assuming no correlation between them and given their maximum exceeds the response criterion  $\lambda$  (i.e., the predicted probability of an incorrect identification given a hit), is monotonically linked to  $\lambda$  itself in the IOM. That is, the smaller  $\lambda$ , the more likely a lure will be identified over a target, given a hit (see Meyer-Grant & Klauer, 2021, who provided a proof of this dependency for  $m = 2$  and different parametrizations of the IOM, including for normally distributed familiarity values<sup>12</sup>). In other words, the smaller the maximum familiarity value, the more likely it is that this value was elicited by a lure rather than a

<sup>10</sup>Note that the dud-alternative effect has sometimes been investigated in the past by simply *adding* a dud to the lineup (Charman et al., 2011). This resulted in an increase in lineup size ( $m$ ) between both conditions. In such a situation, there is yet another mechanism of the IOM that predicts the dud-alternative effect to occur: The maximum of the lineup with a dud will tend to be greater than that of the lineup without one, simply because there are more items available to generate the maximum. However, other studies found an analogous effect even if  $m$  is constant between both conditions (Horry & Brewer, 2016), which is why this mechanism alone is not sufficient to account for past observations.

<sup>11</sup>Importantly, this causal mechanism does not depend on the number of stimuli ( $m$ ) presented in a single lineup. Thus, the IOM's ability to predict the dud-alternative effect does not depend on lineup size. However, analogous to the EM, the magnitude of the effect predicted by the IOM decreases as  $m$  increases.

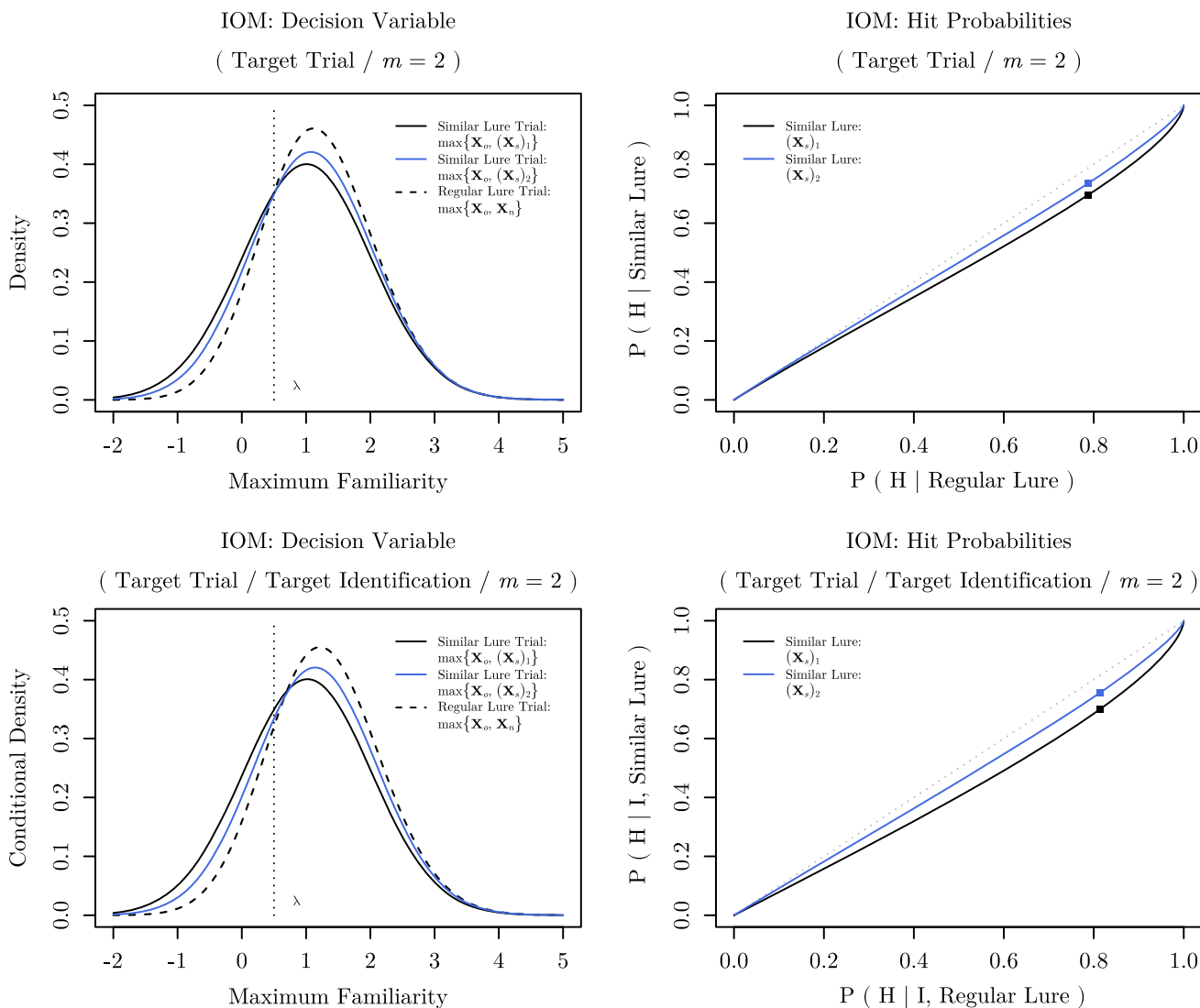
<sup>12</sup>To our knowledge, there is no rigorous proof of this dependency when  $m > 2$ , but simulations suggest that it holds for larger lineups as well.

<sup>9</sup>The aspect of between-item similarity that is noticeable is arguably the within-lineup similarity and not the old–new similarity (i.e., a decision-maker likely discerns whether or not the stimuli of a single lineup resemble each other).

target. Hence, these instances are selectively removed when conditioning on a correct identification response, which in turn results in an observable increase in confidence in situations in which target and lure familiarities are uncorrelated compared to situations in which they are correlated (i.e., the dud-alternative effect). In fact, this can lead to a stronger dud-alternative effect than if lure identifications are also taken into account (see also Fig. 2, bottom row of panels).

### A Critical Test

Whether or not a dud-alternative effect occurs in an SDAI task is therefore not diagnostic for deciding between the IOM and the EM. However, the task can be slightly modified in order to be much more informative in that regard. While (for target trials) the within-lineup similarity cannot be manipulated without affecting old–new similarity, separately manipulating old–new similarity is in fact possible.



**Fig. 2** Graphical representation of the equal-variance Gaussian IOM ( $\mu_{\mathbf{x}_o} = 1.00$ ,  $\mu_{(\mathbf{x}_s)_1} = 0.25$ ,  $\mu_{(\mathbf{x}_s)_2} = 0.10$ ,  $\mu_{\mathbf{x}_n} = 0.00$ ,  $\rho_{\{\mathbf{x}_o, (\mathbf{x}_s)_1\}} = 0.90$ ,  $\rho_{\{\mathbf{x}_o, (\mathbf{x}_s)_2\}} = 0.50$ ,  $\rho_{\{\mathbf{x}_o, \mathbf{x}_n\}} = 0.00$ ) for  $m = 2$ . Left column of panels: PDFs (top left panel) and probability densities conditional on a target identification (bottom left panel) of the decision variable in target trials with a similar (solid lines; black:  $(\mathbf{X}_s)_1$ , blue:  $(\mathbf{X}_s)_2$ ) and a regular (dashed line) lure, respectively. The dotted line indicates the position of the response criterion  $\lambda = 0.5$ . Right column

of panels: The model predictions for the probability of a hit (top right panel) in target trials with (ordinate) and without (abscissa) a similar lure as well as the model predictions for the probability of a hit given a target identification (bottom right panel) in target trials with (ordinate) and without (abscissa) a similar lure (black:  $(\mathbf{X}_s)_1$ , blue:  $(\mathbf{X}_s)_2$ ). The dud-alternative effect corresponds to a curve below the diagonal identity line (dotted gray line). The stronger the effect, the lower the curve. Squares indicate the predicted probabilities for the response criterion  $\lambda$

This simply requires that a similar lure (i.e., a new stimulus which resembles an old one) is not paired with its similar old-item sibling during test, but instead is presented together with another old item (i.e., the similar lure resembles an old stimulus, but not the target presented in the same lineup). This raises the old–new similarity in such lineups while leaving within-lineup similarity unaffected.

Such an experiment (with  $m = 2^{13}$ ) was recently conducted by Meyer-Grant and Klauer (2021). They first presented a sequence of several portrait pictures to the participants, who were asked to memorize them. In a subsequent test phase, they then presented participants with lineups with no systematic within-lineup similarity that either contained a target and a regular lure, a target and a similar lure, a regular lure and a similar lure, or two regular lures.<sup>14</sup> Participants were then asked to provide a four-level confidence rating as to whether or not they believed a target to be present (i.e., the 1-out-of- $m$  detection sub-task), and to subsequently identify the stimulus they most likely believed to be the target (i.e., the  $m$ -alternative forced-choice sub-task).

