Improving lineup effectiveness through manipulation of eyewitness judgment strategies

by

Eric Y. Mah
BA, from Kwantlen Polytechnic University, 2016

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Psychology

Improving lineup effectiveness through manipulation of eyewitness judgment strategies

by

Eric Y. Mah
BA, from Kwantlen Polytechnic University, 2016

Supervisory Committee

Dr. D. Stephen Lindsay, Supervisor
Department of Psychology

Dr. John Sakaluk, Departmental Member
Department of Psychology

Dr. Michelle Lawrence, Outside Member
Faculty of Law

Abstract

Understanding eyewitness lineup judgment processes is critical, both from a theoretical standpoint (to better understand human memory) and from a practical one (to prevent wrongful convictions and criminals walking free). Currently, two influential theories attempt to explain lineup decision making: the theory of eyewitness judgment strategies (Lindsay & Wells, 1985), and the signal detection theory-informed diagnostic-feature-detection hypothesis (Wixted & Mickes, 2014). The theory of eyewitness judgment strategies posits that eyewitnesses can adopt either an absolute judgment strategy (base identification decisions only on their memory for the culprit) or a relative judgment strategy (base identification decision on lineup member comparisons). This theory further predicts that relative judgment strategies lead to an increase in false identifications. Contrast this with the diagnostic-feature-detection hypothesis, which predicts that the lineup member comparisons inherent to relative strategies promote greater accuracy. These two theories have been tested indirectly (i.e., via lineup format manipulations tangentially related to the theory), but there is a lack of direct tests. Across two experiments (*N*s = 192, 584), we presented participants with simulated crime videos and corresponding lineups, and manipulated judgment strategy using explicit absolute and relative strategy instructions and a novel rank-order manipulation meant to encourage lineup member comparisons. We found no substantial differences in identifications or overall accuracy as a function of instructed strategy. These results are inconsistent with the theory of eyewitness judgment strategies but provide some support for the diagnostic-feature-detection hypothesis. We discuss implications for both theories and future lineup research.

Table of Contents

List of Tables

List of Figures

Acknowledgments

First and foremost, I would like to acknowledge my supervisor, Dr. Steve Lindsay, for his support throughout my MSc thesis and program. His sage advice and wealth of research experience have been invaluable—no matter my question or struggles, he always had a relevant paper (or five) or an expert contact to point me in the right direction. He has constantly challenged me to carefully examine my research questions and statistical tools, and to uphold the principles of open science, and this thesis is markedly better as a result. I would also like to thank my committee members Drs. John Sakaluk and Michelle Lawrence. It has been an absolute treat having such insightful committee members from different domains than Steve & I, and their perspectives and feedback have helped me understand the value of interdisciplinary collaboration, and to hopefully link my research to the bigger picture. I would like to thank my graduate student cohort, who have over the course of my degree provided useful for my research and both intellectual and emotional support when times have been tough. Finally, I wish to thank the brilliant Cognition & Brain Sciences group at UVic, whose feedback has been instrumental in designing and refining the study reported therein (as well as many planned future studies).

Introduction

Police lineups & eyewitnesses

For centuries, eyewitness testimony has been a cornerstone of the criminal justice system. The police lineup has served as an important part of the eyewitness testimony procedure. In police lineups, an eyewitness is shown a group of individuals (format varying from in-person to photographs and videos) and asked by a lineup administrator whether or not the perpetrator of the witnessed crime is among the individuals (and if the perpetrator is among the members, which member it is). Though there is evidence that specific aspects of the lineup procedure vary considerable across jurisdictions (e.g., number of members, position of the suspect; Wogalter, Malpass, & Mcquiston, 2010), lineups around the world generally share some core features.

Generally, lineups include one or more *suspects* identified by the police along with several known-to-be-innocent *fillers* or *foils.* If the suspect in a lineup stands out or is highly dissimilar to the other lineup members, the lineup is said to be *unfair*—if this is not the case, the lineup is considered *fair*. Most commonly, lineups are presented either *simultaneously* (all members shown at once) or *sequentially* (lineup members shown one at a time). After an eyewitness views a lineup and make a decision, they are often asked how confident they are in that decision. Of central interest to researchers are eyewitness lineup decisions. An eyewitness choosing from a lineup can make a *correct identification* of a guilty suspect, a *false identification* of an innocent suspect, a known-to-be-erroneous *filler identification*, an *incorrect rejection* (i.e., failure to identify) of a lineup containing a guilty suspect, or a *correct rejection* of a lineup containing an innocent suspect. Of these, false identifications and incorrect rejections are especially important from a practical standpoint—the former because it may lead to the wrongful

conviction of an innocent person, and the latter because it may lead to a guilty criminal walking free.

Lineups and eyewitness errors

Unfortunately, eyewitness errors when judging police lineups are prevalent both in laboratory studies and real-world cases (Brewer, Weber, & Guerin, 2019). In the United States, it is estimated that mistaken identifications play an important role in more than 70% of wrongful convictions that have since been overturned by DNA evidence (Innocence Project, 2019). Given the prevalence of real-world errors, the consequences of these errors, and the evidentiary weight given to eyewitness identifications (e.g., Loftus, 1975), finding ways to reduce eyewitness lineup errors is of interest to policymakers, scientists, and the general public. As a result, it is no surprise that factors affecting eyewitness accuracy have been well-studied. These factors are generally divided into *estimator* variables outside of the control of the justice system (e.g., viewing distance, stress, presence of a weapon, eyewitness individual differences, etc.) and *system* variables that are to some degree within the control of the justice system (e.g., lineup format, interviewing techniques, etc.; Granhag, Ask, & Giolla, 2014; Wells, 1978). Though research on estimator variables provides valuable insights into what witnessing conditions might be indicative of eyewitness reliability, estimator variables are uncontrollable. System variables on the other hand are amenable to manipulation—the key implication being that the appropriate modification of identification procedures can improve eyewitness accuracy.

Lineup-specific factors that influence eyewitness decisions have also been well-researched. Such research has revealed several system variables that can affect eyewitness accuracy, including: lineup construction, instructions given to eyewitnesses, lineup administration, and lineup presentation (Gronlund & Carlson, 2013). Prior research on these and

other system variables have identified several factors that contribute to diminished eyewitness performance. For instance, prior exposure to lineup members before seeing the lineup (e.g., in a mugbook or as an innocent bystander), unfair lineups in which most lineup members do not closely match a suspect description (Colloff, Wade, & Strange, 2016; Gronlund & Carlson, 2013), use of biased instructions, verbal or nonverbal administrator influence and post-identification feedback can all result in more error-prone eyewitnesses (Gronlund & Carlson, 2013; Sauer, Palmer, & Brewer, 2018).

There has been a long and spirited research program examining the influences of lineup *format* on eyewitness accuracy. Such research has revealed that: lineups including multiple members are superior to single-suspect lineups ("showups"), lineups should include only one suspect, lineup fillers should be selected on a combination of similarity to the suspect (match-to-suspect) and similarity to descriptions of the perpetrator (match-to-description) (Gronlund & Carlson, 2013).

Much of lineup format research has compared simultaneous and sequential lineups. A seminal study from Lindsay and Wells (1985) found that sequential lineups resulted in a significantly lower probability of a false identification than simultaneous lineups (a difference of 26%), with no significant differences in correct identifications (8% lower for sequential). This striking finding prompted a flurry of research comparing the two formats, culminating in two meta-analyses of this "sequential superiority effect". The first (Steblay, Dysart, Fulero, & Lindsay, 2001) included 30 tests of the effect, and resulted in an estimated 23% lower probability of false identifications with sequential relative to simultaneous, but also a 15% lower probability of correct identifications. The second meta-analysis (from the same authors; Steblay, Dysart, & Wells, 2011) involved 72 tests of the effect, with a sequential false identification

reduction of 22% and a corresponding correct identification reduction of only 8%. Thus, it appears that the sequential presentation of lineup members results in a robust increase in eyewitness accuracy.

Theories of eyewitness judgment strategies: Absolute and Relative judgments

In order to explain *why* sequential lineups appear to increase eyewitness accuracy, Lindsay and Wells (1985; see also, Wells, 1984) proposed a theory of eyewitness judgment strategies. Specifically, they argued that eyewitnesses can adopt either an "absolute" or "relative" judgment strategy when viewing a lineup. The eyewitness who adopts an absolute strategy will compare each lineup member *only* to their memory of the culprit. In contrast, the eyewitness who adopts a relative strategy will compare each lineup member not only to their memory, but also to the other lineup members. These between-lineup comparisons are argued to be extraneous—the only information required to decide whether or not a lineup member is the culprit is one's memory of the culprit and each specific lineup member (Lindsay & Wells, 1985). Furthermore, the lineup comparisons inherent to a relative judgment strategy may lead eyewitnesses to choose the lineup member who *best* resembled the culprit, even if that lineup member is not a good match to their memory for the culprit.

To further outline these strategies and the proposed disadvantages of a relative judgment strategy, consider the following real-world example. In 2004, Uriah Courtney was brought in as a suspect in the sexual assault of a young woman (Albright, 2017). When presented with the following lineup (Figure 1), the woman identified Courtney (#4).

Figure 1. Lineup from *California v. Uriah Frank Courtney*

This identification led to Courtney's conviction. After serving eight years, he would later be exonerated by DNA evidence. The theory of eyewitness judgment strategies offers a potential explanation for the eyewitness' error in this case. There are two important things to note about this lineup. First, of the lineup members above, only Courtney and two other lineup members had goatees, despite the fact that the eyewitness described the perpetrator as having a goatee. Second, and related, Courtney resembled the culprit more than the other lineup members. Now, consider the same lineup below (Figure 2), with hypothetical "similarity-to-culprit" ratings superimposed.

Figure 2. Lineup from *California v. Uriah Frank Courtney:* Hypothetical similarity-to-culprit ratings

Proponents of the judgment strategies theory argue that if an eyewitness used an absolute judgment strategy, they would evaluate only Courtney's similarity-to-culprit (in this example, 70). Let's also assume that the eyewitness will only identify someone if their similarity-to-culprit is at least 80. So, in this case, the eyewitness would not falsely identify Courtney. Now, consider an eyewitness using a relative judgment strategy. This eyewitness will take into account the absolute similarity (70), but because they compare among lineup members, they will also be

influenced by the degree to which the best-matching member is *more* similar to the culprit than the others. For example, Courtney's similarity is 10 higher than the next-best member. Thus, tye relative strategy eyewitness' total evidence (70+10) meets their decision criteria, resulting in a false identification. If Courtney were in fact guilty, it is likely that the strength of the memory signal that he elicited in the eyewitness would be higher, resulting in a correct identification regardless of strategy. Although these hypothetical values are arbitrary, they illustrate the processes thought to underlie absolute and relative judgment strategies (Clark, 2003).

The link between these judgment strategies and the lineup format debate is straightforward: Sequential lineups discourage lineup member comparisons and relative strategy use by virtue of their individual presentation of lineup members, while simultaneous lineups promote unavoidable lineup member comparisons and relative strategy use (Lindsay & Wells, 1985). The judgment strategy theory, if correct, provides a ready explanation for the sequential superiority effect. Many researchers have reported evidence that they interpret as support for the superiority of absolute judgment strategies. Wells (1993) compared lineups with and without the culprit and found that participants viewing a lineup without the culprit often chose the "next best" lineup member. Based on these findings, he assumed that siphoning of identifications off the culprit and to the "next best" match was due to a relative judgment process. Similarly, Dunning and Stern (1994) found that inaccurate eyewitnesses were more likely to report that their judgments featured more relative characteristics (e.g., "I compared the photos to each other to narrow the choices").

Attempts to formally model absolute and relative judgment processes have resulted in mixed success. Steven Clark's WITNESS model (2003; crudely described in the Uriah Courtney example) fit data from three lineup experiments reasonably well and in line with judgment

strategy predictions. However, others have modelled absolute and relative decision processes and found no compelling differences in accuracy (Kaesler, Dunn, & Semmler, 2017; Fife, Perry, & Gronlund, 2014).

Finally, one noteworthy limitation of the eyewitness judgment strategy literature is the lack of direct tests of the judgment strategies. All of the studies mentioned above take sequential and simultaneous lineups to be direct proxies for absolute and relative judgment strategies. To date, there is only one published study that has attempted to manipulate judgment strategy through the use of strategy-informed instructions (Moreland & Clark, 2019). In this study, eyewitnesses were given strategy instructions based on Clark's (2003) WITNESS model—that is, those in the absolute strategy condition were instructed to identify the best-matching lineup member only if they were a "sufficiently good" match to their memory of the perpetrator, and those in the relative strategy condition were instructed to identify the best-matching lineup member only if they were a "sufficiently better match than anyone else in the lineup" to their memory of the perpetrator (Moreland & Clark, 2019). Despite the use of simple and targeted instructions, these researchers did not find compelling differences between the strategies. This is unlikely to be due to a lack of instruction efficacy—there is promising evidence that other lineup and/or witness instruction manipulations may increase eyewitness accuracy. For instance, for certain lineup types meta-memorial instructions about the phenomenology of accurate memories can improve discriminability and eyewitness accuracy (Guerin, Weber, & Horry, 2018).

Signal Detection Theory: An alternative perspective

The inconsistent judgment strategy results in computational modeling studies and the lack of strategy-based differences in the single published direct test point to potential problems with the theory of eyewitness judgment strategies. Some researchers have proposed the adoption

of *signal detection theory* (SDT), an influential paradigm originating in psychophysics that has been used to describe decision-making under uncertainty (Green & Swets, 1966; Link, 1994). Essentially, SDT seeks to describe the processes by which humans decide whether a signal is present under noisy conditions—e.g., "Is that a predator in that tall grass?", "Is that the person I'm supposed to meet in that crowd, or just someone that looks like them?", "Is this black spot on an X-ray a tumour?". SDT has been commonly applied to memory tasks where participants must classify stimuli as *old* ("I saw that before") or *new* ("I didn't study that before"). Thus, SDT can be naturally extended to lineups, where one must determine the presence of a previously-seen target (culprit) in the presence of noise (never-before-seen similar-looking innocent suspects and multiple known-to-be-innocent foils).

SDT posits that performance in tasks like the ones described above depend on two independent processes: *response criterion* and *discrimination*. Response criterion refers to the threshold of evidence or memory strength (e.g., "culprit-likeness") beyond which a stimulus is classified as signal. Response criterion can be conservative, where a greater degree of evidence is required for classification, or liberal, where a lesser degree of evidence is required for classification. Discrimination refers to the ability to distinguish between signal (e.g., the witnessed culprit) and noise (e.g., someone who merely resembles them). Discrimination depends on the degree of overlap between the evidence strength distributions of signal and noise—for example, if our friend is in a crowd of many people who look quite like them, discrimination will be low. In a sense, discrimination can be thought of as the diagnostic accuracy of an individual or procedure. Increasing or decreasing the response criterion of a procedure simply decreases or increases identification rates, whereas increasing or decreasing discrimination results in substantive shifts in the accuracy of that procedure. Thus, proponents of

SDT in lineup research are generally interested in comparing the discrimination of different lineup procedures.

Crucially, SDT assumes the existence of hypothetical memory strength distributions for signal/targets and lures/noise, and that response criterion and discrimination are independent. Figure 3 below illustrates these SDT concepts.

Figure 3. Illustration of SDT concepts: Response criterion and discrimination (Wixted & Mickes, 2014).

Each of the three panels depicts memory strength distributions for targets (dotted lines) and lures (solid lines). Memory strength will tend to be higher for targets (because they have been seen before), but there will be some overlap between the distributions. The vertical lines in the panels represent varying levels of response criterion from liberal (low evidence required to identify) to conservative (high evidence required to identify). Note that while response bias varies, discrimination (the distance between the center of the distributions) remains constant.

What does all of this have to do with the theory of eyewitness judgment strategies? The answer lies in *how* lineup outcomes have been measured in previous studies. Much of the early research (e.g., Lindsay & Wells, 1985) quantified lineup performance in terms of the *diagnosticity ratio*: the ratio between correct and false IDs, with a larger ratio denoting a better lineup procedure. Rates of correct and false IDs and the diagnosticity ratio are simple, intuitive

measures of lineup outcomes. However, they may not fully capture the accuracy or utility of lineup procedures. To see why, we again consider the hypothetical memory strength distributions in Figure 3. At any point along the *x*-axis, one can compute a diagnosticity ratio for that hypothetical response criterion by dividing the height of the target distribution (i.e., the # of correct IDs at that criterion) by the height of the lure distribution (i.e., the # of false IDs at that criterion). Once we see this, we can also see that by simply increasing response criterion *while holding discrimination constant*, the diagnosticity ratio is increased. The upshot? The previous findings of a sequential advantage may be explained by a conservative shift in response bias—in other words, sequential lineups may simply decrease choosing with no substantive effects on accuracy (Wixted & Mickes, 2014). Thus, proponents of SDT argue that the theory of eyewitness judgment strategies is a theory of response bias, remaining silent on the much more important issue of discrimination (2014).

Luckily, SDT provides a metric for measuring discriminability in the form of *receiver operating characteristic* (ROC) curves. ROC curves were originally used in WWII to measure the ability of radar operators ("receivers") to distinguish enemy aircraft from radar noise, but were readily adapted into SDT (Green & Swets, 1966) and more recently, lineup research (Wixted & Mickes, 2014). In the lineup context, ROC curves plot the diagnosticity ratio across all possible levels of response bias (see Figure 4 below for an example).

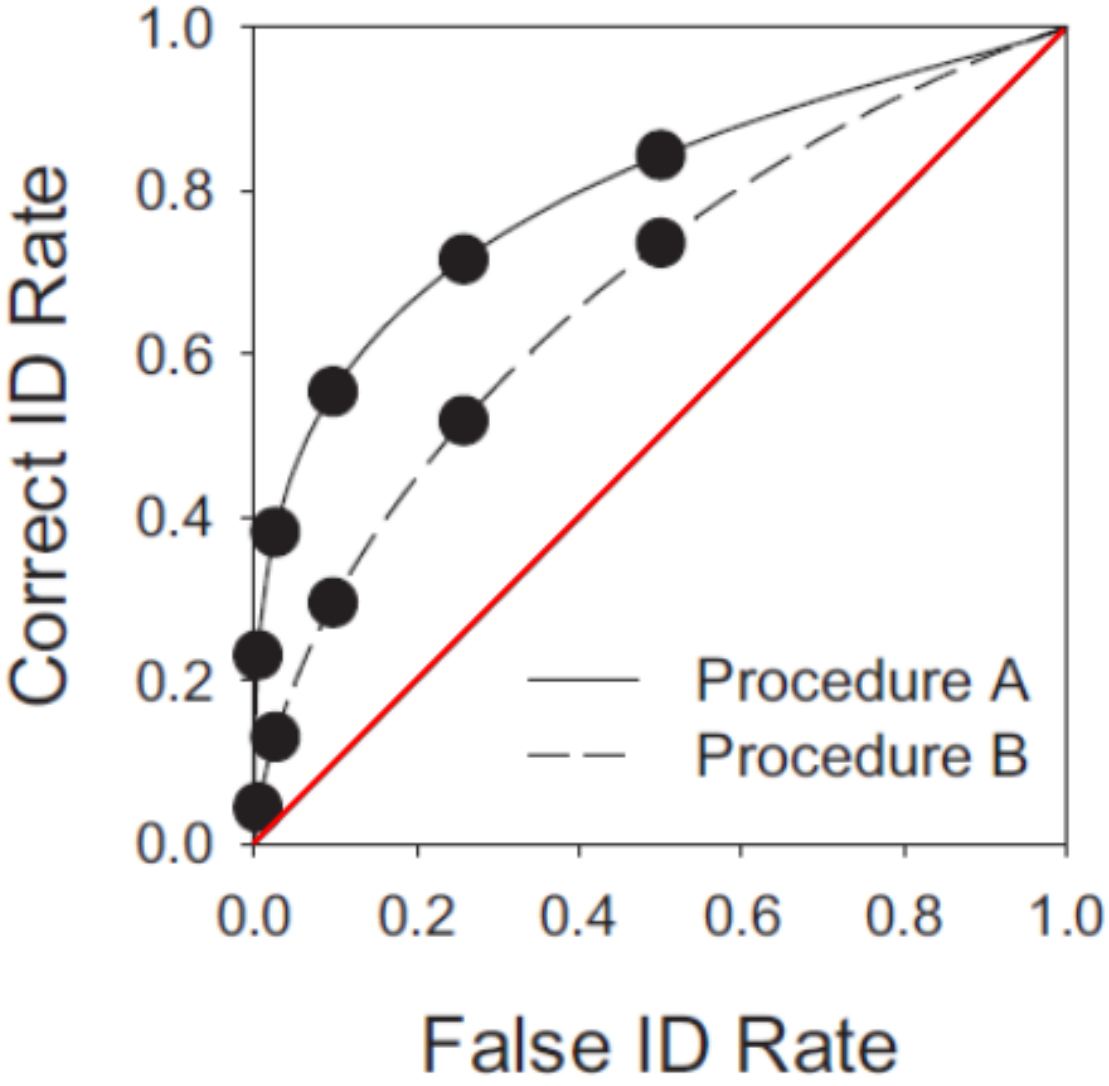


Figure 4. Example ROC curves for two lineup procedures (Wixted & Mickes, 2014).

Each point on the ROC curves represents a different level of response criterion, with the rightmost point representing the most liberal criterion and the leftmost point representing the most conservative criterion. To provide a concrete example, ROC curves in the medical literature are often used to measure the diagnostic accuracy of a given test. Here, the rightmost point might

represent the diagnosticity ratio when using only one test indicator (e.g., fever), while the leftmost point might represent the diagnosticity ratio when using all test indicators (e.g., fever, coughing, shortness of breath). Lineup measure do not include indicators, but response bias can be indexed by using confidence as a proxy. Because most lineups elicit both a decision and confidence in that decision, one can approximate response bias. For example, the rightmost point on the ROC curve (the most liberal criterion) would be calculated using identifications made at *any* level of confidence (e.g., 0 – 100%), the next point might be calculated using only identifications made with at least 10% confidence, the next with identifications made with at least 20% confidence, and so on. An ROC curve constructed in this way provides information about accuracy across all hypothetical levels of response criterion, thus unconfounding criterion and discrimination and providing a more complete picture of lineup performance. The discrimination of a given procedure can be measured by calculating the *area under the curve* (AUC), which can then be used to compare procedures. The larger the AUC, and the further away from the diagonal line (chance performance), the more discriminating the procedure.

ROC curves have the further advantage of incorporating information about eyewitness confidence. Not all identifications are created equal—it is now widely accepted that all else being equal, low-confidence decisions tend to be error-prone, while under "pristine" conditions, high-confidence decisions (particularly IDs) tend to be accurate (Wixted & Wells, 2017). Additionally, jurisdictions may vary in the stock they put in various levels of eyewitness confidence (Wixted & Mickes, 2015). ROC curves allow one to compare the relative utility of lineup procedures under a variety of different criteria. Recently, ROC curves have been criticized by some that argue that AUC-based analyses can lead to erroneous conclusions about lineup procedure diagnosticity—particularly when the lineup procedures under scrutiny lead to different

innocent-suspect ID rates (Smith, Lampinen, Wells, Smalarz, & Mackovichova, 2019). The novel measure proposed by these researchers, dubbed *deviation from perfect performance* (DPP), shows some promise as a new gold standard. Overall, SDT-informed ROC-based measures offer lineup researchers a new and promising way to analyze identification data.

Re-examining the Absolute vs. Relative literature

With the advent of ROC-based lineup measures, researchers raised an important point about the previously observed sequential superiority effect. Namely, the previously reported diagnosticity ratio-based findings were compatible with very different explanations (see Figure 5 below).

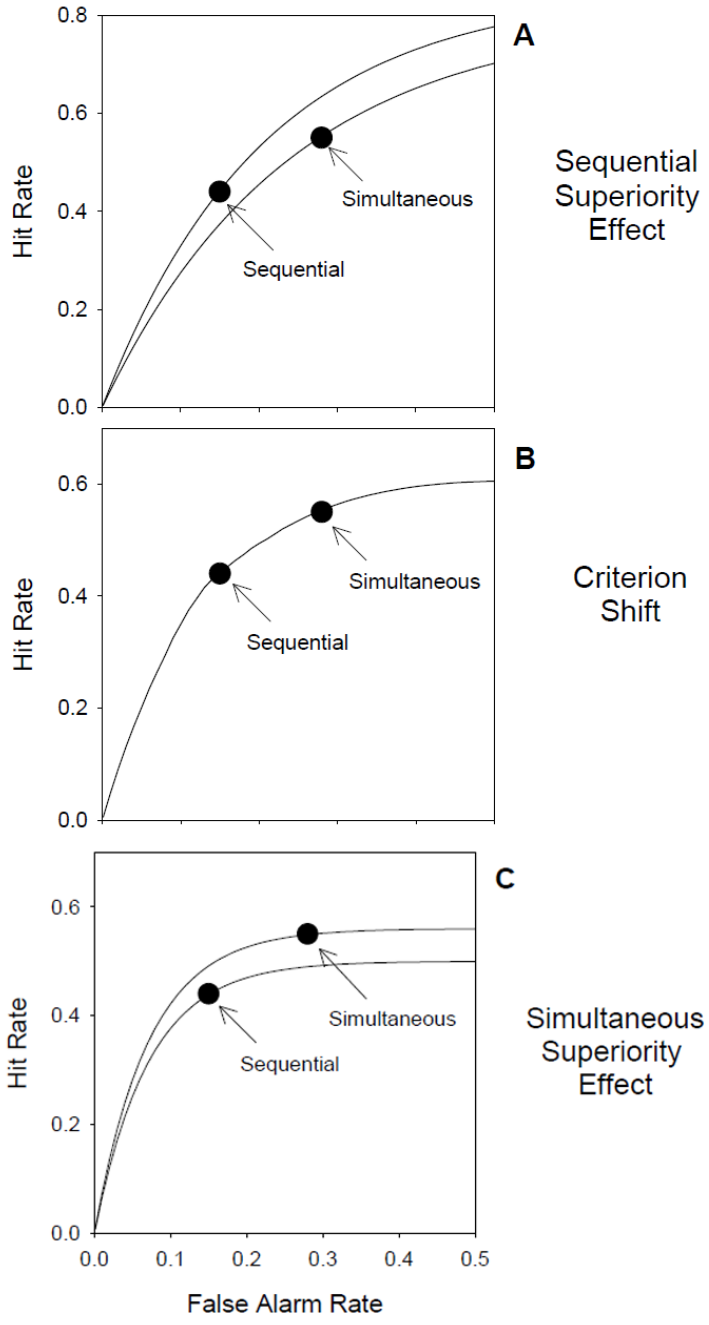


Figure 5. Potential explanations for previous lineup data (Mickes, Flowe, & Wixted, 2012)

The observed sequential superiority effect could be due to an increase in discrimination with the sequential procedure (Panel A). As suggested, the results may be explained by a simple criterion shift (Panel B). Finally, it is also possible that the previous results are explained by a

simultaneous superiority effect (Mickes et al., 2012). Without constructing ROC curves, none of these possibilities can be eliminated. Thankfully, researchers have done just that, with surprising findings. Increasingly, studies show evidence for a simultaneous superiority effect—that is, simultaneous lineups afford more discrimination than sequential lineups (Dobolyi & Dodson, 2013; Mickes et al., 2012; Wixted & Mickes, 2014).

In order to explain this surprising finding, researchers have married the theory of eyewitness judgment strategies and SDT, with the resulting union being the *diagnostic-feature-detection (DFD)* hypothesis of eyewitness identification (Wixted & Mickes, 2014). The DFD hypothesis asserts that the comparisons inherent to relative judgment strategies (and simultaneous lineups) allow eyewitnesses to better identify distinctive (i.e., unique) facial features that are diagnostic of guilt and to identify and *discount* non-distinctive facial features that are *not* diagnostic of guilt.

Consider an illustrative example (Wixted & Mickes, 2014)—an eyewitness witnesses a culprit who is a White male in his early 20s with an oval face and small eyes. The description they provide to the police is of a "White male in his early 20s", and the police assemble a six-person lineup based on this description. When judging the lineup and deciding whether the culprit is present, the witness may differentially weight various features—the age, race, shape of face, shape of eyes, etc. Because the lineup members are presented simultaneously, the witness will be more likely to notice that some features (e.g., age, race) are non-diagnostic of guilt because everyone in the lineup shares them, and there can be only one guilty culprit. Thus, they may be more likely to discount these features, and focus on unique, diagnostic features (e.g., shape of face, shape of eyes). By doing so, their ability to discriminate between guilty and innocent suspects will be increased. Conversely, if the lineup members are presented

sequentially, it will be harder for the witness to extract and discount shared, non-diagnostic features.

DFD-based predictions have been substantiated in subsequent studies. One novel and promising lineup manipulation involves *rank-order lineups* (Carlson et al., 2019a). With this manipulation, participants rank-order lineup members in order of perceived similarity to the culprit (presumably also encouraging lineup comparisons; Carlson et al., 2019a). When compared to standard simultaneous lineup, rank-order participants showed increased discrimination. Drawing on the DFD hypothesis, Carlson and colleagues argue that rank-order procedures facilitate comparisons among lineup members, which reduces reliance on shared features that are non-diagnostic of guilt. Findings of rank-order benefits dovetail nicely with research on *elimination lineups*. In elimination lineups, eyewitnesses select the lineup member who is most similar to the culprit then decide whether the best match is in fact the culprit (Guerin et al., 2018; Pozzulo & Lindsay, 1999). Prior research suggests that the elimination procedure reduces false (but not correct) identifications (Pozzulo & Lindsay, 1999). Further recent support for the DFD hypothesis comes from Colloff and Wixted (2019), who found that discriminability in a showup was enhanced by presenting the showup suspect alongside five "foils" that could not be selected (i.e., the decision was only whether or not the showup suspect was the culprit). By doing so, they showed that the increase in discriminability afforded by simultaneous presentation is not simply due to "filler-siphoning" (the siphoning of some incorrect identifications from the innocent suspect to fillers). Finally, in non-eyewitness domains, there have been a few studies showing discriminability advantages when similar objects (including faces) are presented simultaneously rather than alone or sequentially (Wixted & Mickes, 2014).

All in all, recent emerging findings increasingly suggest that relative judgment strategies informed by the DFD account may improve the accuracy of eyewitness lineup decisions. However, these findings contrast with current recommendations from the Public Prosecution Service of Canada (2018) to police agencies, most notably: “A photo-pack should be provided sequentially, and not as a package, thus preventing ‘relative judgments’.” More studies providing more conclusive evidence for (or against) the benefits of relative judgments could serve to improve police practice.

In the specific case of rank-order procedures, we are aware of only one published study. Other researchers have examined similar lineup manipulations, with varying results. Similar work where witnesses make confidence judgments about all lineup members has been done, with evidence for accuracy benefits (Brewer, Weber, Wootton, & Lindsay, 2012). Framed in terms of the DFD hypothesis, it may be that making individual lineup member confidence judgments facilitates comparisons among lineup members (e.g., confidence in one lineup member relative to another). Other researchers have tested “grain-size lineups”, where participants eliminate lineup members to construct a final set to choose from (Horry, Brewer, & Weber, 2016). Though witnesses tended to calibrate grain/set size to memory (e.g., coarser grain for crimes with poorer viewing conditions), grain-size lineups did not outperform standard simultaneous lineups. In the context of the DFD hypothesis, it is possible that finer-grained lineups (i.e., smaller sets) reduced the amount of possible comparisons, and subsequently the amount of non-diagnostic feature discounting. However, fine-grained lineup decisions were generally more probative than coarse-grained ones (Horry et al., 2016), which one might not expect if the DFD hypothesis holds. Given the paucity of direct tests of rank-order lineups and mixed findings tangential to the DFD

hypothesis, more research is warranted. Such research might also help to rule out potential alternate explanations.

Though some of these studies and theories are suggestive, the overall evidence is inconclusive, and there has been only one published experimental test of the hypothesis that relative judgment strategies inherently lead to more false identifications (Moreland & Clark, 2019). Instead, most studies dance around the central point—the strategies themselves—via examinations of related but distal manipulations (e.g., rank-order lineups, elimination lineups, grain size lineups). In addition to providing stronger, more direct tests of the hypotheses surrounding absolute and relative judgment strategies, further experiments could provide insight into the degree to which eyewitness judgment strategies can be manipulated.

The Current Research

At this point, it's worth taking stock of the literature on judgment strategies, SDT, and the various relevant theories proposed to explain lineup decision making:

1. Early research proposed a sequential (absolute) advantage, based on the theory of eyewitness judgment strategies
2. New research informed by signal detection theory suggests a simultaneous (relative) advantage, which is explained using the diagnostic-feature-detection hypothesis.
3. There is a lack of direct tests of the judgment strategies, with most studies using manipulations not directly related to the judgment strategies. Furthermore, many of these manipulations (e.g., rank-order lineups; Carlson et al., 2019a) show promise, but have not been replicated.

In sum, eyewitness lineup researchers are poised between two influential, conflicting theories of lineup decision making. Increasingly, evidence seems to weight in favor of SDT/DFD-based theories. However, without more direct tests of the strategies, it is impossible to conclusively falsify one of these theories.

More broadly, examinations of whether and how strategy instruction manipulations affect eyewitness lineup performance represent a potentially valuable and untapped area of system variable research—one that speaks directly to the ongoing theoretical discussions about eyewitness judgment strategies. If the use of one strategy (absolute, relative, or some other judgment strategy) promotes greater eyewitness accuracy, specially tailored strategy instructions could be implemented in current lineup practices with few costs. If strategy instruction manipulations are unable to shift judgment strategies, it is still possible that other strategy-related changes to lineup procedures could improve eyewitness decision making.

To address these points, we conducted the current study with the following aims:

1. Directly compare the merits of absolute and relative judgment strategies by explicitly instructing participant-witnesses to use one strategy or another when completing lineups.
2. Provide a strong test of the hypothesis that relative judgments lead to an increase in false identifications. We did this by creating a "worst-case scenario" for relative judgments—lineups in which an innocent suspect resembled the culprit far more than the fillers (a "Uriah Courtney" scenario).
3. Examine the robustness of a promising relative-judgment-themed lineup manipulation: rank-order lineups (Carlson et al., 2019a). For our relative judgments experimental condition, we paired strategy instructions with the rank-order procedure.

Hypotheses

We conducted two similar main experiments and three pilot experiment with these aims. Both main experiments were pre-registered via the Open Science Framework (osf.io). The pre-registration for the first experiment can be viewed here: https://osf.io/t89vg, and the second here: https://osf.io/jtqf2. Though we deviated from our pre-registration for several analyses (deviations subsequently detailed either in-text or in footnotes), our core hypotheses remained the same. Because the hypotheses, design, and results were similar in both experiments, we combined their hypotheses and analyses. We report separate analyses by experiment in the Supplementary Material (Section E for Experiment 1, Section F for Experiment 2), and note in-text and in footnotes where there were substantive differences between experiments. The combined hypotheses were as follows:

H1: We hypothesize either no difference or fewer false identifications (IDs)[1] when participants are instructed to use a relative judgment strategy (with rank-ordering) than when participants are instructed to use an absolute judgment strategy.

H2: We hypothesize either no difference or an increase in guilty-innocent discriminability (as measured using ROC curve AUC and DPP) when participants are instructed to use a relative judgment strategy (with rank-ordering) than when participants are instructed to use an absolute judgment strategy.

With these hypotheses, we tentatively throw in with the SDT/DFD camp. This hypothesis combines our Experiment 1 prediction of no strategy differences (based on the only other published direct test, Moreland & Clark, 2019, as well as our own prior, unpublished work across three experiments) with our Experiment 2 predictions of a relative strategy reduction in

[1] For Experiment 2, our pre-registered hypotheses were framed in terms of *accuracy* instead of *identifications*. To provide a more intuitive interpretation and a better match with the literature, we reframed the hypotheses in terms of identifications (the substantive predictions regarding strategy differences did not change).

false IDs (informed by the recent rank-order study and DFD work). Experiment 1 focused on the hypothesis that relative judgments increase false IDs, and only included culprit-absent lineups. Experiment 2 focused more on the hypothesis that relative judgments increase discriminability, and included culprit-absent and culprit-present lineups (i.e., to allow construction of ROC curves).

H3: We make no a priori predictions about the effects of strategy on rates of correct IDs or filler IDs.

If a relative judgment strategy + rank-order procedure leads to better discrimination of diagnostic features (as per the DFD hypothesis and new rank-order studies), then subjects instructed to use a relative judgment strategy with rank-ordering may, relative to subjects instructed to use an absolute judgment strategy, have a similar or higher probability of correctly identifying the culprit, and a similar or lower probability of incorrectly identifying a filler. Alternatively, if relative judgments + rank-ordering merely increases response criterion (i.e., promotes conservative responding), then those in the relative condition may have a lower probability of making a correct ID.

Given the discussion of lineup accuracy measures—particularly the advantages of SDT measures over simple correct and false ID rates and diagnosticity ratios, one might argue that H1 and H3 are redundant with H2. One reason to conduct the simpler analyses is that correct and false ID rates are more intuitive to laypeople. However, conducting these analyses also permits comparisons with previous research that *only* used these measures (e.g., the original absolute vs. relative work of Lindsay and Wells, 1985). By conducting both sets of analyses, we also hope to provide converging evidence for or against our hypotheses.

H4: We predict that both the culprit and our designated "innocent suspect" will be most often chosen as the person who looks most similar to the culprit.

This hypothesis serves as a manipulation check to ensure that the lineups we created did in fact approximate the "worst case" scenario for relative judgments.

Finally, we pre-registered two supplementary hypotheses related to potential relative vs. absolute differences in confidence and difficulty, and potential lineup order effects. These hypotheses, along with analysis results, can be found in the Supplementary Material (Section D).

Analytic strategy

Bayesian estimation

To evaluate these hypotheses, we adopted a Bayesian estimation approach (McElreath, 2019). Because this approach differs from traditional NHST and Bayes Factors methods, we devote the following section to an explanation of Bayesian estimation. As the name suggests, the focus of Bayesian estimation is on the estimation of models and parameter values. This approach is similar to the "New Statistics" extolled by Geoff Cumming (among others; 2012), which eschews the binary hypothesis testing of traditional NHST in favor of accurate estimation of effect sizes. Estimation approaches avoid many of the pitfalls of NHST and binary hypothesis testing (e.g., dichotomous thinking, bias against the null hypothesis, confusion about *p*-values and associated statistics; Cumming, 2012).

Bayesian estimation is, at its core, "no more than counting the number of ways that things can happen, according to our assumptions. Things that can happen more ways are more plausible" (McElreath, 2019, p. 11). The "assumptions" of Bayesian estimation are represented by *prior distributions* on effects, and can be thought of as our pre-data knowledge and expectations about a particular effect or difference. In Bayesian estimation, prior distributions are

mathematically combined with the data (*likelihood*) to yield *posterior distributions*, which represent the state of our knowledge after having seen the data. To illustrate these concepts further, consider the following simple example. Imagine that we have a coin, and do not know whether it is fair or not, and we want to estimate the probability of flipping heads. Imagine also that we have no prior expectations about the coin's fairness (i.e., we view all probabilities as equally plausible). These expectations can be represented by the following "flat" **prior** distribution in Figure 6 below:

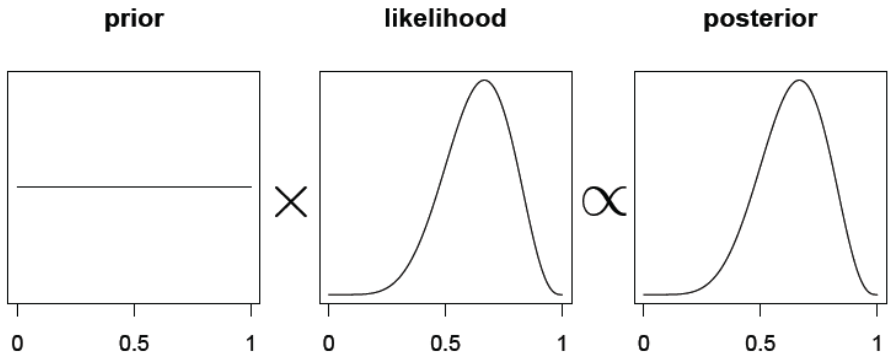


Figure 6. Example of prior, likelihood, and posterior distributions: Flat prior (McElreath, 2019)

Now, imagine that we flip the coin 100 times and obtain 70 heads. Based on this data, we construct the **likelihood** distribution, which represents the distribution of likely "probability of heads" values based purely on our data. By combining the prior and posterior distributions, we arrive at the **posterior**, which shows the distribution of likely probabilities based on both our prior and current knowledge. This posterior can then be carried forward as a prior for new analyses (e.g., if we were to flip the coin another 100 times). Notice that in this example, the posterior is nearly identical to the likelihood. When one has broad or nonspecific priors, the data tend to shape the posterior (especially with larger sample sizes). With a more specific prior, such

as one that assumes that a fair coin is much more likely than an unfair one (Figure 7), we can see how priors can influence the posterior.

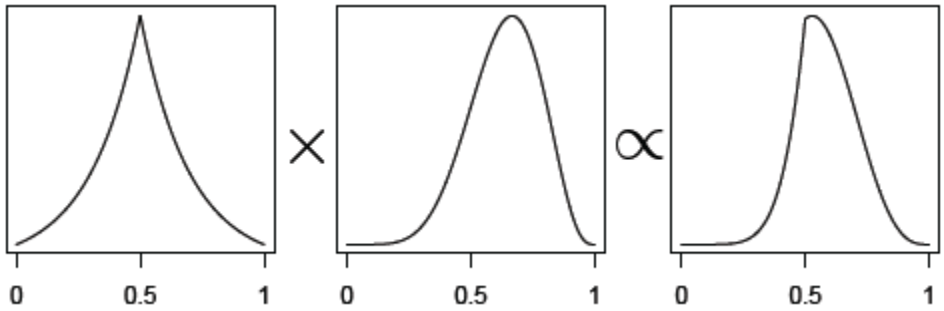


Figure 7. Example of prior, likelihood, and posterior distributions: Specific prior (McElreath, 2019)

With this technique, the focus of inference is on the posterior distribution. For example, one might calculate the most probable value of a parameter, or the range of values that we are X% certain that the true parameter value falls within (referred to as an X% *credible interval*—hereafter denoted by CR). The coin example extends to more complex models (like the multi-level models we will analyze in this study), except that instead of estimating the most likely *single* parameter values, complex models estimate the most likely *combination* of parameter values. In these cases, it is not plausible to arrive at the posterior distribution analytically (like in the simple example). For complex models, the posterior can be approximated by sampling repeatedly from the multidimensional parameter space using an innovative technique known as Hamiltonian Markov Chain Monte Carlo[2]. Details aside, the main takeaway here is that Bayesian

[2] See https://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/ for an excellent discussion and examples of Bayesian sampling techniques, including interactive demonstrations. For all of our estimation analyses, we ran 4 chains of 2,000 iterations (1,000 warmup) each.

estimation has the advantages of estimation approaches and also allows one to incorporate explicitly specified prior knowledge.

Admittedly, there is a disconnect between the binary hypotheses we propose and the analytic methods we adopted. For our estimation analyses, we did not have any hard cut-off criteria that will indicate evidence for/against strategy effects. Rather, we interpreted posterior distributions of effect sizes and based any conclusions on the distributions and the precision of the results obtained (i.e., whether most of the distribution favors an effect size within a certain range, whether the distribution is too wide to draw any unambiguous conclusions). This approach allowed us to provide a more detailed picture of our results while still requiring that any conclusions we draw be defensible to the skeptical outsider (we hope!).

We note that the adoption of Bayesian estimation as our main analytic technique was pre-registered for Experiment 2, but not Experiment 1. Originally, we planned standard Bayesian ANOVAs and *t* tests (Rouder, Morey, Speckman, & Province, 2012; Rouder, Speckman, Sun, Morey, & Iversion, 2009), but after preliminary data analyses (and subsequent learning), it became clear that generalized mixed effects linear models were the best way to analyze our data. These models take into account the dichotomous nature of our outcome variable (Yes/No IDs), allowed us to account for dependencies in our data (multiple responses per participant), as well as item-level variability (differences among our lineups). To complement our pure estimation analyses, we supplemented all analyses with Bayes Factors, which we discuss next.