Their results revealed that, first, people were more likely to correctly identify the target when they were more confident that a target was present; second, people were more likely to falsely identify a similar lure (a so-called *pseudo-identification*) over a regular lure when they were more confident that a target was present; and third, people were more likely to think a target was present in a target trial when the target was accompanied by a similar lure than when the target was accompanied by a regular lure instead (Meyer-Grant & Klauer, 2021). They further showed that the IOM is consistent with these observations. Here, we will derive the respective predictions of the EM and show that, unlike the IOM, it cannot account for this pattern of effects.

Let us therefore initially consider only the first two critical predictions identified by Meyer-Grant and Klauer (2021). It can be shown that both predictions follow from certain rank order probabilities (e.g.,  $P(\mathbf{D}_o > \mathbf{D}_n | \max(\mathbf{D}_o, \mathbf{D}_n) > \lambda)$  in case of the EM) exhibiting monotonicity under changes

in the response criterion (Meyer-Grant & Klauer, 2021). As we will demonstrate next, this property indeed also holds for the Gaussian EM.

In order to provide a rigorous proof of this claim, we first note that for a target trial and  $m = 2$

$$\mathbf{D}_o = \mathbf{X}_o - \mathbf{M} = \mathbf{X}_o - \frac{\mathbf{X}_o + \mathbf{X}_n}{2} = \frac{\mathbf{X}_o - \mathbf{X}_n}{2}$$

and

$$\mathbf{D}_n = \mathbf{X}_n - \mathbf{M} = \mathbf{X}_n - \frac{\mathbf{X}_o + \mathbf{X}_n}{2} = \frac{\mathbf{X}_n - \mathbf{X}_o}{2} = -\mathbf{D}_o.$$

Because the normal distribution is stable under convolution, we further find that

$$\mathbf{D}_o \sim \mathcal{N} \left( \frac{\mu_{\mathbf{X}_o} - \mu_{\mathbf{X}_n}}{2}, \sqrt{\frac{\sigma_{\mathbf{X}_o}^2 + \sigma_{\mathbf{X}_n}^2 - 2\text{Cov}(\mathbf{X}_o, \mathbf{X}_n)}{4}} \right)$$

in the Gaussian case, where  $\text{Cov}(\mathbf{X}_o, \mathbf{X}_n) = \rho_{\{\mathbf{X}_o, \mathbf{X}_n\}} \sigma_{\mathbf{X}_o} \sigma_{\mathbf{X}_n}$ . Without loss of generality (except for the two degenerate cases with perfect correlation of either  $\rho_{\{\mathbf{X}_o, \mathbf{X}_n\}} = 1$  or  $\rho_{\{\mathbf{X}_o, \mathbf{X}_n\}} = -1$ ) we set

$$\sigma_{\mathbf{D}_o} := \sqrt{\frac{\sigma_{\mathbf{X}_o}^2 + \sigma_{\mathbf{X}_n}^2 - 2\text{Cov}(\mathbf{X}_o, \mathbf{X}_n)}{4}} = 1$$

in the following and define

$$\mu_{\mathbf{D}_o} := \frac{\mu_{\mathbf{X}_o} - \mu_{\mathbf{X}_n}}{2}.$$

As previously mentioned, the EM assumes that the decision in the 1-out-of- $m$  detection sub-task is determined by whether or not  $\max\{\mathbf{D}_o, \mathbf{D}_n\}$  exceeds the response criterion  $\lambda$ . We can therefore express the predicted probability of a so-called *miss* ( $\mathbf{M}$ ; i.e., a “target absent” response if an old item is present<sup>15</sup>) as

$$\begin{aligned} P_{EM}(\mathbf{M}) &= P(\max\{\mathbf{D}_o, \mathbf{D}_n\} < \lambda) = P(\max\{\mathbf{D}_o, -\mathbf{D}_o\} < \lambda) \\ &= P(|\mathbf{D}_o| < \lambda). \end{aligned}$$

However, in the Gaussian case this expression is equivalent to the cumulative distribution function (CDF) of a *folded normal distribution* (more specifically, of a folded normal distribution with scale parameter  $\sigma_{\mathbf{D}_o} = 1$  and location parameter  $\mu_{\mathbf{D}_o}$ ). Thus, it holds that

$$P_{EM}(\mathbf{M}) = \begin{cases} \Phi(\lambda - \mu_{\mathbf{D}_o}) - \Phi(-\lambda - \mu_{\mathbf{D}_o}), & \text{if } \lambda \geq 0 \\ 0, & \text{otherwise} \end{cases},$$

where  $\Phi$  is the CDF of the standard normal distribution.

<sup>13</sup>Henceforth we will confine ourselves to an SDAI task with  $m = 2$  as we want to focus on the data by Meyer-Grant and Klauer (2021) and choosing  $m = 2$  conforms to their experimental design. Furthermore, it will suffice for the central research objective of the present work, that is, to assess the tenability of the EM. An interesting side note is that for  $m = 2$ , the so-called *best vs. rest* model (Clark et al., 2011)—a variant of the EM in which the maximum familiarity is compared to the mean familiarity value of the remaining items (Wixted et al., 2018)—is equivalent to the so-called *best vs. next* model (Clark et al., 2011), in which the maximum familiarity is compared to the maximum familiarity of the remaining items. Hence, the critical test presented here applies to these models as well.

<sup>14</sup>Note that in order to construct these different types of lineups, it was essential to present multiple stimuli during study. Only in this way it was possible to construct lineups with a target and a similar lure, that is, a lure which resembles another studied stimulus other than the target of the current lineup.

<sup>15</sup>Like the terms “hit” and “false alarm,” the term “miss” refers exclusively to the responses in the 1-out-of- $m$  detection sub-task. That is, in the context of SDAI, the term “miss” denotes incorrectly rejecting the presence of a target in a target trial, regardless of whether that decision is followed by the selection of a target or a lure in the  $m$ -alternative forced-choice sub-task.



Furthermore, we can express the predicted probability of a miss and a subsequent correct identification as

$$P_{EM}(I, M) = P(\mathbf{D}_o < \lambda, \mathbf{D}_o > 0)$$

since  $\mathbf{D}_o > 0$  implies  $\mathbf{X}_o > \mathbf{X}_n$  if  $m = 2$ . Through conditioning on  $\mathbf{D}_o = x$  we find that

$$P(\mathbf{D}_o < \lambda, \mathbf{D}_o > 0) = \begin{cases} \int_0^\lambda \varphi(x - \mu_{\mathbf{D}_o}) dx, & \text{if } \lambda \geq 0 \\ 0, & \text{otherwise} \end{cases},$$

where  $\varphi$  is the PDF of the standard normal distribution. Thus, it holds that

$$P_{EM}(I, M) = \begin{cases} \Phi(\lambda - \mu_{\mathbf{D}_o}) - \Phi(-\mu_{\mathbf{D}_o}), & \text{if } \lambda \geq 0 \\ 0, & \text{otherwise} \end{cases}.$$

The predicted probability of a correct identification given a miss can therefore be expressed as

$$P_{EM}(I|M, \lambda) = \frac{P_{EM}(I, M|\lambda)}{P_{EM}(M|\lambda)} = \frac{\Phi(\lambda - \mu_{\mathbf{D}_o}) - \Phi(-\mu_{\mathbf{D}_o})}{\Phi(\lambda - \mu_{\mathbf{D}_o}) - \Phi(-\lambda - \mu_{\mathbf{D}_o})} \\ =: \tau_{M(m=2)}(\lambda)$$

for all  $\lambda > 0$  according to the Gaussian EM.

Furthermore, one finds by means of elementary probability theory that

$$P_{EM}(H) = \begin{cases} 1 - \Phi(\lambda - \mu_{\mathbf{D}_o}) + \Phi(-\lambda - \mu_{\mathbf{D}_o}), & \text{if } \lambda \geq 0 \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

and

$$P_{EM}(I, H) = \begin{cases} 1 - \Phi(-\mu_{\mathbf{D}_o}) - \Phi(\lambda - \mu_{\mathbf{D}_o}) + \Phi(-\mu_{\mathbf{D}_o}) = 1 - \Phi(\lambda - \mu_{\mathbf{D}_o}), & \text{if } \lambda \geq 0 \\ \Phi(\mu_{\mathbf{D}_o}), & \text{otherwise} \end{cases}.$$

Thus, it follows immediately that

$$P_{EM}(I|H, \lambda) = \frac{P_{EM}(I, H|\lambda)}{P_{EM}(H|\lambda)} \\ = \frac{1 - \Phi(\lambda - \mu_{\mathbf{D}_o})}{1 - \Phi(\lambda - \mu_{\mathbf{D}_o}) + \Phi(-\lambda - \mu_{\mathbf{D}_o})} \\ =: \tau_{H(m=2)}(\lambda). \quad (2)$$

for all  $\lambda > 0$ .

**Theorem 1**  $\tau_{H(m=2)}(\lambda)$  and  $\tau_{M(m=2)}(\lambda)$  are both strictly monotonically increasing in  $\mathbb{R}^+$  if  $\mu_{\mathbf{D}_o} > 0$  (i.e., if  $\mu_{\mathbf{X}_o} > \mu_{\mathbf{X}_n}$ ).

*Proof* See the [Appendix](#).  $\square$

Consequently, both the increases in correct rate and—as can be seen through exchanging  $\mathbf{X}_o$  with  $\mathbf{X}_s$  in the above derivations and assuming  $\mu_{\mathbf{X}_s} > \mu_{\mathbf{X}_n}$ —in pseudo-identification rate with confidence are consistent with the EM.