Bayes Factors

For our analyses, we computed Savage-Dickey Bayes Factors (BFs), which simply represent the ratio of evidence in favor of a given effect under two models, typically the prior and posterior (Wagenmakers, Lodewycx, Kuriyal, & Grasman, 2010, see also

https://vuorre.netlify.com/post/2017/03/21/bayes-factors-with-brms/). To make this more concrete, consider again the coin toss example and our prior and posterior distributions (Figure 8):

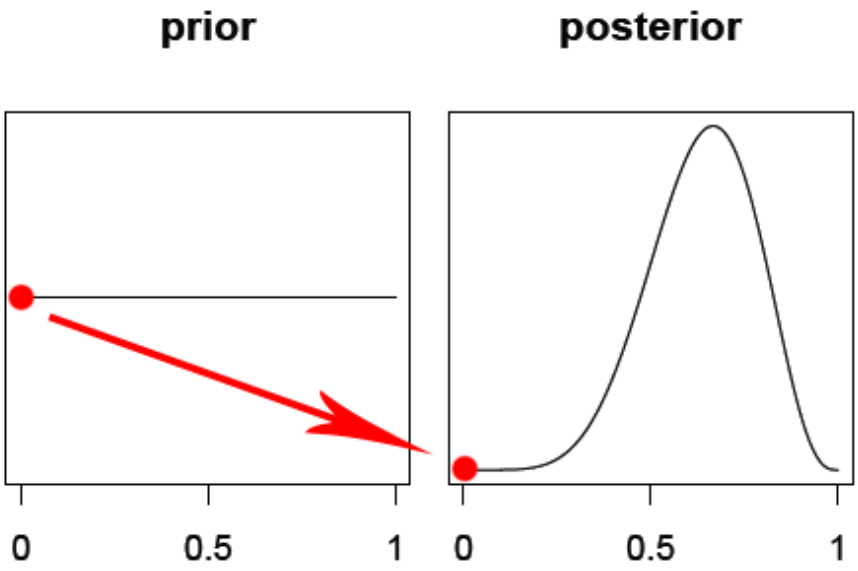


Figure 8. Savage-Dickey BF computation (modified from McElreath, 2019)

Say we want to evaluate the evidence in favor or against a 0% probability of flipping heads. To do this, we simply divide the height of the probability distribution at 0% under the posterior by the height under the prior. In this example, the resulting BF might be something like (.05/.5) = .01. The interpretation of this number is straightforward: *after seeing the data, a 0% probability of flipping a heads is 10 times less likely*.

BFs like these often provide a more intuitive interpretation of results than posterior distributions and represent the degree to which we should update our beliefs in a given effect size. However, they are *highly* dependent on the prior distribution (contrast with posterior distributions, which are generally dominated by the data). With broader priors, Savage-Dickey BFs become less readily interpretable. Because we primarily use broad, minimally constraining priors for our analyses, we advise caution in interpreting our reported BFs. Finally, although BFs

are fully continuous evidence ratios, some researchers adopt categorical cut-offs similar to *p*-values (e.g., a BF > 3 constitutes a meaningful effect; Jeffreys, 1961, in Wagenmakers, Wetzels, Borsboom, and van der Maas, 2011). Though we initially planned to adopt this approach in our pre-registration for Experiment 1, we did not use categorical cut-offs in subsequent analyses (to remain consistent with our overall estimation philosophy). Instead, the goal of computing BFs was to provide additional information about estimated effects beyond the posterior distribution.

WAIC model evaluation

The final aspect of Bayesian estimation that we adopted was model comparison. Model comparison is valuable because it allows one to evaluate the relative contribution of effects to a model's predictive ability (e.g., comparing a model without a main effect to a model with that main effect). In order to compare models, one needs a criterion on which to compare them. We adopted the *widely applicable information criteria* (WAIC). Without going into too much detail, WAIC approximates cross-validation and provides an estimate of the out-of-sample predictive ability (i.e., how well a model would predict hypothetical new data), while adjusting for the amount of parameters in a model (see Gelman, Hwang, & Vehtari, 2014 and McElreath, 2019 for more details). WAIC and other information criteria are preferable to other metrics (e.g., R^2) because information criteria balance against model overfitting via the parsimony consideration. Among the various information criteria available (e.g., AIC, DIC, BIC, WAIC), there are several reasons to prefer WAIC. It does not make assumptions about the shape of the posterior distribution (as AIC and DIC do), allows estimation of prediction at the pointwise level (unlike BIC), and provides estimates that are quite similar to more sophisticated cross-validation techniques (e.g., LOO, LOOIC), while being less computationally intensive (McElreath, 2019).

We used WAIC differences (and intervals on those differences) to compare a variety of models for each hypothesis and evaluate the degree to which an effect (e.g., an absolute vs. relative strategy difference) was important in predicting our dependent variables. As with BFs, we did not have any hard-and-fast criteria for designating an effect as important or not. Rather, WAIC comparisons served as an additional piece of the results puzzle (with posterior distributions and BFs), providing a more complete picture. We note that our use of WAIC for model comparisons was not a part of the pre-registered plan for Experiment 1, and was modified from the plan for Experiment 2. We initially planned to use WAIC to *combine* the predictions of multiple models (each model weighted according to its WAIC), but after further research and preliminary analyses it became clear that this "model stacking" approach (Yao, Vehtari, Simpson, & Gelman, 2018) can be problematic when models do not contain the same parameters. Instead, we opted for the simpler model comparison strategy.

Method

Procedure

Our pilot tests and experiments were programmed in Qualtrics (Qualtrics.com) survey software and had the following general structure:

1. Participants viewed a number of videos of simulated crimes (four in Experiment 1 and two in Experiment 2). The videos were graciously provided by Melissa Colloff, and were used in a previous lineup study (Colloff, Wade, & Strange, 2016). The four videos alternately depicted a carjacking in a parking garage, a theft of a laptop from an office, a mugging in a parking garage, and a graffiti-ing in an empty classroom. Each video lasted approximately 30s, with the exposure to the culprit varying from 5s to 16s. The culprits in each crime video were all white males. All four videos were used in Experiment 1,

while only the carjacking and theft were used in Experiment 2. Video order was counterbalanced across participants.

2. After each video, participants answered basic questions about the event that happened in the video (this was used as an attention check and to later exclude participants).
3. Participants completed a mental rotation distractor task in which they were presented with rotated images of numbers or letters and had to determine whether the rotated number or letter was normal or a mirror-image of its normal presentation
4. Participants were told that they would view photospread lineups for each of the simulated crimes they have viewed (in the same order as the crime videos). They were given detailed descriptions of the lineup procedure, told they should imagine that they were real eyewitnesses, and instructed to be as accurate as possible. They were also told that the culprits of the crimes may or may not be present in each of the lineups.
5. Participants were randomly assigned to use either an absolute or relative judgment strategy when judging the lineups. Participants were given the following strategy instructions:
    a. Absolute: “For each lineup, compare each face one by one to your memory of the culprit. This is called an "absolute judgment" strategy – **it is a matter of comparing each face in the lineup only to your memory of the culprit in the video.** Do not compare the faces within a lineup to one another.”
    b. Relative: “For each lineup, compare the faces to one another and think about which of the faces look more versus less like the culprit. This is called a "relative judgment" strategy – **it is a matter of comparing the faces in the lineup to one another to determine which one of them looks the most like the culprit in the**

video. For each lineup, you will need to order the lineup members in order of most to least similar to the culprit. You will do this by entering a number beside each person, where 1 is the person you think is most similar, 2 is the next most similar and so on, where 6 is the person you think is least similar to the culprit."

6. Participants then viewed either four culprit-absent lineups (Experiment 1) or one culprit-absent and one culprit present lineup (order counterbalanced; Experiment 2) corresponding to the crime videos. Each lineup consisted of six black-and-white photos (also from the Colloff et al., 2016 materials), one of which was the innocent suspect (culprit-absent lineups) or the culprit (culprit-present) lineups and five of which were designated as fillers. The photos were presented in two rows of three photos each (a typical arrangement for simultaneous lineups). The position of the suspect/culprit was randomized. An example lineup (Relative strategy condition) is shown below in Figure 9.

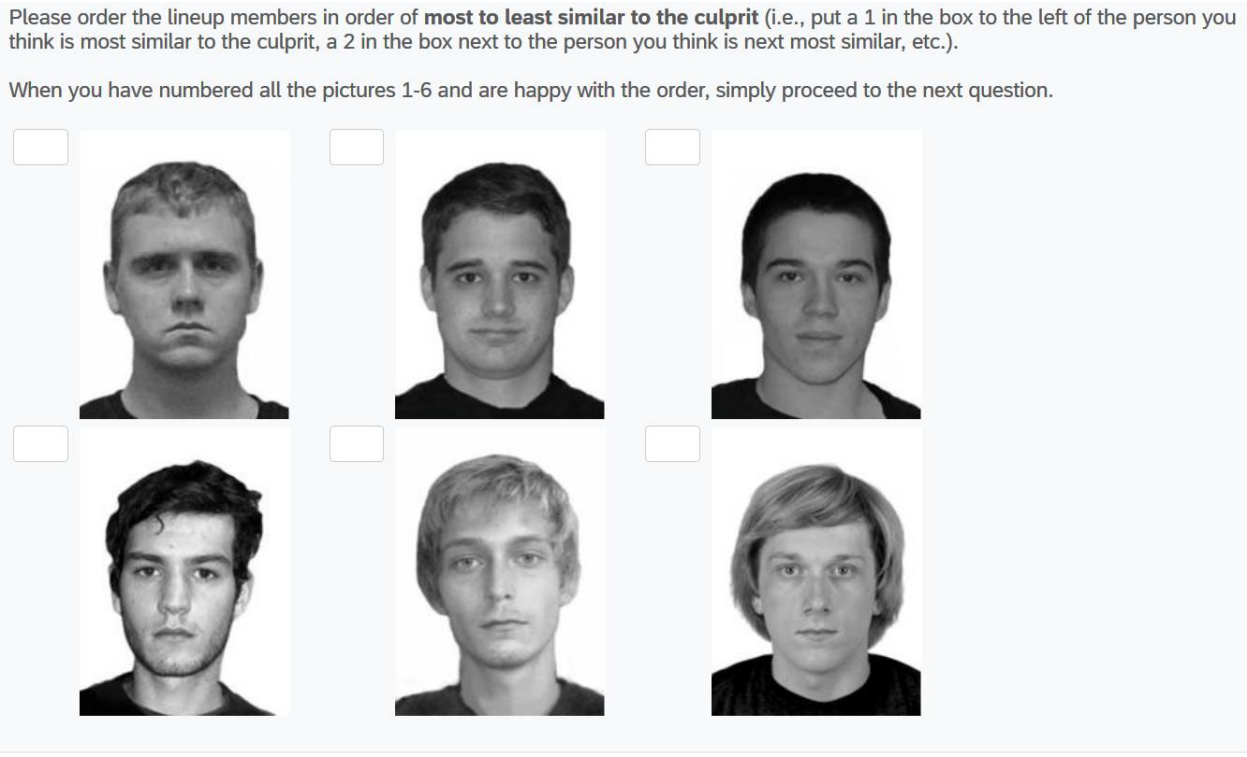


Figure 9. Example lineup (relative rank-order condition)

   a. For each lineup, participants either identified a lineup member as the culprit or responded that the culprit was not present in the lineup. Before making a decision, participants in the relative strategy condition rank-ordered the lineup members in order of most-to-least similar to the culprit by entering a number from 1 (Most similar) to 6 (Least similar) in boxes beside each member's photo. After ranking the lineup members, relative strategy participants indicated whether or not the person they chose as most similar was in fact the culprit.
   b. After making a decision, participants rated their confidence in the decision as well as the decision difficulty. Both ratings were made on 11-point Likert scales from 0-100, with 0 representing "Not at all confident/difficult" and 100 representing "Very confident/difficult".

7. After each lineup, participants were reminded of their assigned strategy and asked to self-report their adherence to that strategy, on a 4-point Likert scale from "Not at all" to "Completely". This question would later be used to exclude lineups on which participants reported not adhering to their assigned strategy at all.
8. For participants in the absolute strategy condition, we presented the same lineups that they had viewed previously and simply asked them to choose the person they thought most resembled the culprit (whether or not that person actually was the culprit). The purpose of these "Forced Relative" lineups was to test our lineup manipulation (i.e., to ensure that our suspects/culprits resembled the culprit more than the fillers). Participants in the relative strategy condition did not complete this phase, as they had already chosen a "most similar" lineup member in the main phase of the experiment.

9. Participants were asked a final set of questions pertaining to attentiveness, technical issues, vision impairments, and whether or not they had completed a similar study before (several of these would be used to exclude participants).
10. Finally, participants were debriefed.

Lineups used in our experiments can be found in the Supplementary Material (Section A). Copies of the full experiment programs are available in Qualtrics and PDF format at https://osf.io/8ac2g/. Our experiments were conducted both online via Prolific.ac, a crowdsourcing website similar to mTurk, and in-person in computer labs with groups of undergraduate students ranging in number from ~1-20.

Lineup pilot testing[3]

As mentioned, our key lineup manipulation involved making the suspect or culprit much more similar to the culprit than the fillers. To accomplish this, we conducted Pilot 1 via Prolific.ac. For this pilot, we presented $N = 54$ online participants with the four crime videos and four culprit-absent lineups. These lineups were designed such that the innocent suspect resembled the culprit much more than the other lineup members (based only on the subjective criteria of the authors). For each lineup, participants simply chose the person they thought most resembled the culprit. Based again on largely subjective interpretation, we replaced several non-suspect lineup members who were viewed as more or similarly similar to the culprit than our designated innocent suspect. We then used these modified lineups for Experiment 1.

Though our suspect similarity manipulation worked well in Experiment 1, we wanted to create lineups using a more principled and true-to-life method. To accomplish this, we conducted Pilot 2 after Experiment 1. We presented $N = 42$ Prolific participants with all four videos and

[3] Pilot tests were not pre-registered

asked them to provide up to 15 different characteristics that they thought best described the culprit. From these model descriptions, the author and two independent RAs extracted and agreed upon a set of 3-5 characteristics[4] most commonly mentioned by participants. These characteristics were used to inform the construction of new lineups for the carjacking and theft that we would use in Experiment 2. Specifically, we chose designated innocent suspects that matched all of the most commonly mentioned characteristics, and selected fillers who matched on only 2 of the characteristics. Again, we note that interpretation of descriptors and facial features was somewhat subjective here. The data for this pilot test (including all suspect descriptors given and various stages of the coding and characteristic selection process) are available at https://osf.io/8ac2g/.

In a final, larger-scale pilot test (Pilot 3, after Experiment 1 and Pilot 2), we reduced the total number of lineups by two and conducted additional pilot testing with the aim of examining general response patterns on new culprit-absent and culprit-present versions of our two lineups (carjacking and theft, the lineups for which our suspect similarity manipulation had worked the best in Experiment 1). This pilot test was conducted in a sample of $N = 240$ undergraduates.

Sample

Pre-registered exclusion criteria. For Pilot 3 and Experiments 1 & 2, we pre-registered a number of exclusion criteria at the participant and lineup level. We excluded participants who self-reported inattentiveness, vision problems, or technical issues, or who did not correctly complete the rank-ordering task on any lineup. We excluded lineups where participants failed to correctly answer at least one video-attention-check question, reported not adhering to their assigned strategy, or for those in the absolute condition, any lineup where a participant identified

[4] Excluding “white” and “male”, as well as any non-facial characteristics (e.g., clothing)

a member as the culprit but later selected a different member as being most similar to the culprit. The full list of exclusionary criteria, along with the justification for each, is presented in the Supplementary Material (Section B)

Pilot tests. Pilot 1 included $N = 54$ online participants ages 18-30 ($M = 23.9$, $SD = 3.3$), with 57% of participants identifying as male (1 participant identified as something other than male or female). Pilot 2 included $N = 42$ online participants ages 18-35 ($M = 24.3$, $SD = 5.7$), with 62% of participants identifying as male. Beyond excluding participants who did not complete Pilots 1 and 2, we did not exclude any other participants. Participants in Pilots 1 and 2 received $1.54 USD for participating.

For Pilot 3, we collected data from 264 undergraduate participants. From this sample, we excluded 3 participants who chose to withdraw after completing the study, 3 who reported vision problems, 1 who reported completing a similar study, 1 who reported video size problems, 3 who reported other technical problems, and 13 who self-reported as being "Very inattentive". This left us with a final sample of $N = 240$ undergraduate participants ages 16-46 ($M = 21.1$, $SD = 4.4$), with 27.5% of participants identifying as male (3 participants identified as something other than male or female)[5]. After participant exclusions, this left us with 480 total lineups. From these, we dropped 5 where participants did not get at least one video question right, 13 where participants did not report at least partly adhering to their assigned strategy, and 5 absolute condition lineups where there was a mismatch between main lineup and Forced Relative lineup choices. This left us with 457 lineups for analysis. Undergraduate participants received bonus credit towards an eligible course for participating.

[5] For Pilot 3 and Experiment 2, we used a transgender-inclusive measure of sex/gender (Bauer, Braimoh, Scheim, & Dharma, 2017).

Sample sizes for the pilot tests were not based on any principled criteria and were selected arbitrarily.

Experiment 1. For Experiment 1, we collected data from 238 undergraduate participants. From this sample, we excluded 21 participants who chose to withdraw after completing the study, 13 who reported vision problems, 2 who reported completing a similar study, 6 who reported technical problems, and 4 who self-reported as being "Very inattentive". This left us with a final sample of N = 192 undergraduate participants ages 18-47 (M = 20.6, SD = 4.3), with 24.5% of participants identifying as male. After participant exclusions, this left us with 768 total lineups. From these, we dropped 19 where participants did not get at least one video question right, 13 where participants did not report at least partly adhering to their assigned strategy, and 17 absolute condition lineups where there was a mismatch between main lineup and Forced Relative lineup choices. This left us with 719 lineups for analysis.

The original pre-registered sampling plan for Experiment 1 was to collect an initial sample of N = 200, begin analyzing the data via default Bayesian ANOVA and t test, and conduct sequential Bayesian sequential hypothesis testing with optional stopping if default BFs for our main hypothesis tests exceeded 4 in favor or against H1 (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017). Sequential hypothesis testing is an efficient alternate to a priori static sample sizes (2017). However, when we analyzed the data at N = 192 (after exclusions), we decided that extending the research via Experiment 2 was more efficient (more on this with the Experiment 1 results).

Experiment 2. For Experiment 2, we collected data from 642 online and 44 undergraduate participants. From this sample, we excluded 22 participants who chose to withdraw after completing the study, 7 who reported vision problems, 4 who reported

completing a similar study, 6 who reported video size problems, 41 who had duplicate rankings on at least one relative lineup, 16 who reported other technical problems, and 6 who self-reported as being “Very inattentive”. This left us with a final sample of $N = 584$ participants (21 undergraduates) ages 17-75 ($M = 30.4$, $SD = 10.4$), with 56% of participants identifying as male (4 participants identified as something other than male or female). After participant exclusions, this left us with 1,168 total lineups. From these, we dropped 20 where participants did not get at least one video question right, 23 where participants did not report at least partly adhering to their assigned strategy, and 30 absolute condition lineups where there was a mismatch between main lineup and Forced Relative lineup choices. This left us with 1,095 lineups for analysis. Combining Experiments 1 and 2 resulted in a final analysis sample of 1,814 lineups from 776 unique participants. Online participants in Experiment 2 received $2.00 USD for participating, and undergraduate participants received bonus credit.

Sample size planning for Experiment 2 was driven by practical constraints. Our pre-registered target $N = 600$ was based on a study budget upper limit. Because our primary analytic focus was on estimation, power as typically conceived was a non-issue for us. Rather, we hoped to collect enough data to obtain reasonably precise estimates of the effects we were interested in. Simulations to determine the likely estimation precision we would obtain with this sample size showed that we could expect to obtain 95% CRs on strategy effects with widths between 14% (i.e., +/- 7%) and 20% (i.e., +/- 10%).

However, we did pre-register several fail-safe stopping criteria based on data quality (specific criteria reported in Supplementary Material Section B). Starting at $n = 100$ and after each additional 100 participants, we checked our data against these criteria. Though there were slight violations of a few of the criteria at several points in data collection, we judged that the

violations were not severe enough to preclude continued testing (see Supplementary Material Section B for more details).

Results

Data files and analysis scripts (in R format) are available at https://osf.io/8ac2g/.

Pilot 3

The primary aim of this final pilot test was threefold:

1) To ensure that our designated innocent suspects and culprits were chosen as most similar to the culprit at least ~70% of the time
2) To ensure that no filler was chosen more than ~10% of the time
3) To replicate basic/expected response patterns observed in our prior research—namely, that culprit choosing rates will be higher than innocent suspect choosing rates (i.e., at least ~10%)

Although we did not pre-register the analyses for Pilot 3, here we adopt relevant pre-registered methods from Experiment 2. To test 1) and 2), we computed the proportion of times each lineup member was chosen as most similar to the culprit (separately for the carjacking and theft lineups, and for culprit-absent and culprit-present lineups). For each proportion, we then computed bootstrapped 95% CRs. We then evaluated means and 95% CRs against our criteria[6]. The results are shown below in Figures 10-13.

[6] This differed from our pre-registered plan to conduct a full Bayesian estimation model with all lineup members as separate predictors. Instead of this overly complex analysis, we evaluated bootstrapped proportions, and conducted a *single* Bayesian model estimating the probability that our designated innocent suspect/culprit was chosen as most similar. This single model still allowed us to answer the main question of interest: Are our suspect/culprits chosen as most similar more often than one would expect in a fair lineup?

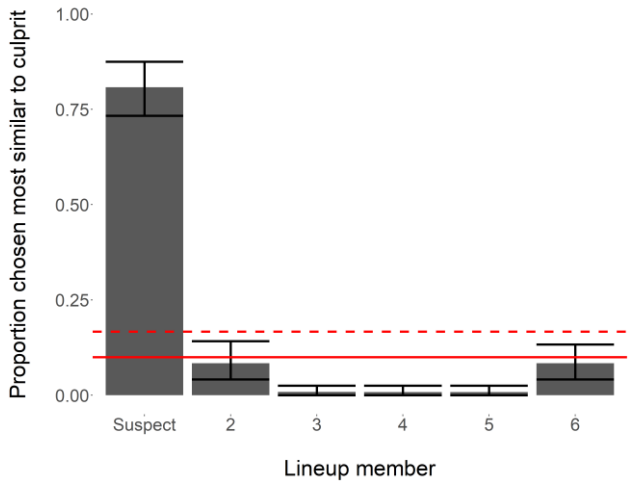


Figure 10. Carjacking culprit-absent lineup: Lineup member choosing. Errors bars = Bootstrapped 95% CRs. Solid red line = 10%, dashed red line = 16.7%.

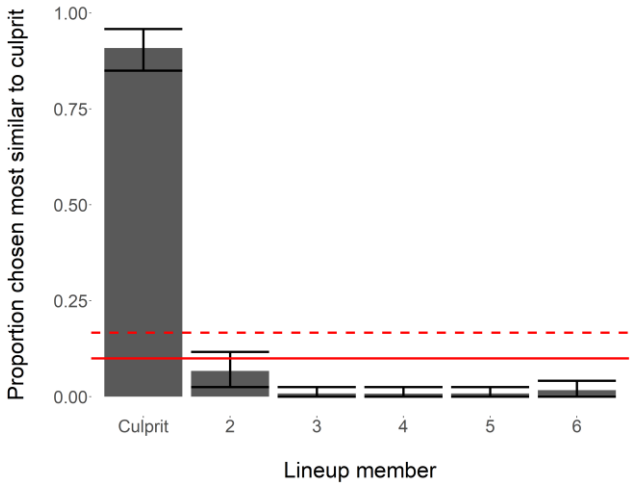


Figure 11. Carjacking culprit-present lineup: Lineup member choosing. Errors bars = Bootstrapped 95% CRs. Solid red line = 10%, dashed red line = 16.7%.

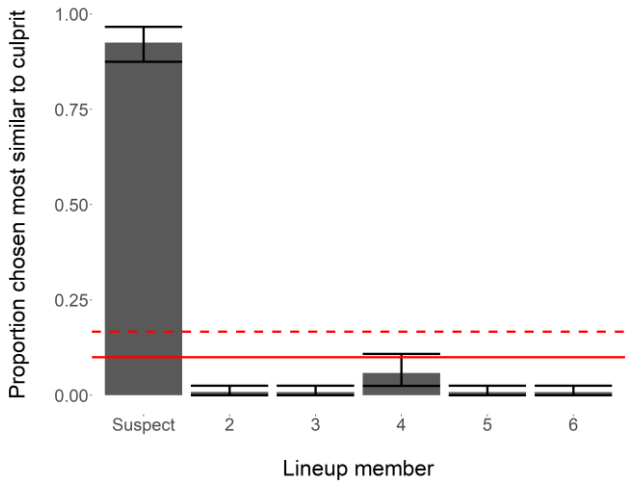


Figure 12. Theft culprit-absent lineup: Lineup member choosing. Errors bars = Bootstrapped 95% CRs. Solid red line = 10%, dashed red line = 16.7%.

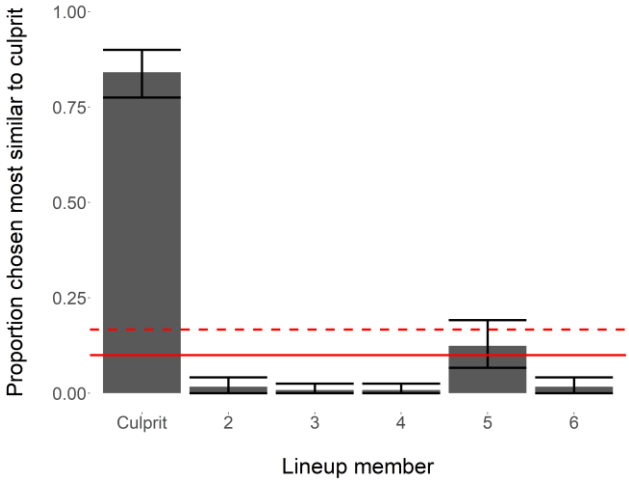


Figure 13. Theft culprit-present lineup: Lineup member choosing. Errors bars = Bootstrapped 95% CRs. Solid red line = 10%, dashed red line = 16.7%.

Descriptively, we can see that all culprits and designated innocent suspects were chosen the vast majority of the time (and >70%), and in only one lineup was another lineup member chosen more than 10% of time (solid red line). Crucially, no lineup fillers were chosen more than

the rate one would expect in a fair lineup where all lineup members are chosen equally (16.7%, dashed red line).

We supplemented each of these descriptive analyses with a single Bayesian estimation model per lineup, estimating the probability that the culprit/innocent suspect was chosen as most similar. The primary purpose of this analysis was to compute a Bayes factor (BF) comparing the posterior probability to a prior *assuming that the culprit/innocent suspect would be chosen near the level of chance in a completely fair lineup (16.7%)*[7]. These BFs represent the relative likelihood that each culprit/innocent suspect was chosen at a chance level. All BFs in favor of a choosing probability > 16.7% exceeded 100; in other words, after seeing the data, the probability that our culprits and innocent suspects were chosen at fair lineup levels is < 1/100. Together, these results show that our lineup manipulation in the pilot was highly successful.

To assess 3) and get a basic idea of culprit and innocent suspect choosing across strategies, we constructed and compared several Bayesian estimation models predicting the probability of an ID (p(ID))[8] from Strategy (absolute, relative) and Presence (culprit-present/CP, culprit-absent/CA). In total, we constructed six Bayesian generalized linear mixed effects models:

1) p(ID) ~ Intercept + (1|Participant) + (1|Lineup)
2) p(ID) ~ Intercept + Strategy + (1|Participant) + (1|Lineup)
3) p(ID) ~ Intercept + Presence + (1|Participant) + (1|Lineup)
4) p(ID) ~ Intercept + Presence + Strategy + (1|Participant) + (1|Lineup)

[7] Specific priors for this and all other Bayesian analysis can be found in the Supplementary Material (Section B).
[8] The form of the dependent variable in this and all subsequent experiments was a dichotomous Yes (identified the suspect or culprit) / No (rejected the lineup). Note that for these analyses, we excluded lineups on which a filler ID was made, because the practical significance of filler IDs is lower than that of correct IDs, false IDs, and lineup rejections. In the main experiments, p(Filler ID) was analyzed separately as per H3.

5) p(ID) ~ Intercept + Presence + Strategy + Presence x Strategy + (1|Participant) + (1|Lineup)
6) p(ID) ~ Intercept + Presence + Strategy + Presence x Strategy + (1|Participant)

These models are *generalized* in that they include a binomial outcome variable (whether or not an ID was made), and *mixed effects* in that they include random intercepts (the 1|X terms) for participants and lineups. In other words, the models allowed for baseline p(ID) to vary by participants and by lineup. The inclusion of these terms also allowed the model to account for within-subjects dependencies, between-participant variability (i.e., participant effects), and between-lineup variability (i.e., item effects)[9]. Some (e.g., McElreath, 2019), argue that mixed-effects/multilevel models should be the default (as opposed to models that average over items and participants).

Finally, these models are Bayesian in that they include prior distributions on all effects. Our chosen priors viewed as most plausible: average performance (~50%), small strategy-based differences in either direction (0% on average), a slightly higher p(ID) on CP lineups (~60%), and a small-to-moderate amount of variability due to participants/lineups. These priors were minimally constraining and incorporated only our expectation that culprits will be chosen more

[9] Our pre-registered models did not initially include lineup intercepts (but should have, to account for item-level effects). We rectify that error here. Additionally, with Model 5, we test the predictive gain achieved by including lineup random intercepts to see whether their inclusion is warranted. We do adopt loose criteria for the inclusion of exclusion of lineup random intercepts, because their inclusion greatly increases variability in the estimates of the outcome variables, obscuring other effects. If the exclusion of lineup random intercepts results in a greater decrease in predictive power (WAIC) than the majority of other model changes, and if the 95% CI on the WAIC difference excludes zero, we will consider lineup random intercepts as important and will include them in the model(s). Our pre-registered models also initially include random *slopes* on presence, by participant (i.e., they allowed for the within-subjects effect of presence to differ across participants). We dropped this (and any other random slope terms) due to difficulties in model estimation when including these complex terms.

often than innocent suspects[10]. For a visual depiction of these priors, see the Supplementary Material (Section C).

The above models were constructed and then compared to one another via WAIC[11]. This method allowed for the evaluation of individual effects (in the same way one might do with a standard ANOVA). The results of the model comparison are shown below in Figure 14:

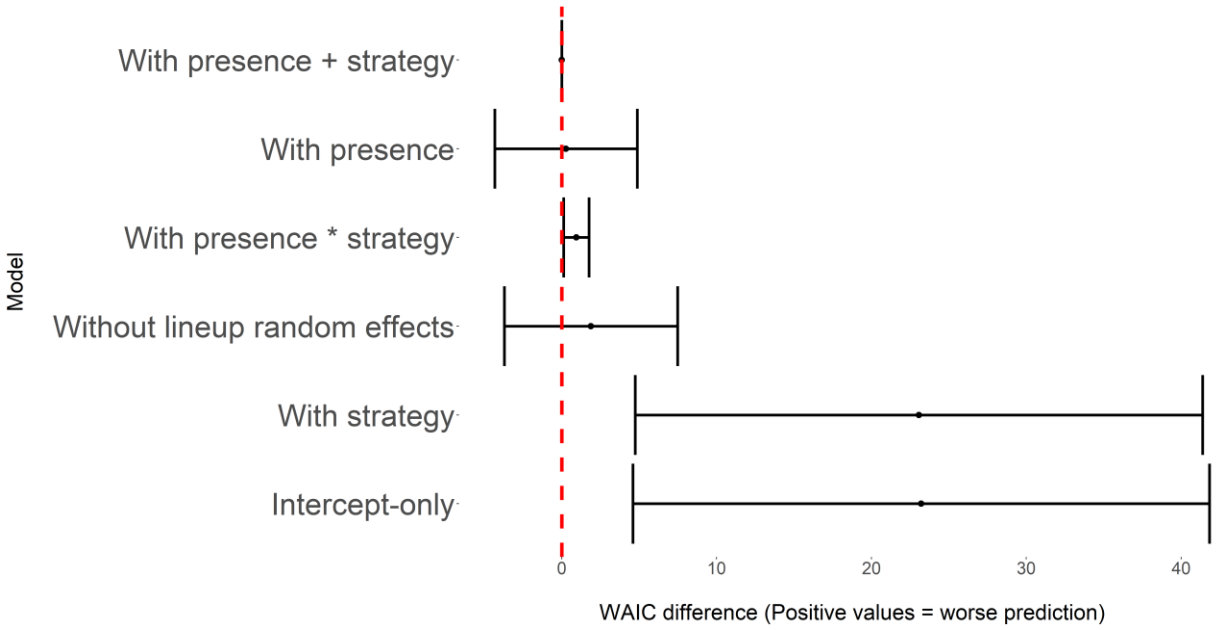


Figure 14. Pilot 3: WAIC model comparison results for predicting p(ID). Models presented in

[10] We deviated slightly from our pre-registered priors for Experiment 2, which allowed for more variability in baseline average p(ID). We reduced the variability in the distribution for this prior to 1) avoid the assumption of more prior variability in one condition than another and 2) make the resulting BFs more interpretable.

[11] This method differs from our pre-registered plan to combine models via model stacking. The original model stacking method was abandoned for reasons mentioned previously. In interpreting the WAIC model comparisons, we adopted a heuristic approach—if the 95% CR on the model difference excludes 0, we consider that model comparison to represent a meaningful difference between models. As a result, these comparisons were somewhat insensitive to the precision with which the WAIC difference was estimated. For example, in Figure 14, the lineup random effects WAIC difference point estimate is larger than the presence * strategy point estimate, but the CR is wider and includes 0. Thus, we are more certain that the presence * strategy interaction reduces model predictive ability slightly, but are less certain about the effects of including or excluding lineup random effects. It is possible that 1) lineup random effects harm model predictions, 2) lineup random effects do not affect model predictions, or 3) lineup random effects benefit model predictions. For two out of the three possibilities, excluding lineup random effects (as we did) will not substantively change model predictive accuracy. Thus, in cases like these we took an Ockham's razor approach to term selection.

descending order of predictive ability relative to the best model (1st Row). Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

This figure shows that the best model in terms of predictive ability is one that includes main effects of Strategy and Presence. However, a model with only Presence was not substantially worse (Row 2) and a model with only Strategy (Row 5) was substantially worse, indicating that Presence effects were doing most of the predictive work. Adding a Strategy * presence interaction (Row 3) resulted in a small but precise decrease in predictive ability, indicating that Presence effects did not vary much across Strategies. Removing lineup-level item-based effects (Row 4) did not substantially reduce prediction, indicating that there may not have been much lineup-based variability. Finally, the baseline intercept-only model had noticeably worse predictive ability, indicating that *something* in the model is predicting p(ID). Based on the above results, it is likely that Presence effects predict p(ID) more than other effects.

To examine p(ID) (i.e., choosing) across Strategy and Presence conditions, we estimated posterior distributions for a model with all effects (excluding lineup random effects as a result of the above comparisons). The results are displayed below in Figure 15.

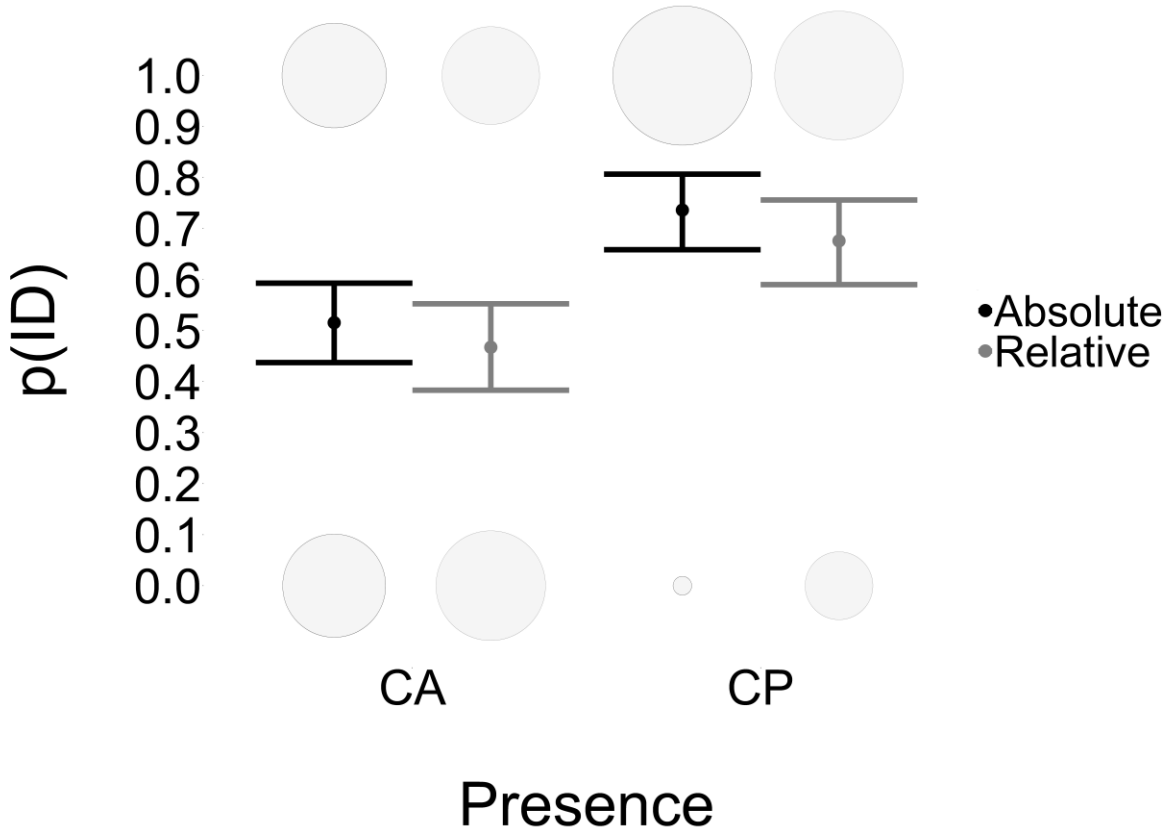


Figure 15. Probability of making an ID by Strategy and Presence. CA = culprit-absent, CP = culprit-present. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models). Shaded circles represent the relative proportion of participants who gave responses at each probability.

These results corroborate the model comparison results—the effect of Presence is noticeably larger than that of Strategy. Recall that our objective with Pilot 3 was to confirm our basic expectation that culprit choosing rates would be higher than innocent suspect choosing rates. To evaluate this, we estimated the difference in culprit/suspect choosing rates separately by

Strategy. In the absolute condition, choosing on CA lineups was 22% lower than on CP lineups [95% CR: 12%, 32%], and in the relative condition, choosing on CA lineups was 21% lower than on CP lineups [95% CR: 10%, 32%]. Savage-Dickey BFs in favor of non-zero presence effects for these comparisons were > 99. Thus, we think it is reasonable to conclude that we observed our basic expected results pattern (choosing at least 10% higher on CP lineups than CA lineups). Though the objective of Pilot 3 was not to examine strategy differences, the results are suggestive of a potential slight reduction in IDs with the relative rank-order procedure. We investigated this possibility further with our combined Experiment 1 and 2 data.

Experiments 1 & 2

H4: We predict that both the culprit and our designated "innocent suspect" will be most often chosen as the person who looks most similar to the culprit. Before moving on to our substantive hypotheses, it was necessary to confirm that our lineup fairness manipulation was successful. Though Pilot 3 suggested that the manipulation would be successful in Experiment 2, we did not conduct extensive pilot testing of the lineups used in Experiment 1. We used the same analytic strategy as in Pilot 3 to estimate the proportion of times that our culprits and innocent suspects were chosen as most similar to the culprit. These proportions, combined across Experiments 1 and 2, are presented in Figures 16-21 below[12].

[12] See the Supplementary Material for experiment-specific figures and corresponding Bayes Factors (Section C for Experiment 1 and Section D for Experiment 2).

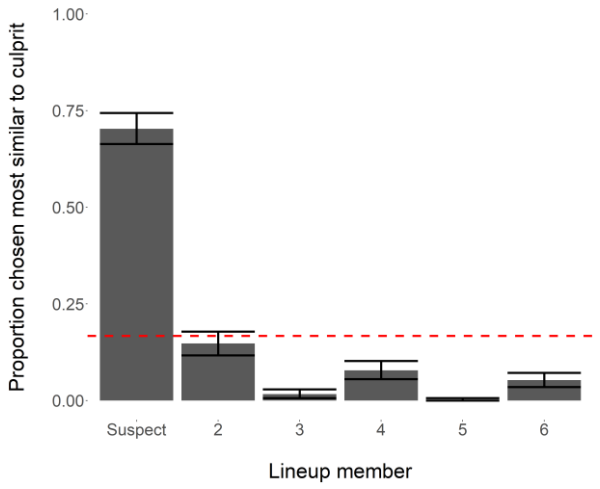


Figure 16. Carjacking culprit-absent lineup for Experiments 1 & 2 combined: Lineup member choosing. Errors bars = Bootstrapped 95% CRs. Dashed red line = 16.7%.

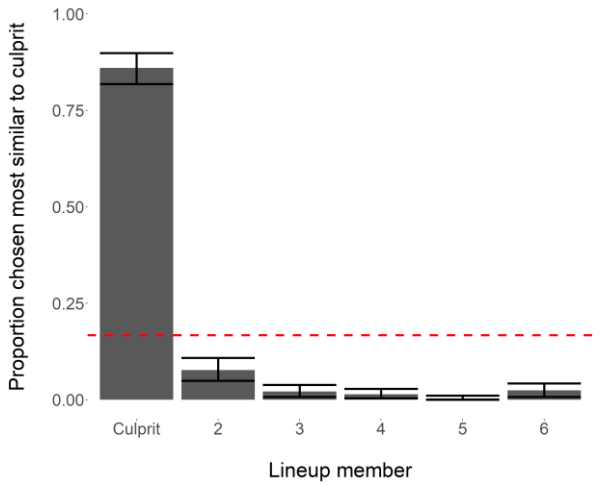


Figure 17. Carjacking culprit-present lineup for Experiments 1 & 2 combined (only includes Experiment 2 data because Experiment 1 did not have culprit-present lineups): Lineup member choosing. Errors bars = Bootstrapped 95% CRs. Dashed red line = 16.7%.

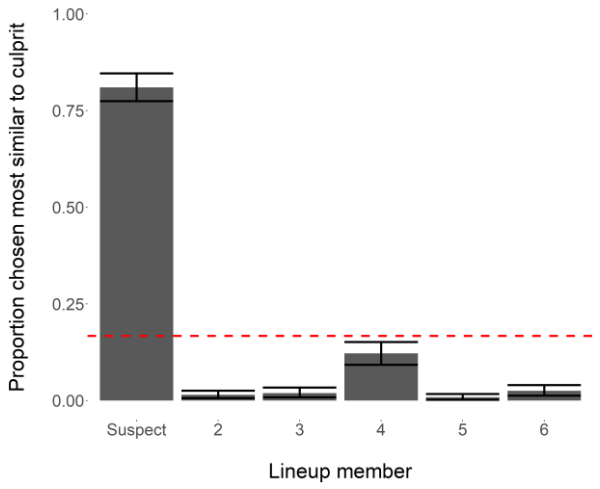


Figure 18. Theft culprit-absent lineup for Experiments 1 & 2 combined: Lineup member choosing. Errors bars = Bootstrapped 95% CRs. Dashed red line = 16.7%.

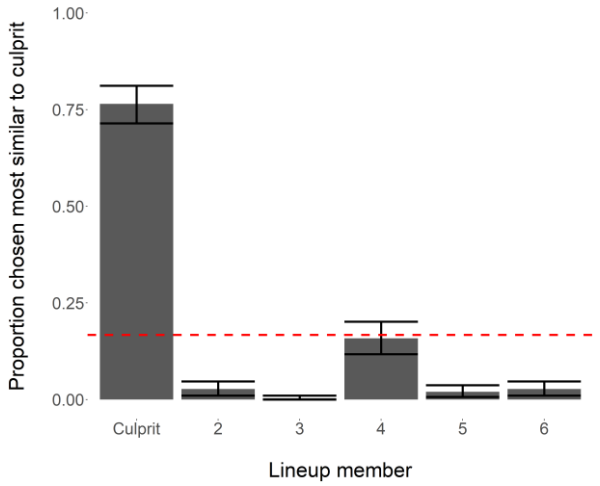


Figure 19. Theft culprit-present lineup for Experiments 1 & 2 combined (only includes Experiment 2 data because Experiment 1 did not have culprit-present lineups): Lineup member choosing. Errors bars = Bootstrapped 95% CRs. Dashed red line = 16.7%.