Interestingly, however, the model predictions regarding the last critical effect discussed by Meyer-Grant and Klauer

(2021) differ between the EM and the IOM. As we established above, if similar lures are included into the design, we must assume that  $\mu_{\mathbf{X}_s} > \mu_{\mathbf{X}_n}$  in order for the Gaussian EM to predict an increase of the pseudo-identification rate with confidence—a direct consequence of Theorem 1.<sup>16</sup> But if this is indeed the case, one can show that according to an equal-variance Gaussian EM (i.e.,  $\sigma_{\mathbf{X}_o} = \sigma_{\mathbf{X}_s} = \sigma_{\mathbf{X}_n} = 1$ ) in which the correlation between familiarity values of a lure and a target does not depend on whether the lure is a similar or regular lure (i.e.,  $\text{Cov}(\mathbf{X}_o, \mathbf{X}_s) = \text{Cov}(\mathbf{X}_o, \mathbf{X}_n)$ , as is to be assumed for the experimental data under consideration here, where the similar lure resembled an old item other than the target; see Meyer Grant & Klauer, 2021), predicted hit rates must be lower in target trials with a similar lure compared to target trials with a regular lure (if  $m = 2$ ). Furthermore, this prediction is independent of the value of the response criterion  $\lambda$  and thus holds regardless of the confidence level.

**Proposition 2** If  $\mu_{\mathbf{X}_o} > \mu_{\mathbf{X}_s} > \mu_{\mathbf{X}_n} = 0$ ,  $\sigma_{\mathbf{X}_s} = \sigma_{\mathbf{X}_n}$ , and  $\text{Cov}(\mathbf{X}_o, \mathbf{X}_s) = \text{Cov}(\mathbf{X}_o, \mathbf{X}_n)$  it must hold that

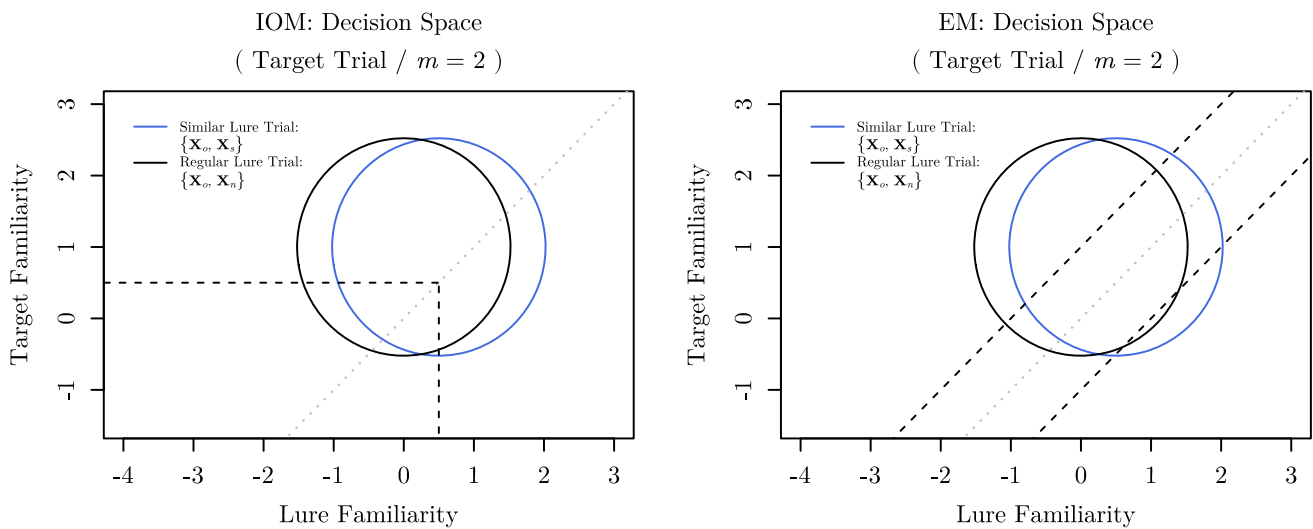
$$P_{EM}(H|\lambda, \text{Similar Lure}) \leq P_{EM}(H|\lambda, \text{Regular Lure}) \quad \forall \lambda \in \mathbb{R}^+. \quad (3)$$

*Proof* See the [Appendix](#).  $\square$

The qualitative prediction of the equal-variance Gaussian EM that is entailed by Proposition 2 is the exact opposite of the corresponding prediction made by any IOM with a monotonic likelihood ratio of similar and regular lure familiarities (see Proposition 9 in Meyer-Grant & Klauer, 2021). This can also be interpreted as an immediate consequence of the central distinguishing aspect between the EM and the IOM framework. That is, in the EM, it is not the absolute familiarity values (as in the IOM) but the relative familiarity values (i.e., how much a single stimulus' familiarity “stands out” compared to the rest of the set) that are relevant for the decision. In the presence of a similar lure (instead of a regular one) the difference between target and lure familiarity values tends to be reduced if  $\mu_{\mathbf{X}_s} > \mu_{\mathbf{X}_n}$ . Thus, the predicted hit rate must decrease in such circumstances according to the EM.

This discrepancy between the EM and the IOM can also be visualized by depicting the two-dimensional decision space of both models (Fig. 3), where each dimension corresponds to one of the  $m = 2$  test-item familiarity values. Suppose that for a target trial, values on the ordinate correspond to the old-item familiarity ( $\mathbf{X}_o$ ), whereas values

<sup>16</sup>Note also that this is an empirical argument for the point we made above that assuming  $\mu_{\mathbf{X}_s} = \mu_{\mathbf{X}_n}$ —as done in previous modeling of the dud-alternative effect—is untenable.



**Fig. 3** Two-dimensional decision space for a target trial (with  $m = 2$ ) of an equal-variance Gaussian IOM (left panel) and EM (right panel). Solid lines depict contour lines (0.05) of the bivariate probability density function of the RVs  $\{\mathbf{X}_o, \mathbf{X}_n\}$  (black) and  $\{\mathbf{X}_o, \mathbf{X}_s\}$  (blue) with  $\mu_{\mathbf{X}_o} = 1.00$ ,  $\mu_{\mathbf{X}_s} = 0.50$ ,  $\mu_{\mathbf{X}_n} = 0.00$ ,  $\rho_{\{\mathbf{X}_o, \mathbf{X}_s\}} = 0.00$ , and  $\rho_{\{\mathbf{X}_o, \mathbf{X}_n\}} = 0.00$ . The dotted gray line indicates equality between target and lure familiarity values. Realizations above and/or to the right of the dashed black lines at least one of the two familiarity values exceeds  $\lambda = 0.5$ , which corresponds to a “target present” response in the IOM. Right panel: Above the upper dashed black line,  $\mathbf{X}_o - 2\lambda$  is larger than the lure familiarity value. Below the lower dashed black line, the lure familiarity values is larger than  $\mathbf{X}_o + 2\lambda$ . Realizations between these two lines correspond to a “target absent” response in the EM

on the abscissa correspond to the new-item familiarity ( $\mathbf{X}_n$ ). Let us first consider the identity line  $\mathbf{X}_o = \mathbf{X}_n$  that passes through the origin. Points above this line corresponds to a correct identification for both the EM and the IOM, as they imply that the target familiarity values exceeds the lure familiarity value.

The key difference between between the EM and the IOM in this representation is how both models partition this decision space into different response categories for the 1-out-of- $m$  detection sub-task. The IOM assumes that the boundary which separates “target present” from “target absent” decisions is defined by the two line segments  $\mathbf{X}_o = \lambda$  and  $\mathbf{X}_n = \lambda$  that terminate at  $(\lambda, \lambda)$ , which is depicted in the left panel of Fig. 3. Any pair  $\{\mathbf{X}_o, \mathbf{X}_n\}$  such that  $\mathbf{X}_o > \lambda$  or  $\mathbf{X}_n > \lambda$  is detected as a trial containing a target—the larger the value of  $\lambda$ , the greater the confidence level. In contrast, according to the EM, the relevant boundaries are defined by the two lines  $\mathbf{X}_o = \mathbf{X}_n + 2\lambda$  and  $\mathbf{X}_n = \mathbf{X}_o + 2\lambda$ , which is depicted in the right panel of Fig. 3. Any pair  $\{\mathbf{X}_o, \mathbf{X}_n\}$  such that  $\mathbf{X}_o > \mathbf{X}_n + 2\lambda$  or  $\mathbf{X}_n > \mathbf{X}_o + 2\lambda$  is detected as a trial containing a target. There is thus a band centered on the identity line  $\mathbf{X}_o = \mathbf{X}_n$  in which item pairs are not detected as containing a target.