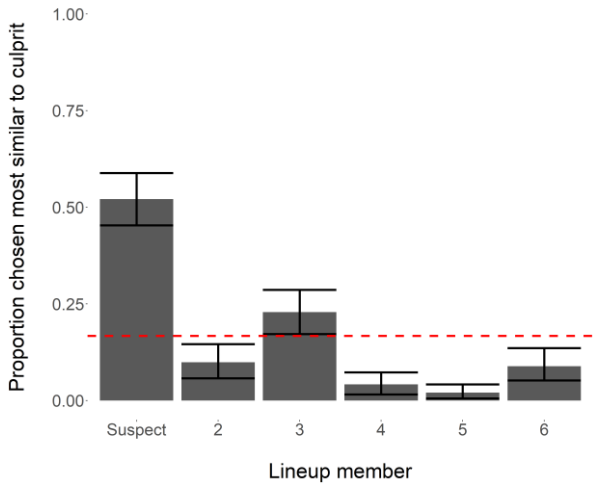


Figure 20. Mugging culprit-absent lineup for Experiments 1 & 2 combined (only includes Experiment 1 data because Experiment 2 did not include the mugging video/lineup): Lineup member choosing. Errors bars = Bootstrapped 95% CRs. Dashed red line = 16.7%.

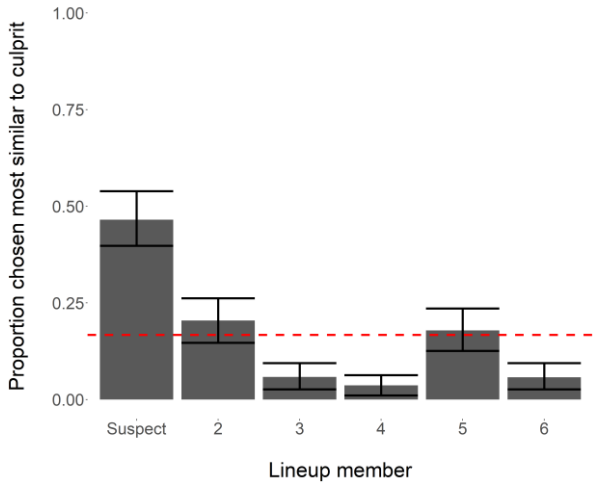


Figure 21. Graffiti-ing culprit-absent lineup for Experiments 1 & 2 combined (only includes Experiment 1 data because Experiment 2 did not include the graffiti-ing video/lineup): Lineup member choosing. Errors bars = Bootstrapped 95% CRs. Dashed red line = 16.7%.

BFs in favor of a choosing proportion higher than 16.7% for the culprits/innocent suspect are displayed below in Table 1.

Table 1. Savage-Dickey BFs in favor of "unfair" culprit/suspect choosing

Lineup	Culprit/Suspect?	BF supporting manipulation
Carjacking	Culprit	> 100
	Innocent suspect	> 100
Theft	Culprit	> 100
	Innocent suspect	> 100
Mugging	Innocent suspect	50
Graffiti-ing	Innocent suspect	25

Based on the descriptive statistics and BFs, most of the lineups appear to be working as intended, with the possible exception of the Mugging and Graffiti-ing CA lineups, where one or more fillers were chosen at or above the chance level. Despite this, in both of these lineups the innocent suspect was still chosen the majority of the time, suggesting that on the whole our lineup manipulation was successful.

H1: We hypothesize either no difference or fewer false identifications (IDs) when participants are instructed to use a relative judgment strategy (with rank-ordering) than when participants are instructed to use an absolute judgment strategy.

H3: We make no a priori predictions about the effects of strategy on rates of correct IDs or filler IDs. We evaluated these hypotheses in our combined Experiments 1 & 2 data using the same model comparison and estimation approach detailed in Pilot 3[13]. We first focused on

[13] Individual experiment results are presented in the Supplementary Material (Section C for Experiment 1 and Section D for Experiment 2).

models predicting correct and false IDs. The results of the model comparison are displayed below in Figure 22.

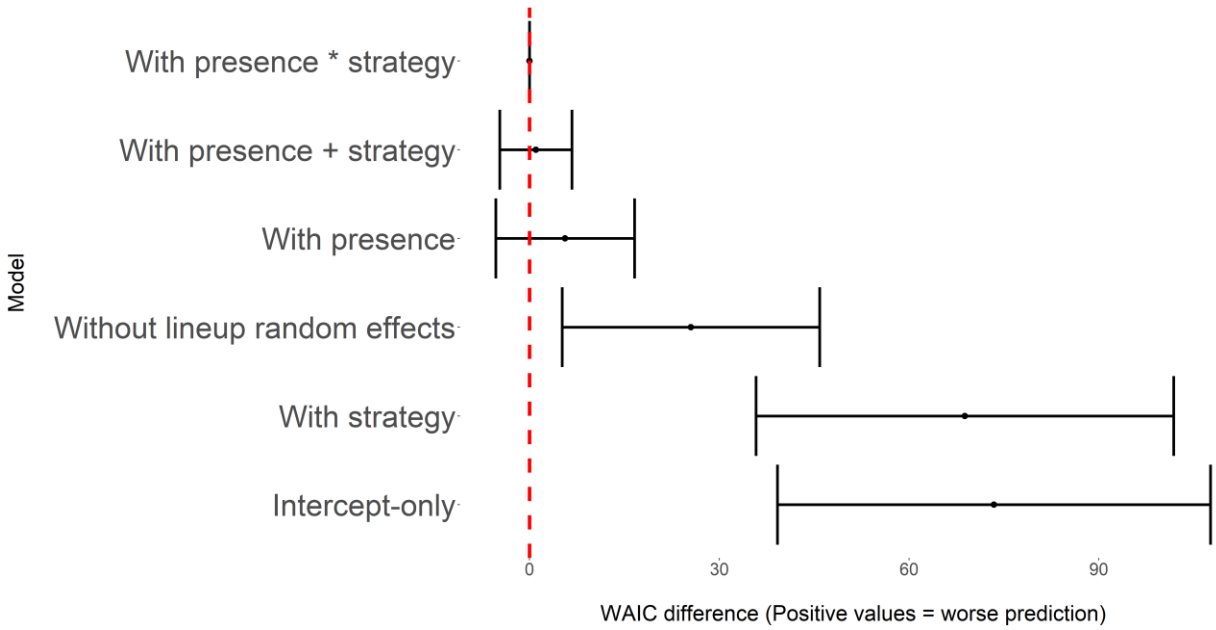


Figure 22. Experiments 1 & 2: WAIC model comparison results for predicting p(ID). Models presented in descending order of predictive ability relative to the best model (1st Row). Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

In interpreting these results, we can see that although a model including Strategy and Presence main effects and their interaction is the best model, removing the interaction and Strategy main effect do not decrease predictive accuracy substantially. Removing presence (i.e., the Strategy- and intercept-only models) does seem to noticeably degrade accuracy. Together, these again suggest small or insubstantial strategy differences and a larger Presence effect. Here,

excluding lineup random effects from the full model does seem to negatively impact performance, suggesting that by-lineup variability plays an important role in the model.

To examine differences in correct and false ID rates between strategies, we estimated posterior distributions for a model with all effects (including lineup random effects as a result of the above comparisons). The results are displayed below in Figure 23.

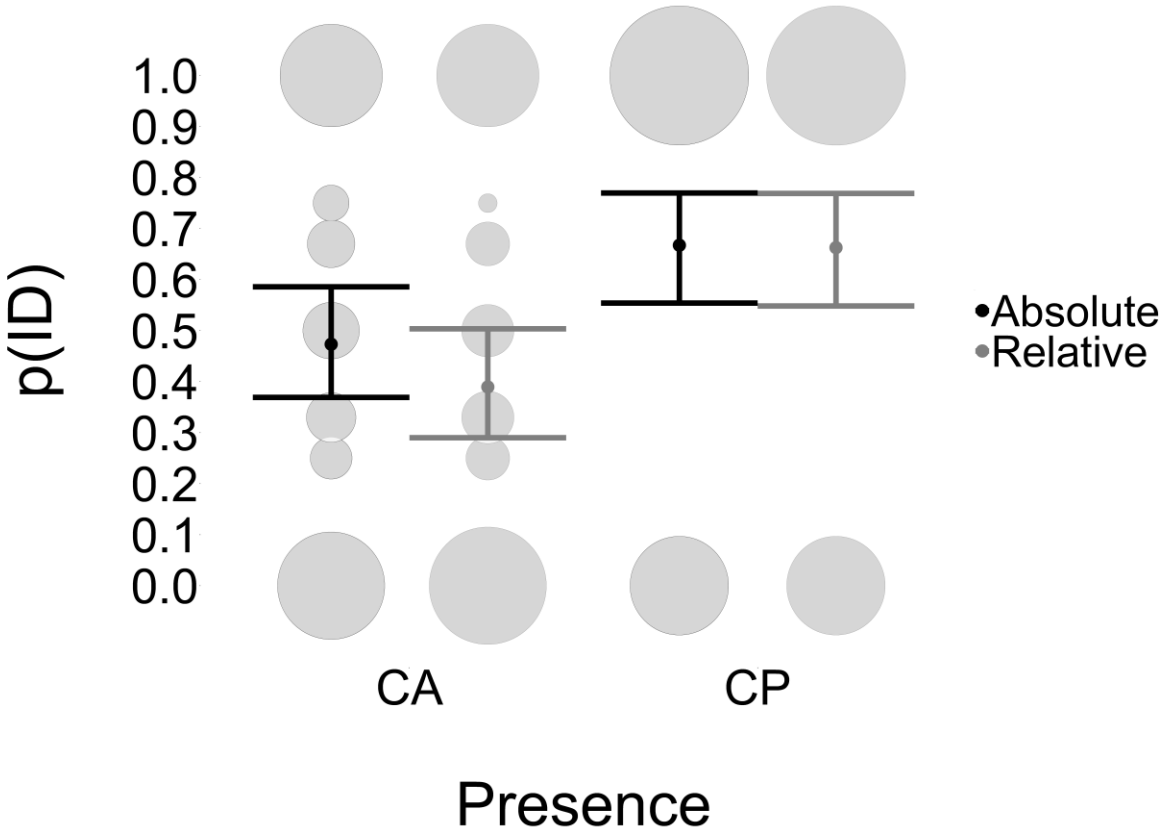


Figure 23. Experiments 1 & 2: Probability of making an ID by Strategy and Presence. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models). Shaded circles represent the relative proportion of participants who gave responses at each probability.

One thing that is immediately apparent is the amount of variability in the estimates—this is due to the fact that posterior estimates were allowed to vary by lineup. Even with the increased variability, we see some evidence of a slightly lower rate of CA false IDs in the relative condition compared to the absolute (estimated difference = 8%, 95% CR[3%, 14%])[14]. The BF in favor of this difference was 25. Conversely, there do not appear to be strategy differences for CP correct IDs (estimated difference = 0%, 95% CR[-8%, 7%]), with a BF of 4 in favor of *no* difference. Thus, relative rank-order lineups may slightly reduce false IDs while leaving correct IDs largely intact.

Because of the high amount of between-lineup variability, we also conducted an exploratory model including lineup as a fixed factor. The estimation results for that model are shown below in Figure 24:

[14] This difference was more evident in Experiment 1, perhaps due to the larger lineup sample size (4 CA lineups per participants instead of 1). However, estimates of the CA strategy difference were similar in both experiments.

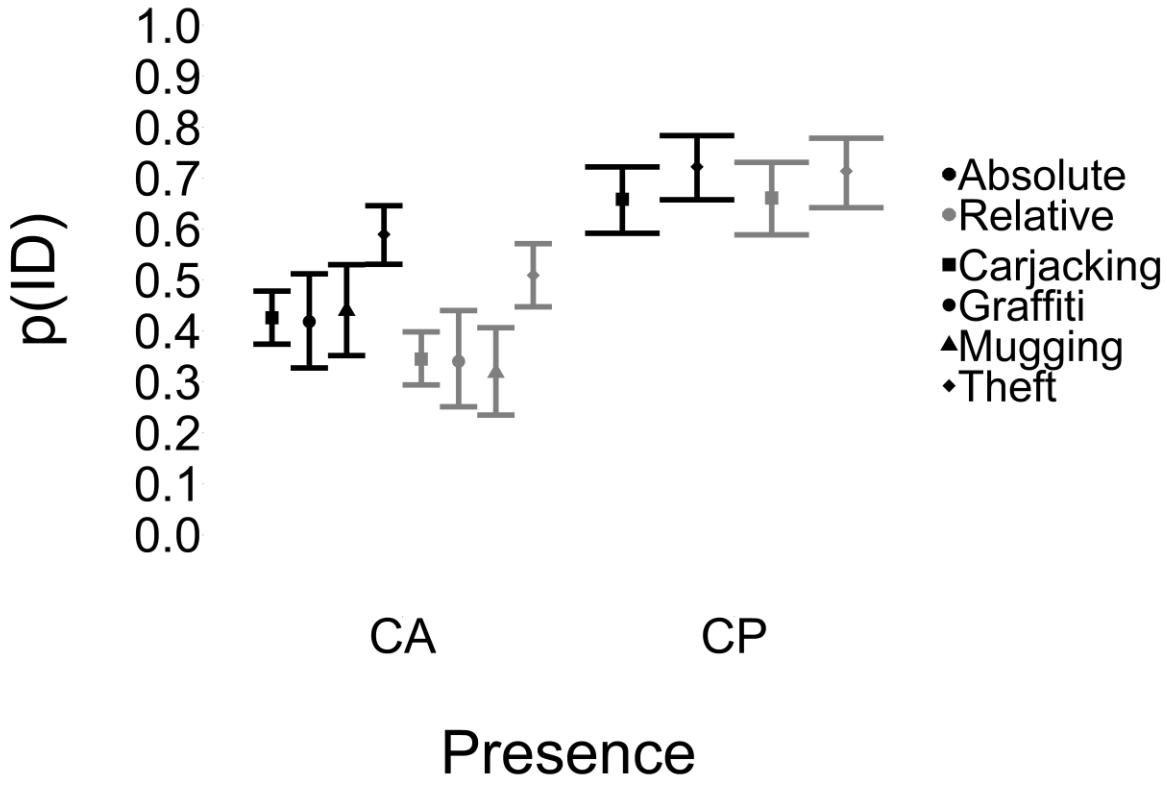


Figure 24. Experiments 1 & 2: Probability of making an ID by Strategy, Presence, and Lineup. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models).

Despite the variability among lineups, the patterns described above appear to be relatively consistent across the lineups: For all CA lineups, the probability of making a false ID is slightly reduced in the relative rank-order condition, and for all CP lineups, the probability of making a correct ID is quite similar.

In order to examine whether the judgment strategies affected filler IDs, we repeated the same analyses, but with p(Filler ID)[15] as the dependent variable. For these analyses, everything but the priors were identical to the analyses of p(ID). Our filler analysis priors differed only in the assumptions we made about Presence effects—instead of assuming an increase in p(ID) for CP lineups, our filler priors assumed an equivalent decrease in p(Filler IDs). This prior change was based on the assumption that participants would be less likely to identify a filler if the true culprit was in the lineup. Otherwise, we conducted model comparisons, estimated posterior distributions, and computed BFs as usual. The results of the model comparison are shown below in Figure 25.

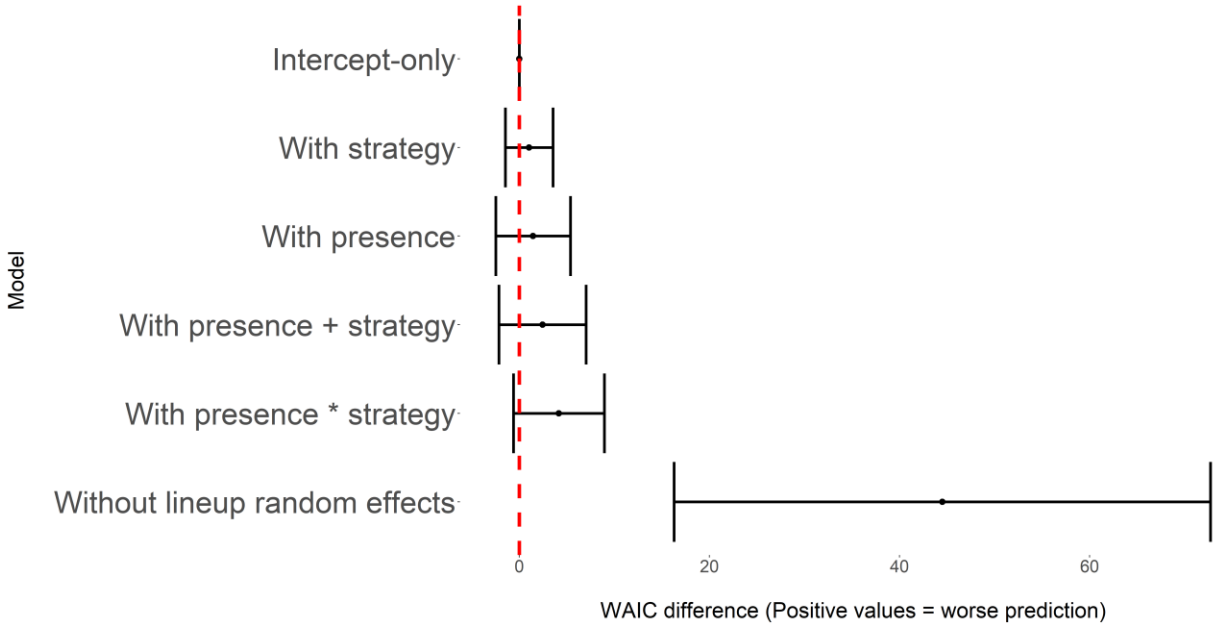
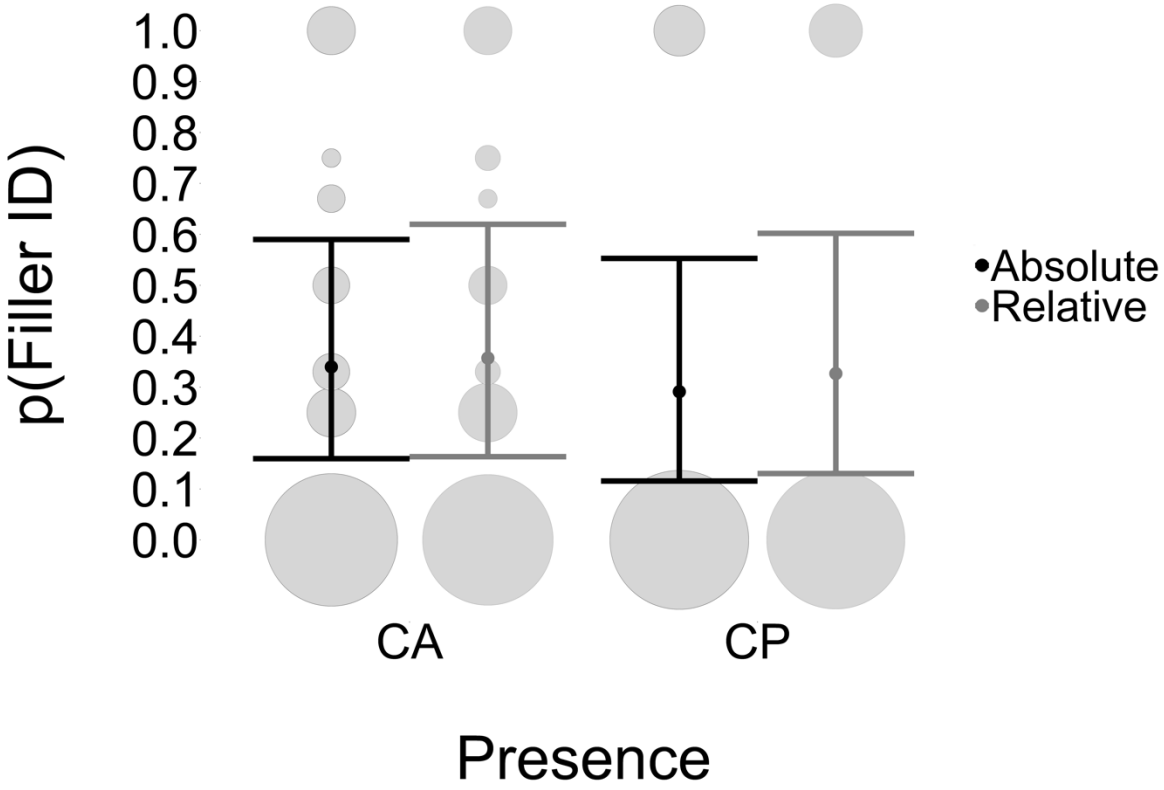


Figure 25. Experiments 1 & 2: WAIC model comparison results for predicting p(Filler ID). Models presented in descending order of predictive ability relative to the best model (1st Row).

[15] A dichotomous dependent variable with possible values Yes (identified a filler) and No (identified a suspect/culprit or rejected the lineup).

Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

The fact that the intercept-only model (without any Strategy or Presence effects) emerged as the most predictive suggests minimal if any substantive effects. Including or excluding these effects had little impact on the model's predictive ability. However, removing lineup random effects noticeably degraded predictive ability[16], again suggesting substantial by-lineup variability. The results of the estimation analysis for the full model (including lineup random effects) are shown below in Figure 26.



[16] Moreso for Experiment 1, again likely due to the greater number of lineups to introduce variability.

Figure 26. Experiments 1 & 2: Probability of making a filler ID by Strategy and Presence. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models). Shaded circles represent the relative proportion of participants who gave responses at each probability.

Again, the by-lineup variability makes interpretation somewhat difficult. It appears that the relative rank-order condition may slightly increase the probability of a filler ID. However, in this model the estimated strategy differences were quite small, for CA filler IDs the difference was 2% (95% CR[-5%, 9%]), and for CP filler IDs the difference was 4% (95% CR[-7%, 15%]). The BF in favor of *no* difference was 2.7 for CA lineups and 2.1 for CP lineups. The combination of small estimated effects and weaker BF evidence against effects suggests that there may be a small but unreliable strategy difference in p(Filler IDs). As with p(IDs), we ran a version of this model including lineup as a fixed effect. The results are displayed below in Figure 27.

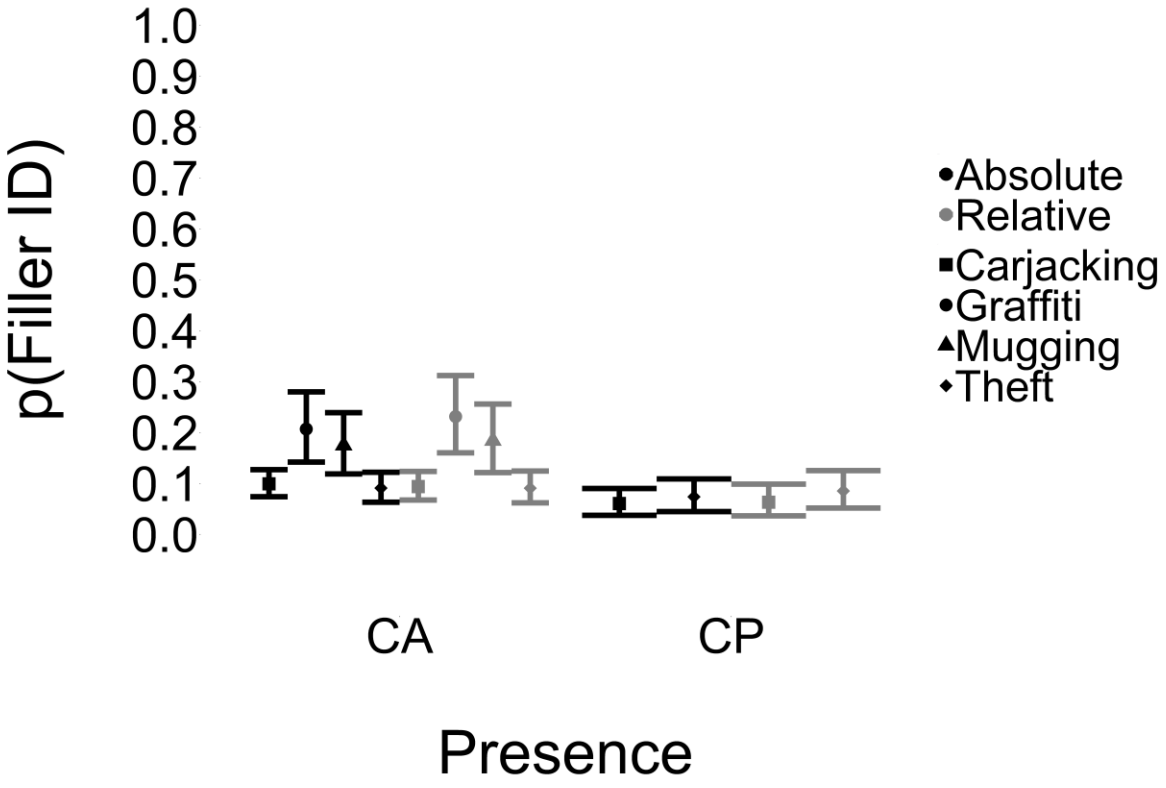


Figure 27. Experiments 1 & 2: Probability of making a filler ID by Strategy, Presence, and Lineup. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models).

As with culprit and suspect IDs, the fixed-lineups analysis reveals a decent amount of by-lineup variability, but consistently trivial strategy effects across lineups. Interestingly, they also show lower overall filler ID rates than the full model analysis (which might suggest that the model has trouble with random lineup intercepts, or with the positively skewed distribution of filler IDs). Overall, we think it unlikely that strategy substantially affects filler ID rates.

Together, our correct, false, and filler ID analyses point to a slight relative strategy accuracy advantage—namely, a small reduction in false IDs with no meaningful differences in correct or filler IDs. However, measures of correct and false ID rates cannot give us the whole picture when it comes to lineup accuracy. For that, we needed to consider discriminability using SDT-based measures.

H2: We hypothesize either no difference or an increase in guilty-innocent discriminability (as measured using ROC curve AUC and DPP) when participants are instructed to use a relative judgment strategy (with rank-ordering) than when participants are instructed to use an absolute judgment strategy. To estimate guilty-innocent discriminability for absolute and relative judgment strategies, we constructed ROC curves. These curves are displayed below in Figure 28:

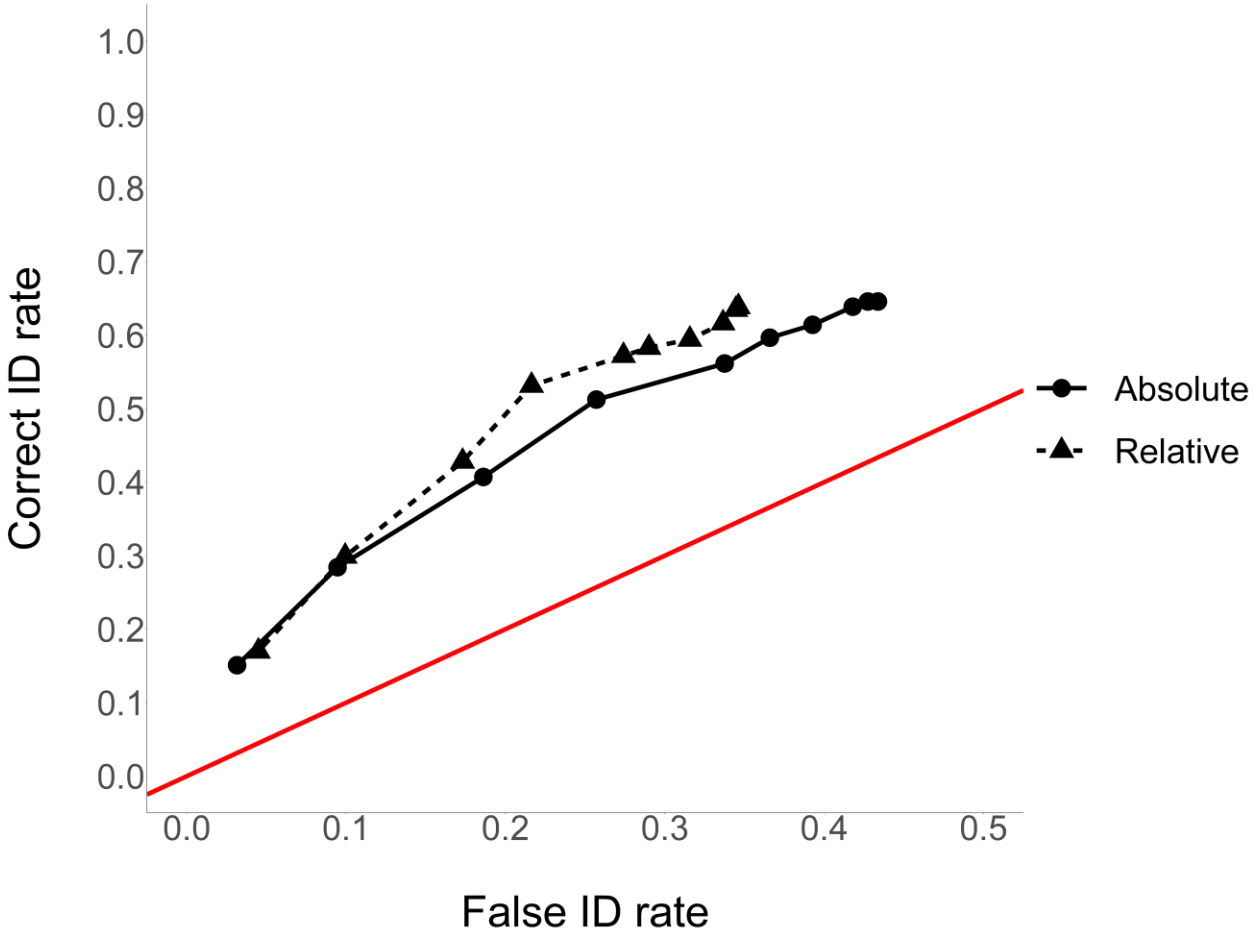


Figure 28. Experiments 1 & 2: ROC curves for absolute and relative judgment strategies.

Consistent with our previous findings that relative rank-order lineups appear to result in a slightly lower false ID rate with no difference in correct ID rate, there appears to be a slight discriminability advantage for relative rank-order lineups[17]. However, this difference was small—the bootstrapped mean estimated difference in partial area under the curve (pAUC) was .02 (95% CR[-.02, .07]. As revealed by the only pre-registered NHST analysis we report, this

[17] When analyzing the Experiment 2 results alone, this advantage was only present at more liberal levels of response criterion. At more conservative levels, absolute lineups had a slight advantage. However, the overall difference in pAUC in Experiment 2 was small, and not statistically significant (see Supplementary Material).

difference was not significant, $p = .33$[18]. We also estimated deviation from perfect performance (DPP; Smith et al., 2019), an alternate measure of lineup utility that better accounts for differential ID rates (as we had with our experiment). This difference was also small, with a mean bootstrapped estimate of the difference of .05 (95% CR[.02, .09])[19]. Although the estimated CR on the DPP did not contain zero, the lack of a significant difference in pAUC and the visually similar ROC curves weigh against a discriminability advantage for the relative rank-order procedure. If the procedure does increase discriminability, this increase is slight, and less than that observed in other experiments (Carlson et al., 2019a).

Discussion

Taking all our results together, we propose the following general conclusions:

1) **Contra the theory of eyewitness judgment strategies, the use of a relative judgment strategy (with rank-order lineups) does not inherently increase false identifications. If anything, this strategy may slightly reduce the likelihood of a false identification.**

The primary theory that we set out to test was the theory of eyewitness judgment strategies, which posits that relative judgment strategies lead to an increase in false IDs, relative to absolute judgment strategies (Lindsay & Wells, 1985). Our results do not support this conclusion—we observed a relative strategy decrease in false ID rates of 8%, with the true decrease likely between 3% and 14%. If sequential and simultaneous lineups are meant to be proxies for absolute and relative strategies, respectively, we would expect to see a relative strategy increase in false IDs in the neighborhood of 22% (Steblay et al., 2011). Our evidence against a relative

[18] We adopted NHST for this analysis because it was not clear that Bayesian estimation of the bootstrapped data would be appropriate.

[19] We deviated from the original pre-registered plan to analyze DPP data via Bayesian estimation and compute a Bayes factor for the DPP difference. However, as with the pAUC analysis, it was not clear that Bayesian estimation of bootstrapped data would be appropriate.

strategy advantage is particularly damning in light of our lineup manipulation. In our culprit-absent lineups, the suspect was chosen to be much more similar to the culprit than the other lineup members. By maximizing the suspect-filler similarity gap, we theoretically increased the discrepancy between absolute and relative strategy outcomes as predicted by models of the strategies (Clark, 2003). However, even in this "worst-case" scenario for relative judgments, we failed to observe any evidence for an increase in false IDs.

Neither did we observe the expected increase in correct IDs (estimated at 8%; Steblay et al., 2011) thought to accompany simultaneous lineups and relative strategy use, or any compelling strategy effects on filler choosing. Rather, our results were more in line with the only other published direct test of the two strategies (Moreland & Clark, 2019). In this study (where judgment strategy was manipulated via instructions), the authors observed similar rates of false IDs and slightly higher correct IDs with relative strategy instructions.

Two possible explanations for the discrepancy between these results and the theory of eyewitness judgment strategies seem plausible: First, instructions-based manipulations are not strong enough to shift eyewitness judgment processes, and second, the judgment strategies do not operate in the way predicted by the theory (Moreland & Clark, 2019). We consider the first possibility in our limitations section. For now, we turn to the second possibility by considering an alternate theory.

2) **However, a relative judgment strategy with rank-order lineups does not appear to increase discrimination/accuracy (as per signal detection theory and the diagnostic-feature-detection hypothesis would predict).**

The second theory we sought to test was the SDT-informed diagnostic-feature-detection model of lineup decision making (Wixted & Mickes, 2014). Recall that this theory proposes that the

comparisons inherent to relative judgment strategies allow eyewitnesses to pick out and discount non-diagnostic information (i.e., shared features). By highlighting diagnostic and non-diagnostic information, relative judgment strategies and simultaneous lineups should theoretically increase guilty-innocent discrimination (Wixted & Mickes, 2014).

However, we failed to observe compelling evidence in favor of a relative strategy discrimination advantage. This was the case even with our rank-order manipulation, which was previously resulted in a dramatic discrimination advantage (Carlson et al., 2019a; see Figure 29 below).

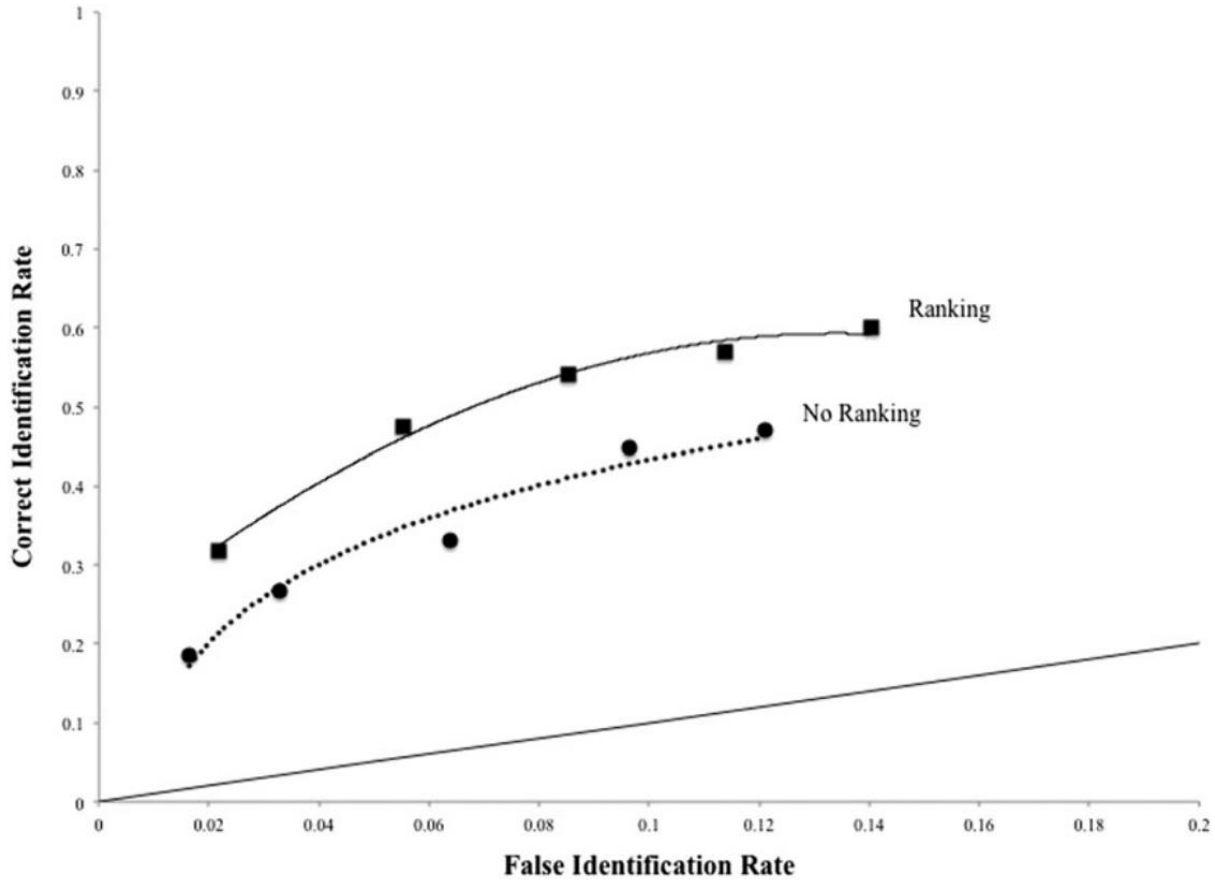


Figure 29. ROC curves for rank-order and no-rank-order lineups (Carlson et al., 2019a)

This study, and others comparing sequential and simultaneous lineups via ROC (Mickes et al., 2012; Wixted & Mickes, 2014), have found a large discriminability advantage for simultaneous lineups/relative strategies. Why then did we observe a trivial difference between the strategies (Figure 28)? Again, it is possible that instructions-based manipulations are too weak to substantially affect lineup outcomes. The similarly unimpressive discrimination results from the previously-mentioned instructions-based study (Moreland & Clark, 2019) suggest that this possibility is a salient one (see Figure 30 below).

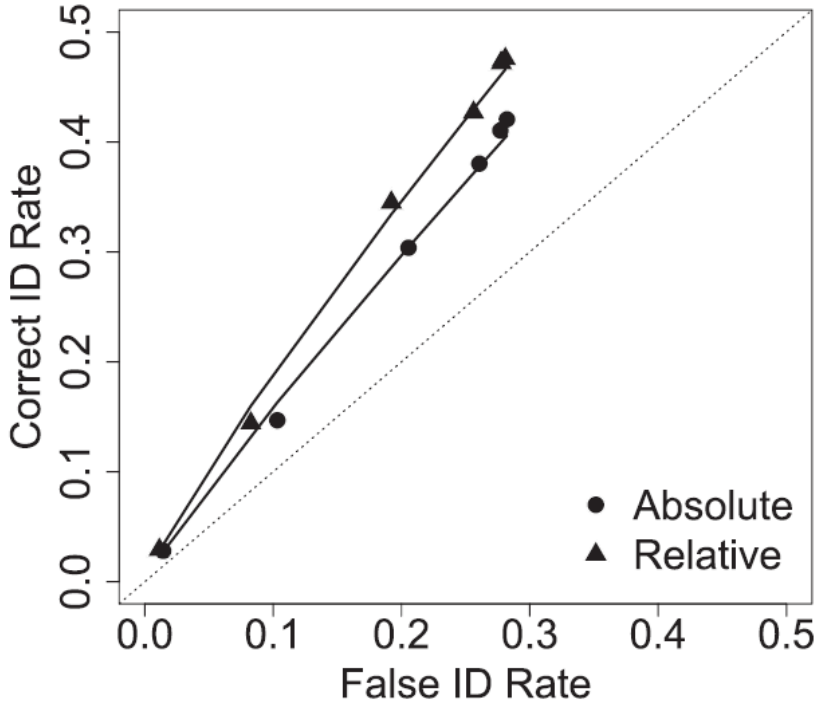


Figure 30. ROC curves for absolute and relative lineups (Moreland & Clark, 2019)

For reasons we will discuss in the limitations section, we do not think this possibility fully explains our discrepant results.

Rather, we suggest the following: *the nature of our lineup manipulation—specifically the fact that the lineups were unfair—reduced the benefits of a relative judgment strategy*. Our lineups

were importantly different from those in the previously mentioned studies, which used fair lineups in which the suspect/culprit did not stand out (Carlson et al., 2019a; Mickes et al., 2012; Moreland & Clark, 2019). How might lineup fairness moderate strategy-based effects? Recall that the diagnostic-feature-detection model explains relative strategy advantages in terms of shared and non-shared feature identification in lineups (Wixted & Mickes, 2014). To the extent that eyewitnesses are able to identify shared features in a lineup, their reliance on these features will be reduced, increasing discrimination. However, this discrimination increase relies on all lineup members sharing the general characteristics of the perpetrator so eyewitnesses can discount these coarse features in favor of features distinctive to their memory for the culprit (Carlson et al., 2019b). The importance of lineup "propitious heterogeneity" (Wells, Rydell, & Seelau, 1993)—variability among lineup members within the constraints of a general description—has been supported empirically. Carlson and colleagues (2019b) found that as fillers became *too* similar to each other and the suspect, discrimination decreased. At the other end of the spectrum, Colloff and colleagues (2016) found that discrimination suffered when the suspect stood out as a result of a distinctive feature.

Thus, the benefits of relative judgments may be contingent on a delicate balance between a basic level of within-lineup similarity and a healthy amount of within-lineup variation. Our lineups decidedly did not strike this balance—they were explicitly designed so that fillers did *not* match a basic description of the culprit. Because lineup members did not share basic features, our lineups did not allow participants to easily extract diagnostic and non-diagnostic information from the lineups. As a result, even if our relative rank-order manipulation did promote relative strategy use, our lineups stacked the deck against relative strategy benefits. This explanation is consistent with the previous findings of rank-order benefits with *fair* lineups (Carlson et al.,

2019a) and with the predictions of the DFD hypothesis (Carlson et al., 2019b; Wixted & Mickes, 2014).

It is worth noting that even if the relative judgments rank-order procedure did not increase discriminability in our study, it did not decrease it (compared to absolute judgments)—another mark in favour of the DFD account and against the theory of eyewitness judgment strategies. With unfair lineups, instead of increasing discriminability, the rank-order procedure may operate differently. One possibility is that as eyewitnesses spend more time with the lineup (and presumably more time comparing lineup members), they become more aware of the *dissimilarity* in the lineup and the fact that one member stands out. This increased awareness of lineup meta-features may make eyewitnesses more cautious and less likely to make a false ID.

Additionally, we believe our data justify the following supplementary conclusions:

3) **Absolute and relative judgment strategies do not substantially affect eyewitness metacognitive judgments of confidence and difficulty. Decisions made on relative rank-order lineups may be made with slightly less confidence and slightly more difficulty (emphasis on "slightly").**

We conducted supplementary analyses of decision confidence and difficulty as a function of strategy (Supplementary Material Section D), and did not find compelling evidence for differences. Admittedly, we had little reason to believe that the strategies would differ—current theory and models of the strategies (e.g., Clark, 2003; Fife et al., 2013) do not predict that one strategy results in higher confidence or more difficult decisions. Our results suggest that confidence and difficulty may not be useful markers for identifying differential strategy use (versus other potential markers such as reaction time, argued to be faster for absolute and accurate decisions; Dunning & Stern, 1994; Sporer, 1993).