To reiterate, the effect of interest concerns changes in the hit rate if a target is presented alongside a similar lure (with familiarity values  $\mathbf{X}_s$ ) instead of a regular lure (with familiarity value  $\mathbf{X}_n$ ). As long as  $\mu_s > \mu_n$ , the bivariate distribution corresponding to target trials with a similar lure is simply the distribution for a target trial with a regular lure shifted by  $\mu_s - \mu_n$  to the right along the abscissa, which can

be easily seen in Fig. 3. This has the effect of increasing the proportion of realizations that are detected in the IOM, given  $\lambda$  remains unchanged (i.e., an increase in the predicted hit rate; see left panel of Fig. 3). For the EM model, on the other hand, the same rightward shift entails that more probability mass crosses into the non-detection band surrounding the identity line, thereby decreasing the proportion of realizations that lead to detection (i.e., a decrease in the predicted hit rate; see right panel of Fig. 3).

In their empirical investigation of this situation, Meyer-Grant and Klauer (2021) found clear evidence for a higher hit rate when including a similar lure instead of a regular one, regardless of the confidence level.<sup>17</sup> This raises some serious doubts regarding the validity of the central cognitive mechanism proposed by the equal-variance Gaussian EM in particular.

## Reanalysis of Previously Published Data

However, the EM generally not being able to predict this qualitative effect hinges on certain model assumptions. More specifically, if the assumptions regarding variances and/or covariances (as specified in Proposition 2) are relaxed, the model can no longer be decisively ruled out based on the occurrence of the critical effect alone. Moreover,

<sup>17</sup>Even if  $\mu_{\mathbf{X}_s} = \mu_{\mathbf{X}_n}$  is assumed, the prediction that follows in this case (i.e.,  $P_{EM}(H|\lambda, \text{Similar Lure}) = P_{EM}(H|\lambda, \text{Regular Lure})$  for all  $\lambda \in \mathbb{R}^+$ ) is thus rejected.

an unequal-variance assumption, in particular, is rather prominent in Gaussian SDT models of recognition memory (Jang et al., 2009; Mickes et al., 2007; Wixted, 2007; Starns et al., 2012; Spanton & Berry, 2021; but see ; Wixted et al., 2018) and should therefore be considered in the present work. Hence, it seems worthwhile to complement our argument with an additional quantitative model comparison between the unequal-variance Gaussian IOM and the unequal-variance Gaussian EM in order to ensure that our criticism of the EM framework is indeed warranted. We therefore conducted a reanalysis of the data by Meyer-Grant and Klauer (2021) to address this particular issue, mirroring their model comparison strategy.

Thus, we first fitted the parameters of both the unequal-variance Gaussian IOM and the unequal-variance Gaussian EM to the data aggregated across participants using a maximum likelihood approach (with  $\sigma_{X_o}$ ,  $\sigma_{X_s}$ , and  $\sigma_{X_n}$  allowed to differ from each other). Calculating the Pearson's  $\chi^2$  test statistics for both models reveals that the EM ( $\chi^2(17) = 392.71$ ) provides a clearly inferior goodness of fit compared to the IOM ( $\chi^2(17) = 149.17$ ). The top and middle row of panels in Fig. 4 depict the respective model predictions (for the ROC and the IOC as well as the hit probabilities for target trials with and without a similar lure, respectively) together with the corresponding experimental data (Meyer-Grant & Klauer, 2021).

However, analyses of data aggregated across participants neglect possible variability in parameters between participants and are thus susceptible to aggregation biases (Morey et al., 2008; Juola et al., 2019; Smith et al., 2017). We therefore additionally conducted a Bayesian hierarchical model analysis (Rouder & Lu, 2005; Rouder et al., 2017), in which the parameter vector for each participant is separately drawn from a population distribution. Again following the procedure outlined by Meyer-Grant and Klauer (2021), we used a latent-trait approach (Klauer, 2010) and assessed model fit by means of a cross-validation index (CVI; Browne, 2000; Gelman et al., 2013; Vehtari et al., 2017). The selection of (hyper-)prior parameters, the cross validation procedure, and the calculation of the CVI adheres exactly to the specifications in Meyer-Grant and Klauer (2021). We also provide model weights ( $W$ ) calculated through stacking of predictive distributions (Yao et al., 2018). All calculations are repeated two times with different seeds of the random number generator (Meyer-Grant & Klauer, 2021). Results clearly suggest that the IOM ( $M_{CVI} = -43\,749.34$ ,  $SD_{CVI} = 1.38$ ,  $M_W = 0.992$ ,  $SD_W = 0.011$ ) provides a better account of the data than the EM ( $M_{CVI} = -43\,985.90$ ,  $SD_{CVI} = 1.64$ ,  $M_W = 0.008$ ,  $SD_W = 0.011$ ). This corroborates the preliminary conclusion drawn from the analysis of the aggregated data.

Additionally, the bottom row of panels in Fig. 4 depicts the model predictions for the hit probabilities

given a correct identification for different confidence levels together with the corresponding experimental data (Meyer-Grant & Klauer, 2021). This reveals yet another critical effect pattern that distinguishes the two model frameworks: Including a similar lure that does not resemble the target of the current lineup appears to increase the hit rate even when conditioning on a correct identification response (i.e., the confidence in a correct identification increases), as opposed to just increasing the overall hit rate regardless of the identification response.<sup>18</sup> The IOM predicts such an effect (see bottom left panel of Fig. 4) since targets selected over similar lures should tend to be associated with a stronger absolute memory-strength signal than targets selected over regular lures, whereas the EM is once more unable to adequately account for this pattern (see bottom right panel of Fig. 4).

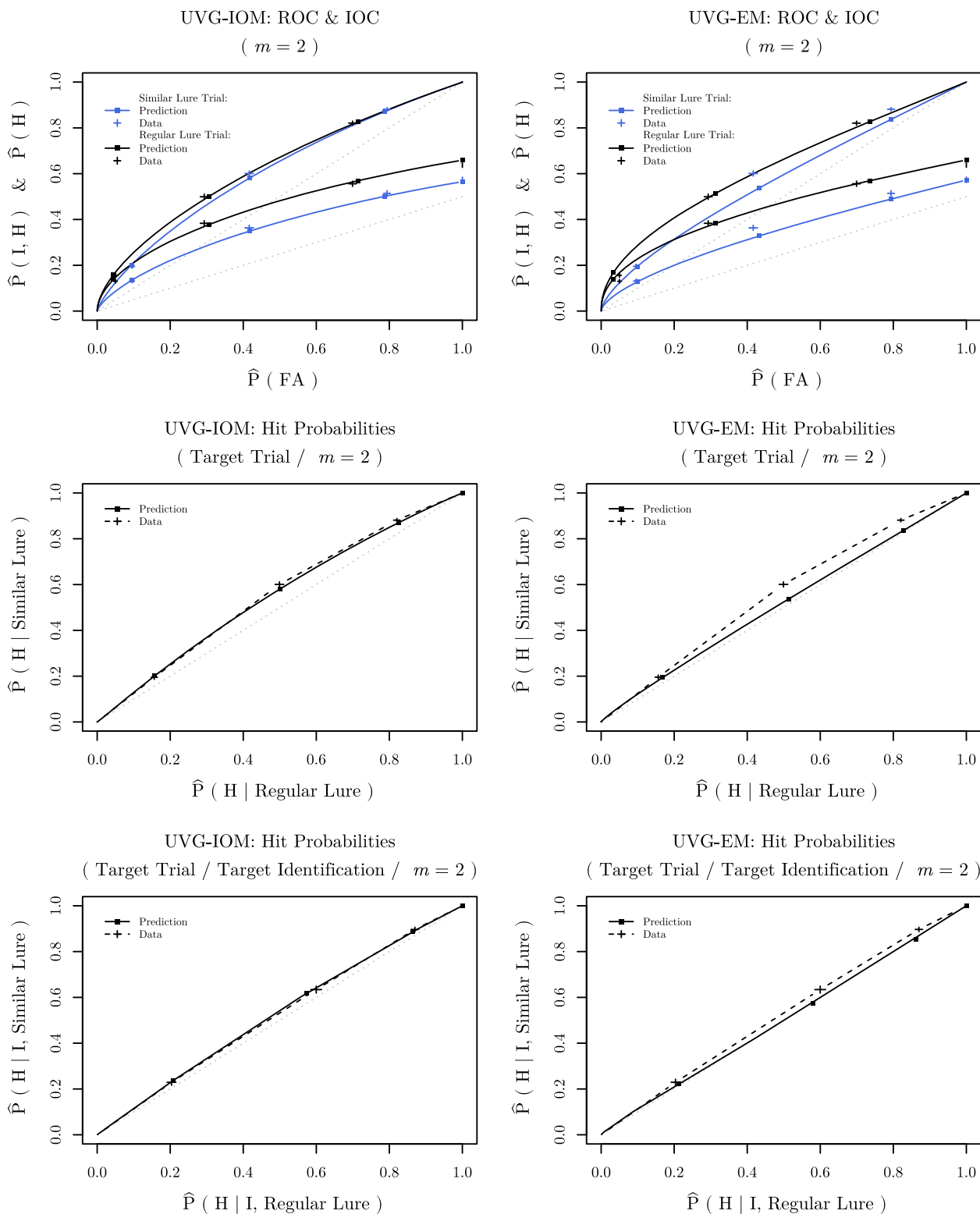
In order to further test this effect, we conducted three generalized linear mixed model analyses (Singmann & Kellen, 2019) with a logistic link function comparing the relative frequencies for hits given a correct identification response between target trials with and without a similar lure. We therefore included the fixed-effect within-subject factor “similar lure” (present vs. absent) as well as crossed random-effect factors for participants as well as target materials (Judd et al., 2012). The random effects structure was determined by a backwards selection (for details, see Meyer-Grant & Klauer, 2021).<sup>19</sup> Each of the three analyses addressed one of three different confidence levels. The first analysis only treated responses with the highest confidence as hits, the second one additionally treated responses with the second highest confidence as hits, and the third one treated all but the lowest confidence rating as hits. In all three analyses, the hit rate was significantly larger in trials with a similar lure (results are reported in Table 1; for a pictorial representation, see Fig. 5).