4) **Multi-lineup designs represent a departure from real-world eyewitness scenarios. At the very least, those using multi-lineup designs should examine both lineup-specific effects and order effects.**

We observed both substantial by-lineup variability in our main analyses as well as a suggestions of potential order effects in our supplementary analyses (Supplementary Material Section D). Although the (lack of) strategy effects we observed were relatively consistent across lineups, there was substantial variability in the baseline probability of culprit/suspect/filler IDs. This is not surprising—no two crimes are the same in terms of the estimator variables (e.g., presence of a weapon, viewing distance) that can affect eyewitness accuracy independent of any manipulations. Manipulations *should* be examined using a variety of crime video and lineup stimuli. However, a failure to account for stimulus-level variability may obfuscate more substantive effects. This point speaks more broadly to the need for more sophisticated mixed-effects models (McElreath, 2019).

Previous research suggests that decision processes for multi-lineup designs are no different from those in single-lineup designs (Mansour et al., 2017). However, Mansour and colleagues (2017) tested a multi-lineup design with lineups after each crime video. In our experiments, participants viewed all crime videos before they saw any lineups. In our multi-lineup design, we found some evidence for order effects, with slight but consistent reductions in choosing on the 2nd lineup. One possible explanation for this pattern is interference buildup, where each additional crime video interferes with memory for the last (Mansour et al., 2017). It could be that multi-lineup interference reduced participants' memory strength on subsequent lineups, making them more reluctant to choose. However, if this explanation is correct, it is not clear why we did not observe a clear decrease in choosing over the four lineups in Experiment 1

(the most evident decrease was from the 1st to 2nd lineup). A Mansour-and-colleagues-type study specifically testing this multi-lineup design would be a valuable addition to the literature. In the meantime, researchers using multi-lineup designs should investigate potential order effects. Alternatively, with the increasing availability of large online convenience samples, it may be more viable to conduct large-sample single-lineup studies. For our purposes, we note that the order effects we observed were small, and did not appear to moderate our more substantive effects.

Limitations & potential criticisms

1) **Our lack of observed strategy differences can be explained by a weak strategy manipulation**.

As mentioned previously, it is entirely possible that our instructions-based strategy manipulation was not strong enough to have a meaningful effect on eyewitness decision processes. It could be that lineup format (i.e., simultaneous lineups) and construction (i.e., unfair lineups) pushed participants to use a particular decision strategy independent of the instructions we gave them. This is a particularly salient possibility for our absolute strategy condition—though we instructed participants *not* to compare lineup members, it could be that the simultaneous format made such comparisons inevitable. Finally, though participants reported adhering at least partly to their assigned strategy on the vast majority of lineups, it is possible that eyewitness self-reports of strategy use are unreliable, or that metacognitive absolute and relative decision processes are hard to access and judge accurately.

Although we cannot fully rule out the possibility that instructions-based manipulations are insufficient to shift eyewitness judgment strategies, recent research suggests that meta-memorial instructions *can* meaningfully affect eyewitness decision making and lineup outcomes.

For example, if participants are given phenomenological descriptions of how accurate remembering “feels”, they show improved discrimination on lineups (Guerin et al., 2018). As to the possibility that simultaneous lineups inherently promote a relative judgment strategy, we point to 1) research showing that participants completing simultaneous lineups self-reported a variety of different judgment strategy characteristics reminiscent of both absolute and relative strategies (Dunning & Stern, 1994) and 2) research showing that slight modifications to the simultaneous lineup procedure (e.g., confidence ratings for all lineup members, rank-order lineups) can result in substantively different lineup outcomes (Brewer et al., 2012; Carlson et al., 2019a). Even though the only other published study to use absolute and relative strategy instructions found minimal strategy differences, there were reliable differences under specific conditions (e.g., increased accuracy with relative strategy instructions when a suspect was surrounded by high-similarity foils; Moreland & Clark, 2019).

We also think it unlikely that participants could not accurately report on their own strategy use—previous research has shown that self-reported eyewitness decision processes are associated with accuracy, suggesting that the self-reports are not without merit (Dunning & Stern, 1994). Additionally, our strategy instructions were designed to be simple and straightforward, increasing the likelihood that participants would be able to adopt and report on the strategies.

Finally, the strongest point against a weak-manipulation explanation for our results is the fact that another large-sample study found large strategy differences in the predicted direction using the same rank-order lineup manipulation that we adopted (Carlson et al., 2019a). These findings suggest that judgment strategy is not wholly determined by lineup format, and that the rank-order procedure does in fact promote relative judgments. All told, we believe that the most

plausible explanation for our lack of findings is the unfair-lineups explanation that we described previously. That is, unfair lineups reduce the amount of diagnostic information available to the participant, potentially attenuating any diagnostic-feature-detection based benefits of the relative rank-order procedure.

2) **Our lack of strategy differences can be explained by peculiarities of the stimuli that we used.**

One potentially noteworthy feature of the stimuli that we used was the fact that the crime videos were in color but the lineup photos were black-and-white. Aside from the fact that memory performance for studied faces may decrease when color modality differs from study to test (Wogalter & Laughery, 1987), the modality shift may have direct implications in terms of the diagnostic-feature-detection hypothesis. By removing color from the lineup, one removes information (both diagnostic and non-diagnostic) that might otherwise contribute to discrimination. In other words, the black-and-white photos may have reduced the DFD-based benefits of our rank-order condition in a similar (but likely lesser) way that our unfair lineups might have.

Aside from this concern, there are broader constraints on generality that accompany the use of any particular stimulus set. All of the simulated crime videos we used (and the majority of those we have seen in the literature) were non-violent crimes with young, white male perpetrators. Though we have few a priori reasons to expect that our results are idiosyncratic to the stimuli that we used (e.g., the same stimuli have been used to elicit effects with other manipulations; Colloff et al., 2016), it is unclear whether our results generalize to different contexts (e.g., violent crimes or crimes with a weapon, where memory for culprits may be negatively affected; Granhag et al., 2014; Laney & Loftus, 2018). Despite this, the fact that we

used different non-violent crimes that varied meaningfully and found similar results safeguards us against a complete lack of generalizability.

The exclusive use of white male perpetrators in our (and other) experiments does raise interesting questions about ingroup/outgroup biases. For instance, the Own-Race Bias (ORB)—worse memory for other-race than own-race faces—is well-documented (Sporer, 2001; Wilson, Bernstein, & Hugenberg, 2016). We did not collect ethnicity data in our non-pilot studies, but it is likely that our student and online samples were majority-White samples. Though one might expect that overall performance would be reduced in a different sample or with non-White culprits, it is not clear what effect the ORB might have on our strategy manipulation. Because the DFD hypothesis predicts that discriminability will improve to the extent that eyewitnesses can identify shared and non-shared facial features, it is possible that diagnostic feature identification is hampered when viewing other-race faces due to less efficient processing (Wilson et al., 2016). This is a possibility worth investigating.

Although we could not address the ORB in the current study, we did collect self-reported gender for all participants. There is some evidence for an "Own-Gender Bias" (OGB; Wright & Sladden, 2003), whereby females tend to remember more faces than males and have better memory for female than male faces (though meta-analytic evidence suggests there may not be a male own-gender bias; Herlitz & Lovén, 2013). Similar to the ORB, it is possible that the OGB could moderate strategy-based effects. To investigate this possibility, we conducted our primary analysis of identification probability by Strategy and Presence, this time including self-reported gender (male, female)[20] as an interacting factor. With this analysis, we found very similar

[20] Because only 6 participants reported a gender identity other than male or female, we did not include theses participants as a separate category for this analysis. Investigation of potential effects in a larger non-binary sample would be interesting, as these participants might view male and female culprits as outgroup members.

response patterns for males and females, suggesting that gender-based intergroup biases did not substantially influence our results (See Supplementary Material D for more details).

Though we think it unlikely that gender-based intergroup biases affected our results, we concede that there are other potential intergroup factors that could have. Beyond basic face recognition and identification, ingroup/outgroup biases may affect more complex eyewitness judgments. For instance, people tend to view in-group perpetrators as less culpable than out-group perpetrators (holding crime constant; Lindholm & Christianson, 1998). Additionally, people have clear stereotypes about what kind of person commits what kind of crime (e.g., suspects viewed as more stereotypically Black were viewed as more likely to commit a drive-by shooting than serial killings), though ethnic majority/minority status may not affect these stereotypes (Davies, Hutchinson, Osborne, & Eberhardt, 2016; Osborne & Davies, 2012). Furthermore, culprits with high perceived phenotypic stereotypicality are viewed as more culpable of stereotypical crimes (Davies et al., 2016). More generally, evaluative associations of outgroups tend to be more stereotypical and negative (Kawakami, Amodio, & Hugenberg, 2017).

All told, the literature on intergroup relations and biases suggests that the field of eyewitness lineup research would benefit from a greater focus on intergroup factors. For instance, using more varied crime events and lineups with culprits of varying age, gender, and race[21] would permit the examination of intergroup effects and test the generalizability of theories of eyewitness judgments. For certain hypotheses like the DFD hypothesis, intergroup effects on face perception and categorization may be especially relevant. One could also measure in-group/out-group identification (e.g., "What do you think the ethnic identity of the perpetrator

[21] One promising example is the Eyewitness Lineup Identification Database (Ribbers, Rubinova, & Fitzgerald, 2019), which includes images and recordings of hundreds of adult males and females of varying ethnicities. We plan to use this database for future studies.

is?", "How much do you identify with your ethnic identity?", etc.)—relating intergroup identification to forensic identification of culprits and suspects.

That said, we think it unlikely that intergroup factors completely explain our results. For one, we found little evidence that gender—a perceptually "basic" social category (Kawakami et al., 2017)—affected overall accuracy or the (lack of) strategy effects. At the very least, we have shown that the relative rank-order procedure has minimal effects that are consistent along one axis of intergroup-ness (gender), holding another axis (race) relatively constant. Insofar as other factors (race, stereotypicality of crime) affect the memory and face perception processes that are central to the theories under examination, we think that it is more likely that such factors will act on memory at a more general level (e.g., worse memory for other-race faces, worse or better memory for less stereotypical crimes) than a specific level (e.g., the processes by which people search for and identify shared facial features). In other words, the degree of intergroup identification with the culprits might serve as a manipulation akin to lineup fairness or degree of within-lineup similarity, which would affect the *degree* to which participants could make use of diagnostic-feature-detection, but not the *nature* of those processes. Conceptualized in this way, our study was likely conducive to a high degree of intergroup-based diagnostic-feature-detection (based on the assumption that most participants were White, like the culprits)—it is just that this facilitation was counteracted by more substantive lineup factors (i.e., unfairness).

3) **Our lineup experiment (and lineup experiments in general) did not approximate real-world identification procedures and may lack ecological validity.**

Though there are several advantages to using videotaped simulated crimes (e.g., ease of administration to large samples, stimulus consistency), videotaped crimes may lack ecological validity. In the real world, eyewitnesses experience crimes live in three dimensions (Wells &

Penrod, 2011). Real crimes involve richer experiences, and are generally more stressful (Deffenbacher, Bornstein, Pendrod, & McGorty, 2004; Pozzulo, Crescini, & Panton, 2008). Some studies have found that stress during encoding of crimes is associated with reduced recognition accuracy (Davis, Peterson, Wissman, & Slater, 2019; Deffenbacher et al., 2004). However, others have found no stress-induced decrements in recognition accuracy (Sauerland et al., 2016; Pozzulo et al., 2008), the confidence-accuracy relationship (Davis et al., 2019), or discriminability (Sauerland et al., 2016). Real-world lineups also differ from those used in experiments such as ours. Experimental participants are typically given instructions in a text-based format, and their motivation to be accurate in their identification may be low. Contrast this with real-life lineups, where eyewitnesses are typically instructed by a uniformed officer, and where their involvement in the crime and motivation to be accurate are likely higher. As a result, it is unclear whether the results of videotaped-crime studies generalize to real-world scenarios.

Most eyewitness studies use videotaped-crimes, though some have used live staged crimes (e.g., Wells et al., 1993). The use of live staged crimes comes with its own challenges—they require continued re-enactments (potentially leading to session-to-session differences), can be time-consuming to administer, typically limit the size of groups one can test (a problem in lineup research, where sample size requirements are large given the low number of data points per participant), run the risk of participants intervening in the staged crime, and may engender other ethical concerns (Wells & Penrod, 2011). As a result, it is not clear whether the gain in ecological validity is worth the trade offs. To our knowledge, only two studies have directly compared eyewitness performance in live and videotaped crimes (Ihlebaek, Løve, Eilertsen, & Magnussen, 2003; Pozzulo et al., 2008). The first study found lower accuracy in the live staged crime than in the video, but similar patterns of memory errors (Ihlebaek et al., 2003). The second

found no substantial differences between the two modalities (Pozzulo et al., 2008). These results are promising in light of the challenges of live staged crime experiments, but it is worth noting that the sample sizes in these studies were relatively small (N = 106, 104 respectively). As a result, further research on format effects with larger sample sizes is necessary.

One emerging potential avenue for large-sample live-staged-crime experiments is crowdsourced research, which involves running experiments across a distributed network of collaborating laboratories (Moshontz et al., 2018). Crowdsourced research networks such as the *Psychological Science Accelerator* (https://psysciacc.org/) distribute the challenges of hard-to-run experimental designs across multiple labs, allowing one to collect larger samples. It would be particularly exciting to see a large-scale collaborative comparison of live and videotaped crimes, and of text-based versus live (or video-based) lineup strategy instructions. Of course, there will still be challenges with distributed staged-crime research (e.g., ensuring consistency across staged crimes, possible cross-cultural effects), but these challenges should not be insurmountable.

Future directions

Though the current study provided insight into two influential theories in the eyewitness decision making literature, several open questions remain. Here we propose a number of potential fruitful avenues of further research.

1) **Replicating the rank order effect with fair and unfair lineups.**

We argue that our lack of strategy differences is due to an interaction between lineup fairness and strategy manipulation—specifically, that the rank-order procedure results in increased discriminability only under fair-lineup conditions. This conclusion is based on a cross-experiment, cross-material comparison with the sole published rank-order study (Carlson et al.,

2019a). In order to fully support this conclusion, we need a within-experiment comparison where the strategy manipulation is crossed with lineup fairness. Such an experiment would serve to directly test our proposed explanation. It would also test the robustness of the relative rank-order discrimination advantage, which has not yet been replicated.

2) **Testing the diagnostic-feature-detection hypothesis with sequential lineups.**

One of the potential limitations of our study was the exclusive use of simultaneous lineups. Though the idea was to manipulate strategy independent of lineup format, we had to choose *a* format to use, and it is possible that our choice of simultaneous lineups introduced potential format-based confounds. One possibility is to test our hypotheses with sequential lineups. That is, instead of trying to make simultaneous lineups (which may involve some inherent relative strategy use) more absolute, one could try to make sequential lineups more relative. To test the predictions of the diagnostic-feature-detection hypothesis with sequential lineups, one could provide eyewitnesses with the diagnostic information that is normally missing in sequential lineups. This information could be presented in the form of a reminder that all lineup members that participants are about to view will share certain features. If the DFD hypothesis is true, then participants given this information should be able to discount these features when viewing a sequential lineup, resulting in increased discrimination (relative to participants who view a sequential lineup without diagnostic information). This experiment would better allow the disentanglement of strategy- versus format-specific effects.

3) **Further direct experimental tests of the diagnostic-feature-detection hypothesis.**

The DFD literature, while promising, suffers from the same lack of direct tests as the eyewitness judgment strategy literature. Our study, while closer to a direct test than other manipulations (e.g., simultaneous vs. sequential, lineup vs. showup, rank-order without strategy

instructions), is still not as direct as it could be. There are several possibilities for more direct tests of DFD-based predictions. First, one could tailor instructions that are directly informed by the DFD hypothesis. Instead of instructing participants to simply compare lineup members, one could explicitly instruct participants to identify features shared by lineup members, and to discount these features when making an identification decision. Instructions could vary in specificity (e.g., fully explaining the DFD hypothesis, or simply telling participants to identify and discount shared features). Second, researchers can implement more direct control of within-lineup similarity, which would allow the subtle manipulation of similarity variables and testing of more specific DFD predictions (e.g., as in Carlson et al., 2019b where artificial face stimuli were used to allow tight control of similarity). Third, researchers could adopt alternate behavioural and self-report measures in concert with traditional lineup measures to better understand judgment processes. Eye-tracking has shown some promise in lineup studies, with dwell time being related to both within-lineup similarity and the probability of a positive identification (Flowe & Cottrell, 2010), and fixation duration revealing participant concealment of culprit recognition (Millen & Hancock, 2019). In the context of the DFD hypothesis, eye-tracking could be used to determine whether participants exposed to a DFD-informed manipulation engage in more visual comparisons of lineup members, and whether they fixate on diagnostic/non-diagnostic facial features. Future studies could incorporate qualitative debriefing or post-lineup questions querying participant strategy use, which could then be coded for content and compared to theoretical accounts. Fourth, if basic effects can be established in large laboratory samples, the generalizability of these effects could be tested with more varied stimulus sets and measures of intergroup factors. Finally, smaller focused and targeted studies

using staged crimes and officer-administered instructions[22] would verify whether laboratory effects generalize to real-world contexts.

4) Formal computational modeling of the diagnostic-feature-detection hypothesis.

Though psychology as a field is now cognizant of and taking steps to address an ongoing "replication crisis" (Lindsay, 2015; Shrout & Rodgers, 2018), some argue that there is a lurking "theory crisis" (Szollosi et al., 2019; Oberauer & Lewandowsky, 2019). These (and other) authors argue that while questionable research practices harm replicability, the real culprits are vague theories and a lack of strong links between theory and experimental tests. One proposed solution to this crisis is computational modeling. Formal computational modeling of theories forces researchers to construct theories such that they engender more precise testable hypotheses (Oberauer & Lewandowsky, 2019). Though modeling work has been done for the theory of eyewitness judgment strategies (e.g., Clark, 2003; Kaesler et al., 2017), the same is not true for the newer DFD hypothesis. Although the original DFD proponents (Wixted & Mickes, 2014) propose a computational-model-like framework that involves assigning varying weights to diagnostic and non-diagnostic features, a formal model has not been validated. If a formal model was constructed, one could then fit it to the large corpus of existing simultaneous/sequential lineup data as well as newer non-direct tests of the DFD hypothesis. Particularly exciting is the possibility of comparing DFD-based models to absolute/relative models as a direct test of the two theories.

Conclusion

[22] Given current events in the United States and elsewhere, and the shifting landscape of police-community perceptions and relationships, care should be taken in implementing officer-administered manipulations and interpreting the results of such manipulations.

Our study provided a direct test of two influential theories of lineup decision making: the theory of eyewitness judgment strategies (Lindsay & Wells, 1985) and the signal detection theory-based diagnostic-feature-detection hypothesis. By directly manipulating strategy use independent of lineup format, we were able to assess the degree to which absolute and relative judgment strategies differed in terms of meaningful lineup outcomes. We found minimal strategy differences in correct, false, and filler IDs, and importantly, no differences in discrimination and lineup accuracy. These results cast further doubt on the theory of eyewitness judgment strategies. Though further work is necessary to determine whether our results support the diagnostic-feature-detection hypothesis, we believe that our results are compatible with the hypothesis and provide potential boundary conditions on the benefits of DFD-based manipulations. The current study sets the stage for a number of future research directions, primarily aimed at more directly testing the DFD hypothesis.

Finally, beyond the tests of lineup decision making theory, our study has implications for lineup administration in practice. Previous research (with fair lineups) suggests that relative strategy rank-order lineups increase eyewitness accuracy. Our study (with unfair lineups) did not find increased accuracy, but neither did we observe decreased accuracy. As a result, police adoption of rank-order simultaneous lineups in the field may be desirable. At the very least, we hope that our research and emerging findings can inform and improve current practices.

References

Albright, T.D. (2017). Why eyewitnesses fail. *Proceedings of the National Academy of Sciences of the United States of America, 114*(30), 7758-64. https://doi.org/10.1073/pnas.1706891114

Bauer, G.R., Braimoh, J., Scheim, A.I., & Dharma, C. (2017). Transgender-inclusive measures o f sex-gender for population surveys: Mixed-methods evaluation and recommendations. *PLOS-ONE*. https://doi.org/10.1371/journal.pone.0178043

Brewer, N., Weber, N., Wootton, D., & Lindsay, D.S. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science, 23*(10), 1208-14. https://doi.org/10.1177/0956797612441217

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

Carlson, C.A., Jones, A.R., Goodsell, C.A., Carlson, M.A., Weatherford, D.R., Whittington, J.E., & Lockamyeir, R.F. (2019a). A method for increasing empirical discriminability and eliminating top-row preference in photo arrays. *Applied Cognitive Psychology*. Advance online publication. https://doi-org.ezproxy.library.uvic.ca/10.1002/acp.3551

Carlson, C.A., Jones, A.R., Whittington, J.E., Lockamyeir, R.F., Carlson, M.A., & Wooten, A.R. (2019b). Lineup fairness: Propitious heterogeneity and the diagnostic feature-detection hypothesis. *Cognitive Research: Principles and Implications, 4*(2). https://doi.org/10.1186/s41235-019-0172-5

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

Colloff, M.F., Wade, K.A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science, 27*(9), 1227-1239. DOI: 10.1177/0956797616655789

Colloff, M.F., & Wixted, J.T. (2019). Why are lineups better than showups? A test of filler siphoning and enhanced discriminability accounts. *Journal of Experimental Psychology: Applied*. Advance online publication. doi:10.1037/xap0000218

Cumming, G. (2012). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis.* New York: Routledge.

Davies, P.G., Hutchinson, S., Osborne, D., & Eberhardt, J.L. (2016). Victims' race and sex leads to eyewitness misidentification of perpetrator's phenotypic stereotypicality. *Social Psychology and Personality Science, 7*(6), 491-499. https://doi.org/10.1177/1948550616644655

Deffenbacher, K.A., Bornstein, B.H., Penrod, S.D., & McGorty, E.K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law and Human Behavior, 28*(6), 687-706.