## Discussion

In the present work, we have demonstrated that two aspects of between-item similarity, namely old–new similarity and within-lineup similarity, must be treated as distinct concepts and should therefore be separately modeled. By deriving conflicting predictions between the equal-variance Gaussian EM and many IOMs for situations in which old–new similarity is selectively manipulated and  $m = 2$ , we have, furthermore, shown that past empirical results (Meyer-Grant

<sup>18</sup>We thank Jeffrey Starns for suggesting to investigate this effect.

<sup>19</sup>The final random effects structure for all three analyses included random intercepts for participants and targets, whereas the final random effects structure for the two analyses described last also included by-target random slopes for the factor “similar lure.”



**Fig. 4** Data (dashed lines and crosses; lengths of cross lines correspond to 95% bootstrap CIs) of the experiment conducted by Meyer-Grant and Klauer (2021) and the respective model predictions (solid lines and squares) of the best fitting unequal-variance Gaussian IOM (UVG-IOM; left column of panels) and the best fitting unequal-variance Gaussian EM (UVG-EM; right column of panels). Upper row of panels depicts ROCs and IOCs for trials with (blue) and without

(black) a similar lure. Dotted gray lines represent guessing level. Middle row of panels depicts the probability of a hit in target trials with a similar lure on the ordinate and in target trials with a regular lure on the abscissa. Bottom row of panels depicts the probability of a hit conditional on a correct identification in target trials with a similar lure on the ordinate and in target trials with a regular lure on the abscissa. Dotted gray line is the identity line

**Table 1** Results of the three generalized linear mixed model analyses (accounting for different confidence levels) of the hit rate differences between trials with and without a similar lure

Hit	$\widehat{P}(\text{H} I, \text{Similar Lure})$	$\widehat{P}(\text{H} I, \text{Regular Lure})$	$\ln(OR)$	95% CI	$\chi^2(1)$	$p$
Conservative	20.10%	17.51%	0.17	[0.05, 0.28]	8.02	.009
Medium	63.74%	60.11%	0.15	[0.04, 0.27]	7.14	.009
Liberal	92.48%	90.02%	0.31	[0.16, 0.46]	14.89	<.001

All  $p$ -values are Bonferroni–Holm corrected for three test. The log-odds ratio is denoted by  $\ln(OR)$ .

& Klauer, 2021) are irreconcilable with the equal-variance Gaussian EM.

Although we have not elaborated on it thus far, this critical test applies not only for a model parametrization in terms of equal-variance normal distributions, but also to other EMs in cases where the likelihood ratio of similar and regular lure familiarities is a monotonically increasing function (see the alternative proof of Proposition 2 in the Appendix). This is arguably a quite reasonable and defensible assumption to make as it is equivalent to a low familiarity value being “more likely under the familiarity distribution of regular than under the familiarity distribution of similar lures, whereas for high familiarity values it is the other way around” (Meyer-Grant & Klauer, 2021, p.9; see also Kellen et al., 2021; Rouder et al., 2021; Green & Swets, 1966).

A reanalysis of the data from Meyer-Grant and Klauer (2021) additionally revealed that even when relaxing the equal-variance assumption,<sup>20</sup> the EM is still clearly outperformed by the IOM in terms of quantitative model fit.

It is also noteworthy that the key message of the present work does extend beyond the class of SDT models of recognition memory. One could, for example, imagine a similar mechanism within so-called *threshold models* (see, e.g., Kellen & Klauer, 2018; Coombs et al., 1970; Bernbach, 1967). In contrast to SDT models, threshold models assume that the latent memory signal is not directly accessible by the decision-maker. Instead, these models propose mediating mental states of discrete nature (such as “correctly detecting a target to be old” or “correctly detecting a lure to be new”), which can be reached with a certain probability that depends on whether the stimulus is old or new.

Recently, it has been shown that threshold models—in particular the so-called *two-high-threshold model* (Snodgrass & Corwin, 1988)—cannot explain certain response patterns when manipulating memory strength and presenting multiple stimuli simultaneously during test (Kellen et al., 2021; Kellen & Klauer, 2014). This is the case

because—according to the standard interpretation of the two-high-threshold model—a memory-strength manipulation only affects the probability of correctly detecting a target. However, the model predictions for the response frequencies of interest do not depend on this very probability, but only on the probability of correctly detecting a lure. To avoid these problems, one could easily extend the two-high-threshold model to operate in a similar way as the EM, that is, by allowing detection probabilities to depend on the specific context of presented stimuli. This results in altering the probability of a correct lure detection when the memory strength of the target is manipulated (e.g., the stronger the target’s memory signal, the easier it is to detect a lure).

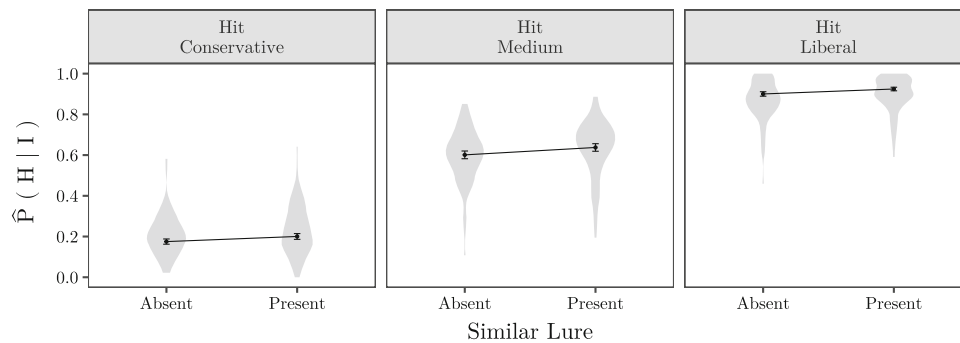
However, this idea falls victim to the same observation as the SDT version of the EM, since increasing the old–new similarity would then decrease the probability of correctly detecting a target, which in turn would tend to lower the hit rate in such a situation. But this again directly contradicts the observations reported by Meyer-Grant and Klauer (2021). Therefore, our results also indirectly strengthen the arguments by Kellen and Klauer (2014) and Kellen et al. (2021).

There is, however, a caveat that concerns the applicability of our findings to research primarily focused on eyewitness identification: The experimental paradigm used to obtain the data on which our argument is based (Meyer-Grant & Klauer, 2021) differs from the paradigm commonly used in that field. More precisely, the data under consideration here are obtained via a *multiple-trial experiment* (i.e., testing each participants on multiple lineups), whereas most research on eyewitnesses identification relies on *single-trial experiments* (i.e., each participant is only tested on one lineup; Mansour et al., 2017).

On the one hand, a multiple-trial design permits analyzing the data by means of Bayesian hierarchical modeling, which helps to avoid aggregation biases that may be present but remain unnoticed in single-trial data. On the other hand, in the task presented here, certain pairs of items that participants encounter will tend to elicit two relatively high familiarity values (i.e., target trials with a similar lure). One could argue that in such a situation a comparative strategy (as instantiated by the EM) is not advisable, as it would lead to many misses in this scenario. Thus, the test

<sup>20</sup>Note that an unequal-variance Gaussian SDT model with  $\sigma_{X_n} \neq \sigma_{X_s}$  does not exhibit a monotonic likelihood ratio of similar and regular lure familiarities.





**Fig. 5** Mean relative frequencies (black dots) of a hit given a correct identification in the absence and presence of a similar lure. In the left panel only the highest confidence level is treated as a hit (hit conservative), in the middle panel the highest and second highest confidence

levels are treated as a hit (hit medium), and in the right panel all but the lowest confidence level are treated as a hit (hit liberal). Error bars depict  $\pm 1SE$  (model based) and the violin plots depict kernel densities of individual relative frequencies

environment may encourage an absolute strategy (as instantiated by the IOM). This would imply that the results may not generalize to single-trial situations, which might encourage other strategies. Future research may address this question by examining these effects in single-trial experiments.

Another aspect of the experiment conducted by Meyer-Grant and Klauer (2021) that differs from most lineup procedures used in forensic practice, is the number of simultaneously presented items (typically,  $m > 2$  for real-life lineups). Although it seems at least reasonable to assume that the basic underlying memory mechanisms do not depend on lineup size, this conjecture may be verified in the future by quantitative model comparisons for larger lineups.

Note also that the focus of the present work was on SDAI, that is, on simultaneous lineup procedures. However, the general theoretical ideas underlying the modeling framework of the EM and the IOM can be extended to sequential lineups as well (see, e.g., Dunn et al., 2022; Kaesler et al., 2020).

Despite these limitations, we are inclined to conclude that the central mechanism of the EM can be questioned on the basis of our results, especially in light of the fact that other pieces of evidence seemingly favoring the EM over the IOM are inconclusive upon closer inspection (Wixted et al., 2018; Akan et al., 2021). In fact, the IOM appears to be the only SDAI extension of the SDT model framework which can account for all observations discussed in the literature on lineup memory to date.

However, this does not necessarily imply that this model is uncontested. It has, for example, been argued that a *low-threshold model* (Kellen et al., 2016; Luce, 1963; Starns, 2020) of SDAI could be a viable competitor to IOMs based on SDT (Meyer-Grant & Klauer, 2021). Moreover, unequal-variance Gaussian SDT models are themselves known to have certain properties that are at least controversial. For instance, they do not possess monotone likelihood ratios, which is—as already mentioned above—often considered

to be implausible (e.g., Green & Swets, 1966; Kellen & Klauer, 2018; Kellen et al., 2021). Thus, while the present work contributes to the debate on what qualifies as a reasonable modeling approach for SDAI by questioning the validity of one prominent candidate model (viz., the EM), further research is clearly necessary to validate and/or refine the models of recognition memory used to study lineup memory.

## Appendix

*Proof of Theorem 1* First, we rearrange terms in Eq. 2 such that

$$\tau_H(\lambda) := \tau_{H(m=2)}(\lambda) = \frac{1}{1 + \frac{\Phi(-\lambda - \mu_{D_o})}{1 - \Phi(\lambda - \mu_{D_o})}}.$$

Clearly, for  $\tau_H(\lambda)$  to be a monotonically increasing function, it suffices to show that

$$\begin{aligned} \mathcal{T}_1(\lambda) &:= \frac{\Phi(-\lambda - \mu_{D_o})}{1 - \Phi(\lambda - \mu_{D_o})} = \frac{1 - \Phi(\lambda + \mu_{D_o})}{1 - \Phi(\lambda - \mu_{D_o})} \\ &= \frac{\int_{\lambda}^{\infty} \varphi(z + \mu_{D_o}) dz}{\int_{\lambda}^{\infty} \varphi(z - \mu_{D_o}) dz} \end{aligned}$$

is monotonically decreasing. Taking the derivative of  $\mathcal{T}_1(\lambda)$  with respect to  $\lambda$  yields

$$\mathcal{T}'_1(\lambda) = \frac{\varphi(\lambda - \mu_{D_o}) \int_{\lambda}^{\infty} \varphi(z + \mu_{D_o}) dz - \varphi(\lambda + \mu_{D_o}) \int_{\lambda}^{\infty} \varphi(z - \mu_{D_o}) dz}{\left( \int_{\lambda}^{\infty} \varphi(z - \mu_{D_o}) dz \right)^2}. \quad (4)$$

The denominator in Eq. 4 is always positive and will therefore be ignored in the following. Thus,  $\mathcal{T}_1(\lambda)$  is strictly monotonically decreasing if and only if

$$\varphi(\lambda - \mu_{D_o}) \int_{\lambda}^{\infty} \varphi(z + \mu_{D_o}) dz - \varphi(\lambda + \mu_{D_o}) \int_{\lambda}^{\infty} \varphi(z - \mu_{D_o}) dz$$

$$\begin{aligned}
 &= \frac{1}{\sqrt{2\pi}} \int_{\lambda}^{\infty} \exp\left(-\frac{1}{2}\left((\lambda - \mu_{\mathbf{D}_o})^2 + (z + \mu_{\mathbf{D}_o})^2\right)\right) \\
 &\quad - \exp\left(-\frac{1}{2}\left((\lambda + \mu_{\mathbf{D}_o})^2 + (z - \mu_{\mathbf{D}_o})^2\right)\right) dz \\
 &= \frac{1}{\sqrt{2\pi}} \int_{\lambda}^{\infty} \exp\left(-\frac{1}{2}\left(\lambda^2 + z^2 + 2(\mu_{\mathbf{D}_o})^2\right)\right) \\
 &\quad \left(\exp(\mu_{\mathbf{D}_o}(\lambda - z)) - \exp(\mu_{\mathbf{D}_o}(z - \lambda))\right) dz < 0 \quad (5)
 \end{aligned}$$

for all  $\lambda \in (0, \infty)$ . Since it holds that

$$\exp\left(-\frac{1}{2}\left(\lambda^2 + z^2 + 2(\mu_{\mathbf{D}_o})^2\right)\right) \geq 0$$

and  $\lambda - z < z - \lambda$  for all  $z > \lambda$ , the inequality in Eq. 5 must hold if  $\mu_{\mathbf{D}_o} > 0$ . Thus,  $\tau_{\mathbf{H}}(\lambda)$  is monotonically increasing in  $\mathbb{R}^+$ .

To show that

$$\begin{aligned}
 \tau_{\mathbf{M}}(\lambda) &:= \tau_{\mathbf{M}(m=2)}(\lambda) \\
 &= \frac{\Phi(\lambda - \mu_{\mathbf{D}_o}) - \Phi(-\mu_{\mathbf{D}_o})}{\Phi(\lambda - \mu_{\mathbf{D}_o}) - \Phi(-\mu_{\mathbf{D}_o}) + \Phi(-\mu_{\mathbf{D}_o}) - \Phi(-\lambda - \mu_{\mathbf{D}_o})} \\
 &= \frac{1}{1 + \frac{\Phi(-\mu_{\mathbf{D}_o}) - \Phi(-\lambda - \mu_{\mathbf{D}_o})}{\Phi(\lambda - \mu_{\mathbf{D}_o}) - \Phi(-\mu_{\mathbf{D}_o})}}.
 \end{aligned}$$

is a monotonically increasing function, it suffices to show that

$$\begin{aligned}
 \mathcal{T}_2(\lambda) &:= \frac{\Phi(-\mu_{\mathbf{D}_o}) - \Phi(-\lambda - \mu_{\mathbf{D}_o})}{\Phi(\lambda - \mu_{\mathbf{D}_o}) - \Phi(-\mu_{\mathbf{D}_o})} \\
 &= \frac{1 - \Phi(\mu_{\mathbf{D}_o}) - (1 - \Phi(\lambda + \mu_{\mathbf{D}_o}))}{\Phi(\lambda - \mu_{\mathbf{D}_o}) - \Phi(-\mu_{\mathbf{D}_o})} \\
 &= \frac{\Phi(\lambda + \mu_{\mathbf{D}_o}) - \Phi(\mu_{\mathbf{D}_o})}{\Phi(\lambda - \mu_{\mathbf{D}_o}) - \Phi(-\mu_{\mathbf{D}_o})} \\
 &= \frac{\int_0^{\lambda} \varphi(z + \mu_{\mathbf{D}_o}) dz}{\int_0^{\lambda} \varphi(z - \mu_{\mathbf{D}_o}) dz}
 \end{aligned}$$

is monotonically decreasing. Taking the derivative of  $\mathcal{T}_2(\lambda)$  with respect to  $\lambda$  yields

$$\mathcal{T}'_2(\lambda) = \frac{\varphi(\lambda + \mu_{\mathbf{D}_o}) \int_0^{\lambda} \varphi(z - \mu_{\mathbf{D}_o}) dz - \varphi(\lambda - \mu_{\mathbf{D}_o}) \int_0^{\lambda} \varphi(z + \mu_{\mathbf{D}_o}) dz}{\left(\int_0^{\lambda} \varphi(z - \mu_{\mathbf{D}_o}) dz\right)^2}. \quad (6)$$

The denominator in Eq. 6 is always positive and will therefore be ignored in the following. Thus,  $\mathcal{T}_2(\lambda)$  is strictly monotonically decreasing if and only if

$$\varphi(\lambda + \mu_{\mathbf{D}_o}) \int_0^{\lambda} \varphi(z - \mu_{\mathbf{D}_o}) dz - \varphi(\lambda - \mu_{\mathbf{D}_o}) \int_0^{\lambda} \varphi(z + \mu_{\mathbf{D}_o}) dz$$

$$\begin{aligned}
 &= \frac{1}{\sqrt{2\pi}} \int_{\lambda}^{\infty} \exp\left(-\frac{1}{2}\left((\lambda + \mu_{\mathbf{D}_o})^2 + (z - \mu_{\mathbf{D}_o})^2\right)\right) \\
 &\quad - \exp\left(-\frac{1}{2}\left((\lambda - \mu_{\mathbf{D}_o})^2 + (z + \mu_{\mathbf{D}_o})^2\right)\right) dz \\
 &= \frac{1}{\sqrt{2\pi}} \int_{\lambda}^{\infty} \exp\left(-\frac{1}{2}\left(\lambda^2 + z^2 + 2(\mu_{\mathbf{D}_o})^2\right)\right) \\
 &\quad \left(\exp(\mu_{\mathbf{D}_o}(z - \lambda)) - \exp(\mu_{\mathbf{D}_o}(\lambda - z))\right) dz < 0 \quad (7)
 \end{aligned}$$

for all  $\lambda \in (0, \infty)$ . Since again

$$\exp\left(-\frac{1}{2}\left(\lambda^2 + z^2 + 2(\mu_{\mathbf{D}_o})^2\right)\right) \geq 0$$

and  $z - \lambda < \lambda - z$  for all  $z < \lambda$ , the inequality in Eq. 7 must hold if  $\mu_{\mathbf{D}_o} > 0$ . Therefore,  $\tau_{\mathbf{M}}(\lambda)$  is also monotonically increasing in  $\mathbb{R}^+$ .  $\square$

*Proof of Proposition 2* First, we note that

$$\begin{aligned}
 &\sqrt{\frac{\sigma_{\mathbf{X}_o}^2 + \sigma_{\mathbf{X}_s}^2 - 2\text{Cov}(\mathbf{X}_o, \mathbf{X}_s)}{4}} \\
 &= \sqrt{\frac{\sigma_{\mathbf{X}_o}^2 + \sigma_{\mathbf{X}_n}^2 - 2\text{Cov}(\mathbf{X}_o, \mathbf{X}_n)}{4}} = \sigma_{\mathbf{D}_o}
 \end{aligned}$$

and that we can again set  $\sigma_{\mathbf{D}_o} = 1$  without loss of generality.

It then follows from Eq. 1 that Eq. 3 is equivalent to

$$\begin{aligned}
 &1 - \Phi\left(\lambda - \frac{\mu_{\mathbf{X}_o} - \mu_{\mathbf{X}_s}}{2}\right) + \Phi\left(-\lambda - \frac{\mu_{\mathbf{X}_o} - \mu_{\mathbf{X}_s}}{2}\right) \\
 &\leq 1 - \Phi\left(\lambda - \frac{\mu_{\mathbf{X}_o} - \mu_{\mathbf{X}_n}}{2}\right) + \Phi\left(-\lambda - \frac{\mu_{\mathbf{X}_o} - \mu_{\mathbf{X}_n}}{2}\right).
 \end{aligned}$$

Because  $\mu_{\mathbf{X}_o} > \mu_{\mathbf{X}_s} > \mu_{\mathbf{X}_n} = 0$ , it must hold that

$$\frac{\mu_{\mathbf{X}_o} - \mu_{\mathbf{X}_n}}{2} > \frac{\mu_{\mathbf{X}_o} - \mu_{\mathbf{X}_s}}{2} > 0.$$

Thus, it suffices to show that

$$\begin{aligned}
 &\frac{\partial}{\partial \mu} (1 - \Phi(\lambda - \mu) + \Phi(-\lambda - \mu)) \\
 &= \frac{1}{\sqrt{2\pi}} \left( \exp\left(-\frac{1}{2}(\lambda - \mu)^2\right) - \exp\left(-\frac{1}{2}(-\lambda - \mu)^2\right) \right) \\
 &\quad \exp\left(-\frac{1}{2}(\lambda^2 - \mu^2)\right) \\
 &= \frac{\exp(\lambda\mu) - \exp(-\lambda\mu)}{\sqrt{2\pi}} \geq 0,
 \end{aligned}$$

which is clearly the case since  $\lambda\mu > -\lambda\mu$  and

$$\frac{\exp\left(-\frac{1}{2}(\lambda^2 - \mu^2)\right)}{\sqrt{2\pi}} \geq 0$$

if  $\lambda > 0$  as well as  $\mu > 0$ .  $\square$

*Alternative proof of Proposition 2* First, let  $f_{\mathbf{X}_o}$ ,  $f_{\mathbf{X}_s}$ , and  $f_{\mathbf{X}_n}$  denote the PDFs of target, similar lure and regular lure familiarities (i.e., the RVs  $\mathbf{X}_o$ ,  $\mathbf{X}_s$ , and  $\mathbf{X}_n$  with the same

support  $\Omega = \mathbb{R}$ ), respectively. Furthermore, let  $F_{\mathbf{X}_s}$  and  $F_{\mathbf{X}_n}$  denote the CDFs of  $\mathbf{X}_s$  and  $\mathbf{X}_n$ , respectively.

If we assume independence, then the complementary CDF (i.e., the tail distribution function) of the difference of two independent RVs, for instance,  $\mathbf{X}_o$  and  $\mathbf{X}_n$  (i.e., the RV  $\mathbf{X}_o - \mathbf{X}_n$ ), can be written as

$$\begin{aligned}
 1 - F_{\{\mathbf{X}_o - \mathbf{X}_n\}}(\lambda) &= \iint_{\mathbb{R}^2} f_{\mathbf{X}_o}(x_o) f_{\mathbf{X}_n}(x_n) \mathbb{1}_{\{(x_o, x_n) \in \mathbb{R}^2 : x_o - x_n > \lambda\}}(x_o, x_n) dx_n dx_o \\
 &= \int_{\mathbb{R}} \int_{-\infty}^{x_o - \lambda} f_{\mathbf{X}_o}(x_o) f_{\mathbf{X}_n}(x_n) dx_n dx_o \\
 &= \int_{\mathbb{R}} F_{\mathbf{X}_n}(x_o - \lambda) f_{\mathbf{X}_o}(x_o) dx_o, \tag{8}
 \end{aligned}$$

where  $\mathbb{1}$  denotes the indicator function. If  $\Omega \subset \mathbb{R}$ , then we apply a strictly monotonic transformation  $\mathcal{G} : \Omega \rightarrow \mathbb{R}$  to  $\mathbf{X}_o$ ,  $\mathbf{X}_s$ , and  $\mathbf{X}_n$  (i.e., a change of variable for all familiarity distributions, such that the support of the transformed RVs becomes  $\mathbb{R}$ ). One immediately sees that Eq. 8 is the probability of a hit as predicted by a nonparametric EM in cases where  $m = 2$ , both item familiarities are independent, and a regular lure is presented alongside the target.

By exchanging  $f_{\mathbf{X}_n}$  and  $F_{\mathbf{X}_n}$  with  $f_{\mathbf{X}_s}$  and  $F_{\mathbf{X}_s}$ , respectively, in Eq. 8, we also find that

$$1 - F_{\{\mathbf{X}_o - \mathbf{X}_s\}}(\lambda) = \int_{\mathbb{R}} F_{\mathbf{X}_s}(x_o - \lambda) f_{\mathbf{X}_o}(x_o) dx_o$$

is the predicted probability analogous to Eq. 8, when instead of a regular lure a similar lure is presented (still assuming that its familiarity is independent from the target familiarity).

If we further assume first order stochastic dominance of similar and regular lure familiarities (i.e.,  $F_{\mathbf{X}_s}(x) \leq F_{\mathbf{X}_n}(x)$  for all  $x \in \mathbb{R}$ ), we easily see that

$$\begin{aligned}
 \int_{\mathbb{R}} F_{\mathbf{X}_s}(x_o - \lambda) f_{\mathbf{X}_o}(x_o) dx_o &\leq \int_{\mathbb{R}} F_{\mathbf{X}_n}(x_o - \lambda) f_{\mathbf{X}_o}(x_o) dx_o \\
 \Leftrightarrow P_{EM}(H|\lambda, \text{Similar Lure}) &\leq P_{EM}(H|\lambda, \text{Regular Lure}) \tag{9}
 \end{aligned}$$

indeed holds for all  $\lambda \in \mathbb{R}^+$ .

It is well known that a monotone likelihood ratio entails first order stochastic dominance (Bapat & Kochar, 1994). Consequently, all distributions exhibiting a monotone likelihood ratio of similar and regular lure familiarity distributions (i.e.,  $f_{\mathbf{X}_s}(x)/f_{\mathbf{X}_n}(x)$  is a monotonically increasing function for all  $x \in \mathbb{R}$ ) also exhibit first order stochastic dominance of the similar and regular lure familiarity distributions (i.e.,  $F_{\mathbf{X}_s}(x) \leq F_{\mathbf{X}_n}(x)$  for all  $x \in \mathbb{R}$ ).

Equal-variance Gaussian SDT models indeed exhibit this monotone likelihood ratio and hence also first

order stochastic dominance of similar and regular lure familiarity distributions if  $\mu_{\mathbf{X}_s} > \mu_{\mathbf{X}_n}$ . Furthermore, for a bivariate Gaussian distribution, the absence of a within-lineup correlation is equivalent to both RVs being independent. Therefore, the inequality in Eq. 9 holds for the equal-variance Gaussian EM, if there is no within-lineup correlation.  $\square$

**Acknowledgements** We would like to thank Andrew Heathcote for motivating the present work.

**Authors Contribution** Constantin G. Meyer-Grant had the idea for the present work, conceived the line of argument, derived the mathematical formalization of the models as well as the proofs of Theorem 1 and Proposition 2, conducted the analyses, prepared all the plots, and drafted the original manuscript. Karl Christoph Klauer provided a vital step for the proof of Theorem 1, checked all mathematical derivations, and edited the original manuscript. Both authors read and approved to the final version of the manuscript.

**Funding** Open Access funding enabled and organized by Projekt DEAL. Constantin G. Meyer-Grant received support from the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation), GRK 2277 ‘‘Statistical Modeling in Psychology.’’

**Data and Code Availability** The data analyzed in the present work were first published by Meyer-Grant and Klauer (2021) and are available in an Open Science Framework repository (<https://osf.io/3qnggh/>).

The R code (R Core Team, 2021) as well as the Stan models (Carpenter et al., 2017) for the reanalyses in terms of the Bayesian hierarchical model comparison are likewise available in an Open Science Framework repository (<https://osf.io/9tyx3/>).

## Declarations

**Ethical Approval** Not applicable as no original data from participants were collected.

**Consent to Participate and Publish** Not applicable as no original data from participants were collected.

**Competing Interests** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

Akan, M., Robinson, M. M., Mickes, L., Wixted, J. T., & Benjamin, A. S. (2021). The effect of lineup size on eyewitness identification. *Journal of Experimental Psychology: Applied*, 27(2), 369–392.

- Bapat, R., & Kochar, S. C. (1994). On likelihood-ratio ordering of order statistics. *Linear Algebra and its Applications*, *199*, 281–291.
- Bernbach, H. A. (1967). Decision processes in memory. *Psychological Review*, *74*(6), 462–480.
- Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear – or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 587–606.
- Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*(1), 108–132.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.
- Charman, S. D., & Wells, G. L. (2007). Applied lineup theory. 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. 219–254). Lawrence Erlbaum Associates.
- Charman, S. D., Wells, G. L., & Joy, S. W. (2011). The dud effect: Adding highly dissimilar fillers increases confidence in lineup identifications. *Law and Human Behavior*, *35*(6), 479–500.
- Clark, S. E., Erickson, M. A., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior*, *35*(5), 364–380.
- Cohen, A. L., Starns, J. J., & Rotello, C. M. (2021). sdtlu: An R package for the signal detection analysis of eyewitness lineup data. *Behavior Research Methods*, *53*(1), 278–300.
- Colloff, M. F., Wade, K. A., Wixted, J. T., & Maylor, E. A. (2017). A signal-detection analysis of eyewitness identification across the adult lifespan. *Psychology and Aging*, *32*(3), 243–258.
- Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Prentice-Hall.
- Dubé, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 130–151.
- Dunn, J. C., Kaesler, M., & Semmler, C. (2022). A model of position effects in the sequential lineup. *Journal of Memory and Language*, *122*, 104297.
- Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology*, *67*(5), 818–835.
- Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis. *Psychology, Public Policy, and Law*, *19*(2), 151–164.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC Press.
- Glanzer, M., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review*, *16*(3), 431–455.
- Green, D. M., & Birdsall, T. G. (1978). Detection and recognition. *Psychological Review*, *85*(3), 192–206.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.
- Green, D. M., Weber, D. L., & Duncan, J. E. (1977). Detection and recognition of pure tones in noise. *The Journal of the Acoustical Society of America*, *62*(4), 948–954.
- Hanczakowski, M., Zawadzka, K., & Higham, P. A. (2014). The dud-alternative effect in memory for associations: Putting confidence into local context. *Psychonomic Bulletin & Review*, *21*(2), 543–548.
- Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101.
- Hintzman, D. L. (2001). Similarity, global matching, and judgments of frequency. *Memory & Cognition*, *29*(4), 547–556.
- Horry, R., & Brewer, N. (2016). How target–lure similarity shapes confidence judgments in multiple-alternative decision tasks. *Journal of Experimental Psychology: General*, *145*(12), 1615–1634.
- Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*(2), 291–306.
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69.
- Juola, J. F., Caballero-Sanz, A., Muñoz-García, A. R., Botella, J., & Suero, M. (2019). Familiarity, recollection, and receiver-operating characteristic (ROC) curves in recognition memory. *Memory & Cognition*, *47*(4), 855–876.
- Kaesler, M., Dunn, J. C., Ransom, K., & Semmler, C. (2020). Do sequential lineups impair underlying discriminability? *Cognitive Research: Principles and Implications*, *5*(1), 1–21.
- Kellen, D., Erdfelder, E., Malmberg, K. J., Dubé, C., & Criss, A. H. (2016). The ignored alternative: An application of Luce’s low-threshold model to recognition memory. *Journal of Mathematical Psychology*, *75*, 86–95.
- Kellen, D., & Klauer, K. C. (2014). Discrete-state and continuous models of recognition memory: Testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1795–1804.
- Kellen, D., & Klauer, K. C. (2018). Elementary signal detection and threshold theory. In E.-J. Wagenmakers, & J. T. Wixted (Eds.) *The Stevens handbook of experimental psychology and cognitive neuroscience* (4th ed., Vol. 5 pp. 161–200). Wiley.
- Kellen, D., Winiger, S., Dunn, J. C., & Singmann, H. (2021). Testing the foundations of signal detection theory in recognition memory. *Psychological Review*, *128*(6), 1022–1050.
- Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*(1), 70–98.
- Lee, J., & Penrod, S. D. (2019). New signal detection theory-based framework for eyewitness performance in lineups. *Law and Human Behavior*, *43*(5), 436.
- Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, *70*(1), 61–79.
- Luus, C. A., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behavior*, *15*(1), 43–57.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user’s guide* (2nd ed.). Earlbaum.
- Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology*, *57*(4), 335–384.
- Mansour, J. K., Beaudry, J. L., & Lindsay, R. C. L. (2017). Are multiple-trial experiments appropriate for eyewitness identification studies? Accuracy, choosing, and confidence across trials. *Behavior Research Methods*, *49*(6), 2235–2254.
- Meyer-Grant, C. G., & Klauer, K. C. (2021). Monotonicity of rank order probabilities in signal detection models of simultaneous detection and identification. *Journal of Mathematical Psychology*, *105*, 102615.
- Mickes, L., & Gronlund, S. D. (2017). Eyewitness identification. In J. H. Byrne, & J. T. Wixted (Eds.) *Learning and memory: A*

- comprehensive reference* (2nd ed., Vol. 2 pp. 529–552). Academic Press.
- Mickes, L., Seale-Carlisle, T. M., Wetmore, S. A., Gronlund, S. D., Clark, S. E., Carlson, C. A., Goodsell, C. A., Weatherford, D., & Wixted, J. T. (2017). ROCs in eyewitness identification: Instructions versus confidence ratings. *Applied Cognitive Psychology, 31*(5), 467–477.
- Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review, 14*(5), 858–865.
- Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology, 52*(6), 376–388.
- Morrell, H. E. R., Gaitan, S., & Wixted, J. T. (2002). On the nature of the decision axis in signal-detection-based models of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(6), 1095–1110.
- Norman, D. A., & Wickelgren, W. A. (1969). Strength theory of decision rules and latency in retrieval from short-term memory. *Journal of Mathematical Psychology, 6*(2), 192–208.
- Parks, C. M., & Yonelinas, A. P. (2008). Theories of recognition memory. In J. H. Byrne, & H. L. Roediger (Eds.) *Learning and memory: A comprehensive reference* (1st ed., Vol. 2 pp. 389–416). Academic Press.
- R Core Team (2021). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Rotello, C. M. (2017). Signal detection theories of recognition memory. In J. H. Byrne, & J. T. Wixted (Eds.) *Learning and memory: A comprehensive reference* (2nd ed., Vol. 2 pp. 529–552). Academic Press.
- Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*(4), 573–604.
- Rouder, J. N., Morey, R. D., & Pratte, M. S. (2017). Bayesian hierarchical models of cognition. In W. H. Batchelder, H. Colonius, E. N. Dzhafarov, & J. Myung (Eds.) *New handbook of mathematical psychology: Foundations and methodology* (pp. 504–551). Cambridge University Press.
- Rouder, J. N., Province, J. M., Swagman, A. R., & Thiele, J.E. (2014). *From ROC curves to psychological theory*. Manuscript submitted for publication.
- Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. H. Spieler, & E. Schumacher (Eds.) *New methods in cognitive psychology* (1st ed., pp. 4–31). Routledge.
- 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.
- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*(1), 34–50.
- Spanton, R. W., & Berry, C. J. (2021). *Variability in recognition memory scales with mean memory strength: Implications for the encoding variability hypothesis*, PsyArXiv Preprint. <https://psyarxiv.com/3sbnh>.
- Starns, J. J. (2020). High-and low-threshold models of the relationship between response time and confidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(4), 671–684.
- Starns, J. J., Rotello, C. M., & Ratcliff, R. (2012). Mixing strong and weak targets provides no evidence against the unequal-variance explanation of zROC slope: A comment on Koen and Yonelinas (2010). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(3), 793–801.
- Starr, S. J., Metz, C. E., Lusted, L. B., & Goodenough, D. J. (1975). Visual detection and localization of radiographic images. *Radiology, 116*(3), 533–538.
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432.
- Voormann, A., Spektor, M. S., & Klauer, K. C. (2021). The simultaneous recognition of multiple words: A process analysis. *Memory & Cognition, 49*(4), 787–802.
- 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.
- Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press.
- Windschitl, P. D., & Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 198–215.
- Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*(1), 152–176.
- Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121*(2), 262–276.
- Wixted, J. T., & Mickes, L. (2018). Theoretical vs. empirical discriminability: The application of ROC methods to eyewitness identification. *Cognitive Research: Principles and Implications, 3*(1), 1–22.
- Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. M. (2018). Models of lineup memory. *Cognitive Psychology, 105*, 81–114.
- Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis, 13*(3), 917–1007.

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