Dobolyi, D.G., & Dodson, C.S. (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification confidence. *Journal of Experimental Psychology: Applied, 19*(4), 345-357. https://doi.org/10.1037/a0034596

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. DOI: 10.1037/0022-3514.67.5.818

Federal/Provincial/Territorial Heads of Prosecutions Subcommittee on the Prevention of Wrongful Convictions (2018). Chapter 3 – Eyewitness Identification and Testimony. In

Innocent at Stake: The Need for Continued Vigilance to Prevent Wrongful Convictions in Canada. Retrieved online at: https://www.ppsc-sppc.gc.ca/eng/pub/is-ip/ch3.html

Fife, D., Perry, C., & Gronlund, S. D. (2014). Revisiting absolute and relative judgments in the WITNESS model. *Psychonomic Bulletin & Review*, *21*(2), 479-487. doi:10.3758/s13423-013-0493-1

Flowe, H., & Cottrell, G.W. (2011). An examination of simultaneous lineup identification decision processes using eye tracing. *Applied Cognitive Psychology*, *25*(3), 443-451. doi: 10.1002/acp.1711

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997-1016.

Granhag, P.A., Ask, K. & Giolla, E.M. (2014). Eyewitness recall: an overview of estimator-based research. In T. J. Perfect & D.S. Lindsay (Eds.), The SAGE handbook of applied memory (pp. 541-558). London: SAGE Publications Ltd. doi:10.4135/9781446294703.n30

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

Gronlund, S.D. & Carlson, C.A. (2014). System-based research on eyewitness identification. In T. J. Perfect & D.S. Lindsay (Eds.). The SAGE handbook of applied memory (pp. 595-613). London: SAGE Publications Ltd. doi:10.4135/9781446294703.n33

Guerin, N., Weber, N., & Horry, R. (2018). Metamemory and lineup instructions: Asking eyewitnesses to recollect reduces false identification. Preprints. doi:10.20944/preprints201812.0012.v1

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition, 21*(9-10), 1306-1336. https://doi.org/10.1080/13506285.2013.823140

Horry, R., Brewer, N., & Weber, N. (2016). The grain-size lineup: A test of a novel eyewitness identification procedure. *Law and Human Behavior, 40*(2), 147-158. doi: 10.1037/lhb0000166

Ihlebaek, C., Løve, T., Eilertsen, D.E., & Magnussen, S. (2003). Memory for a staged criminal event witnessed live and on video. *Memory, 11*(3), 319-327. DOI: 10.1080/09658210244000018

Innocence Project. (2019). 'Eyewitness Identification Reform'. Retrieved on May 15th, 2019, from https://www.innocenceproject.org/eyewitness-identification-reform/

Jeffreys, H. (1961). *Theory of probability*. Oxford, England: Oxford University Press.

Kaesler, M., Dunn, J., & Semmler, C. (2017). Using measurement models to understand eyewitness identification. Paper presented at CogSci 2017: 39th Annual Meeting of the Cognitive Science Society, London, UK.

Kawakami, K., Amodio, D.M., & Hugenberg, K. (2017). Chapter one – Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization. *Advances in Experimental Social Psychology, 55*, 1-80. https://doi.org/10.1016/bs.aesp.2016.10.001

Laney, C., & Loftus, E. F. (2018). Eyewitness memory. In R. N. Kocsis (Ed.), *Applied criminal psychology: A guide to forensic behavioral sciences* (p. 199–228). Charles C Thomas Publisher, Ltd.

Lindholm, T., & Christianson, S-Å. (1998). Intergroup biases and eyewitness testimony. *The Journal of Social Psychology, 138*(6), 710-723. doi: 10.1080/00224549809603256

Lindsay, D.S. (2015). Replication in Psychological Science. *Psychological Science, 26*(12), 1827-32. https://doi.org/10.1177/0956797615616374

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

Link, S.W. (1994). Rediscovering the past: Gustav Fechner and signal detection theory. *Psychological Science, 5*(6), 335-340.

Loftus, E.F. (1975). Reconstructing memory: The incredible eyewitness. *Jurimetrics Journal, 15*(3), 188-193.

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), 2235-54. doi: 10.3758/s13428-017-0855-0.

McElreath, R. (2019). Statistical Rethinking: A Bayesian Course with Examples in R and Stan (2nd Edition).

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

Millen, A.E., & Hancock, P.J.B. (2019). Eye see through you! Eye tracking unmasks concealed face recognition despite countermeasures. *Cognitive Research: Principles and Implications, 4*(23). https://doi.org/10.1186/s41235-019-0169-

Moreland, M.B., & Clark, S.E. (2019). Absolute and relative decision processes in eyewitness identification. *Applied Cognitive Psychology, 34,* 142-156. https://doi.org/10.1002/acp.3602

Moshontz, H., Campbell, L, Ebersole, C.R., IJzerman, H., Urry, H.L., Forscher, P.S., Grahe, J.E., McCarthy, R.J., Musser, E.D., Antfolk, J., Castille, C.M., Evans, T.R., Fiedler, S., Flake, J.K., Forero, D.A., Janssen, S.M.J., Keene, J.R., Protzko, J., Aczel, B., . . . Chartier, C.R. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science, 1*(4), 501-515. doi: 10.1177/2515245918797607

National Research Council (2014). *Identifying the Culprit: Assessing Eyewitness Identification.* Washington, DC: The National Academies Press.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review, 26*(5), 1596-1618. https://doi.org/10.3758/s13423-019-01645-2

Osborne, D., & Davies, P.G. (2012). Eyewitness identifications are affected by stereotypes about a suspect's level of perceived stereotypicality. *Group Processes & Intergroup Relations, 16*(4), 488-504. https://doi.org/10.1177/1368430212454927

Permut, S., Fisher, M., & Oppenheimer, D.M. (2019). TaskMaster: A tool for determining when subjects are on task. *Advances in Methods and Practices in Psychological Science, 2*(2). https://doi.org/10.1177/2515245919838479

Pozzulo, J.D., Crescini, C., & Panton, T. (2008). Does methodology matter in eyewitness identification research?: The effect of live versus video exposure on eyewitness identification accuracy. *International Journal of Law and Psychiatry, 31*(5), 430-437. doi: 10.1016/j.ijlp.2008.08.006

Pozzulo, J.D., & Lindsay, R.C.L. (1999). Elimination lineups: An improved identification procedure for child eyewitnesses. *Journal of Applied Psychology, 84*(2), 167-176. DOI: 10.1037/0021-9010.84.2.167

Public Prosecution Service of Canada (2018). Innocent at stake: The need for continued vigilance to prevent wrongful convictions in Canada. Retrieved from: https://www.ppsc-sppc.gc.ca/eng/pub/is-ip/index.html

Rakoff, J.S., & Loftus, E.F. (2018). The intractability of inaccurate eyewitness identification. *Daedalus, 147*(4), 90-98. https://doi.org/10.1162/daed_a_00522

Ribbers, E., Rubinova, E., & Fitzgerald, R. (2019). The Eyewitness Lineup Identification Stimulus Database. Talk presented at the Annual Conference of the European Association of Psychology and Law, Santiago de Compostela.

Rouder, J.N., Morey, R.D., Speckman, P.L., & Province, J.M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356-374, https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypotheses. *Psychonomic Bulletin & Review, 16*(2), 225-237. doi:10.3758/PBR.16.2.225

Sauer, J.D., Palmer, M.A., & Brewer, N. (2018, in press). Eyewitness identification. In N. Brewer & A. Douglas (Eds.), Psychological Science and the Law. Guilford.

Sauerland, M., Raymaekers, L.H.C., Otgaar, H., Memon, A., Waltjen, T.T., Nivo, M., Slegers, C., Broers, N.J., & Smeets, T. (2016). Stress, stress-induced cortisol responses, and eyewitness identification performance. *Behavioral Sciences and the Law, 34*(4), 580-594. doi: 10.1002/bsl.2249

Sauerland, M, & Sporer, S.L. (2007). Post-decision confidence, decision time, and self-reported decision processes as postdictors of identification accuracy. *Psychology, Crime & Law, 13*(6), 611-625. https://doi.org/10.1080/10683160701264561

Schönbrodt, F.D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes Factors: Efficiently testing mean differences. *Psychological Methods, 22*(2), 322-339. https://doi.org/10.1037/met0000061

Shrout, P.E., & Rodgers, J.L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69,* 487-510. https://doi.org/10.1146/annurev-psych-122216-011845

Smith, A.M., Lampinen, J.M., Wells, G.L., Smalarz, L., & Mackovichova, S. (2019). Deviation from perfect performance measures the diagnostic utility of eyewitness lineups but partial area under the ROC curve does not. *Journal of Applied Research in Memory and Cognition, 8*(1), 50-59. https://doi.org/10.1016/j.jarmac.2018.09.003

Sporer, S.L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology, 78*(1), 22-23. https://doi.org/10.1037/0021-9010.78.1.22

Sporer, S. L. (2001). The own-race bias: Beyond recognition of faces in the laboratory. *Psychology, Public Policy, and the Law, 7*, 170–200. doi: 10.1037/1076-8971.7.1.170

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R.C.L. (2011). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 25*(5), 459-473. DOI:10.1023/a:1012888715007

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

Szollosi, A., Kellen, D., Navarro, D.J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Science, 24*(2), 94-95. https://doi.org/10.1016/j.tics.2019.11.009

Wagenmakers, E.-J., Lodewycx, T., Kuriyal, H., Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology, 60*(3), 158-189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & Van Der Maas, H.L.J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), 426-432. doi: 10.1037/a0022790

Wells, G.L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology, 36*(12), 1546-1557. DOI: 10.1037/0022-3514.36.12.1546

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

Wells, G.L., & Penrod, S.D. (2011). Eyewitness identification research: Strengths and weaknesses of alternative methods. In B. Rosenfeld & S.D. Penrod (Eds.), *Research Methods in forensic psychology* (pp. 237-257). John Wiley & Sons Inc.

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78,* 835–844. https://doi.org/10.1037/0021-9010.78.5.835.

Wilson, J.P., Bernstein, M.J., & Hugenberg, K. (2016). A synthetic perspective on the Own-Race Bias in eyewitness identification. In Bornstein, B., Miller M. (eds). *Advances in Psychology and Law (Volume 2).* Springer.

Wixted, J.T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121*(2), 262-276. DOI:10.1037/a0035940

Wixted. J.T., & Mickes, L. (2015). Evaluating eyewitness identification procedures: ROC analysis and its misconceptions. *Journal of Applied Research in Memory and Cognition, 4*(4), 318-323. https://doi.org/10.1016/j.jarmac.2015.08.009

Wixted, J.T., & Wells, G.L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. Psychological Science in the Public Interest, 18(1), 10-65). https://doi.org/10.1177/1529100616686966

Wogalter, M.S., & Laughery, K.R. (1987). Face recognition: Effects of study to test maintenance and change of photographic mode and pose. *Applied Cognitive Psychology, 1*, 241-253.

Wogalter, M.S., Malpass, R.S., & Mcquiston, D.E. (2004). A national survey of US police on preparation and conduct of identification lineups. *Psychology, Crime & Law, 10*(1), 69-82. https://doi.org/10.1080/10683160410001641873

Wright, D.B., & Sladden, B. (2003). An own gender bias and the importance of hair in face recognition. *Acta Psychologica, 114*(1), 101-114. https://doi.org/10.1016/S0001-6918(03)00052-0

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with Discussion). Bayesian Analysis, 13(3), 917-1007.

Supplementary Material

A. Lineups

a. Pilot 3 & Experiment 2

i. Carjacking CP

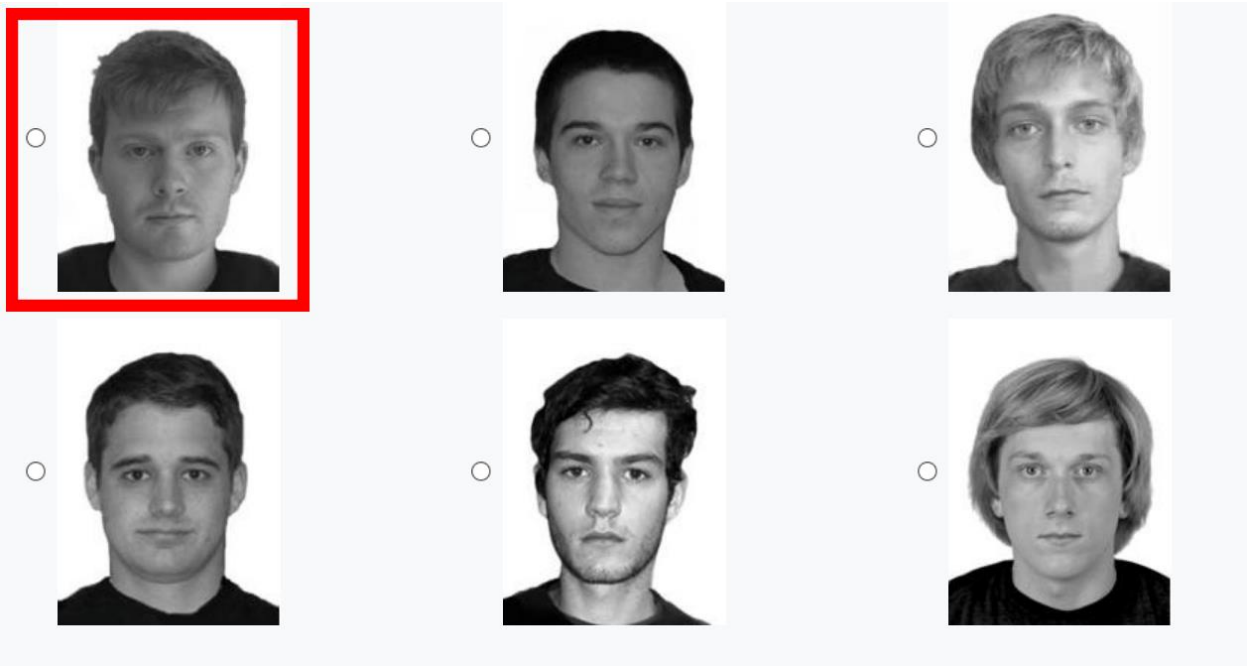


Figure A1. Carjacking culprit-present lineup. Culprit indicated in red. Lineup member numbers corresponding with lineup member choosing data are: 1, 3, 5, 2, 4, 6 (going by row, left to right).

ii. Carjacking CA

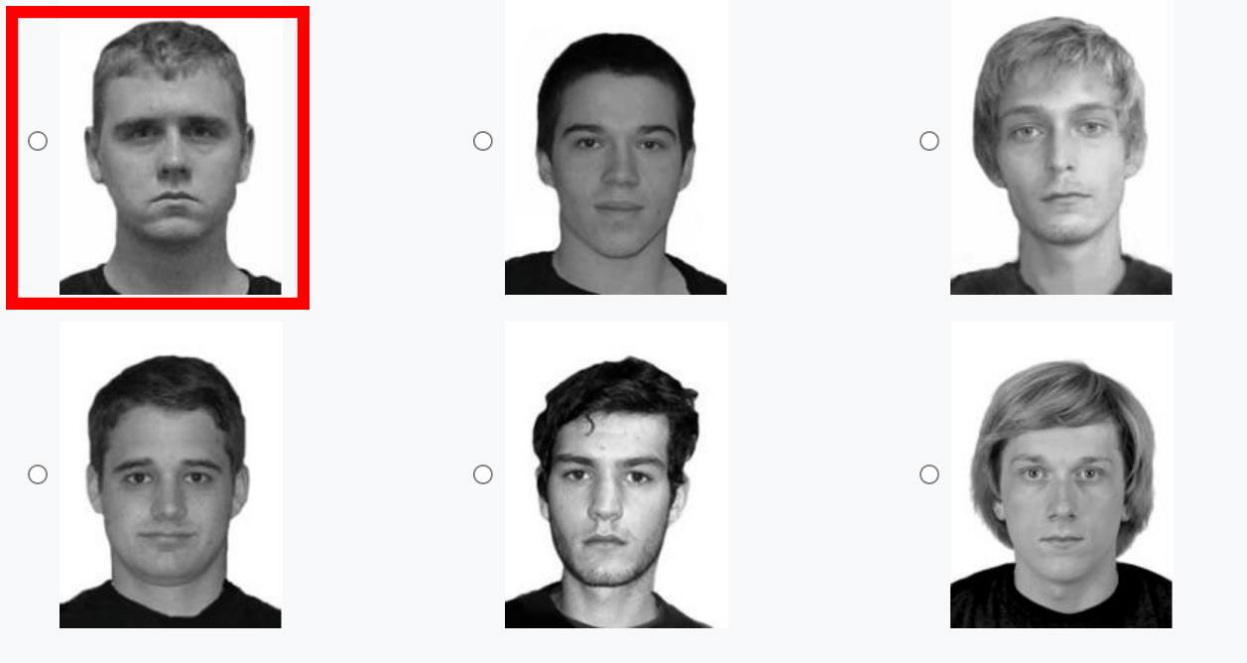


Figure A2. Carjacking culprit-absent lineup. Designated innocent suspect indicated in red. Lineup member numbers corresponding with lineup member choosing data are: 1, 3, 5, 2, 4, 6 (going by row, left to right).

iii. Theft CP

Figure A3. Theft culprit-present lineup. Culprit indicated in red. Lineup member numbers corresponding with lineup member choosing data are: 1, 3, 5, 2, 4, 6 (going by row, left to right).

iv. Theft CA

Figure A4. Theft culprit-absent lineup. Designated innocent indicated in red. Lineup member numbers corresponding with lineup member choosing data are: 1, 3, 5, 2, 4, 6 (going by row, left to right).

b. Experiment 1

i. Carjacking CA

Figure A5. Carjacking culprit-absent lineup. Designated innocent suspect indicated in red. Lineup member numbers corresponding with lineup member choosing data are: 1, 3, 5, 2, 4, 6 (going by row, left to right).

ii. Theft CA

Figure A6. Theft culprit-absent lineup. Designated innocent suspect indicated in red. Lineup member numbers corresponding with lineup member choosing data are: 1, 3, 5, 2, 4, 6 (going by row, left to right).

iii. Mugging CA

1. Culprit

Figure A7. Mugging culprit

2. CA lineup

Figure A8. Mugging culprit-absent lineup. Designated innocent suspect indicated in red. Lineup member numbers corresponding with lineup member choosing data are: 1, 3, 5, 2, 4, 6 (going by row, left to right).

iv. Graffiti-ing CA

1. Culprit

Figure A9. Graffiti-ing culprit

2. CA lineup

Figure A10. Graffiti-ing culprit-absent lineup. Designated innocent suspect indicated in red. Lineup member numbers corresponding with lineup member choosing data are: 1, 3, 5, 2, 4, 6 (going by row, left to right).

B. Exclusionary criteria

We excluded any lineups where participants:

1) Fail to answer at least one video question correctly for the corresponding crime video
    a) The rationale being that if participants didn't catch basic details about the video, they were unlikely to have paid a minimal level of attention to the video. These kinds of questions have been used in other eyewitness studies (e.g., Colloff et al., 2016). Additionally, current US Supreme Court standards for evaluating the reliability of eyewitness testimony include assessing "the witness' degree of attention" (National Research Council, 2014, p. 32).
2) Reported that they did not adhere at all to their assigned strategy
    a) Previous research suggests that eyewitnesses can metacognitively identify and report on identification decision strategies (e.g., Dunning & Stern, 1994; Sauerland & Sporer, 2007). As we have no other indices of strategy use, we will rely on self-reported adherence to identify and exclude participants.
3) For those in the Absolute condition, any lineup where a lineup member is chosen and then subsequently a different lineup member is chosen as "most similar to the culprit" in the Forced Relative phase
    a) We choose to exclude these participants because it is hard to interpret a decision in which 1) a participant initially believes a lineup member is the culprit (and thus should be viewed as "most similar to the culprit") but 2) later chooses a different lineup member as "most similar to the culprit". This criterion captures the current Supreme Court stipulation that judges should examine "the accuracy of the witness' prior description of the criminal"; i.e., the consistency of the witness'

pretrial identifications (National Research Council, 2014, p. 32; Rakoff & Loftus, 2018).

And any participants who:

1) Self-report as being "Very inattentive"
    a) Based on the Supreme Court standards, such a self-report directly calls into question “the witness’ degree of attention” (National Research Council, 2014, p. 32).
2) Self-report as having completed a similar study before
    a) Though we inform participants that the study will involve videos and memory questions, we do not tell participants that they will be watching crime videos and completing lineups. Any participants who have completed a prior eyewitness/lineup study may approach the task differently, and do not represent the typical real-world eyewitness (who has no expectations when witnessing a crime).
3) Self-report as being color-blind or severely visually impaired (specifically, whether participants report that they had great difficulty seeing the videos or the lineup photos).
    a) The former will be excluded due to potential memory differences when study & test stimuli match in color modality; Wogalter & Laughery, 1987). The latter will be excluded on the basis of the Supreme Court standard that reliability depends on “the opportunity of the witness to view the criminal at the time of the crime” (National Research Council, 2014, p. 32), and research on estimator variables showing the negative effects of viewing distance and poor lighting on eyewitness accuracy (Granhag et al., 2014).

4) Report any major technical issues that arose during the study (e.g., videos freezing or not playing at full size)

5) ~~Leave the active survey window during either of the videos or either of the main lineups, as determined using the TaskMaster tool for Qualtrics (Permut, Fisher, & Oppenheimer, 2019).~~[23]

6) For those in the relative condition, any participants who gave duplicate rankings (e.g., two ratings of 1) on any of the lineups. This criterion was not pre-registered.

    a) It is hard to interpret duplicate ratings (e.g., a participant might believe that two lineup members are equally similar, or they may have failed to understand the instructions), but we opted to exclude any participants that met this criterion.

Experiment 2 fail-safe stopping criteria:

We pre-registered that we would cease online data collection at any increment of $n = 100$:

1) If choosing is not at least 10% lower (raw mean difference) on culprit-absent lineups relative to culprit-present lineups

2) If accuracy on either culprit-present or culprit-absent lineups is lower than 30% or higher than 80%

3) If the culprit/suspect are not judged to be "most similar" to the culprit at least 70% of the time, or if any other lineup member is judged to be "most similar" to the culprit at least 20% of the time.

4) If we detect that more than 10% of our participants are bots (via a bot-filter question)

5) If more than 10% of our participants complete the survey too quickly (5 minutes or less)

[23] We dropped this exclusion criterion because we were unable to get the TaskMaster tool to work with our version of the survey.

6) If the error rate on the video attention questions is greater than 10%

7) If total exclusions based on our exclusionary criteria result in more than 10% of the sample being excluded

At several points during data collection, there were minor violations of the above (e.g., at N = 100, total lineup exclusions were 11%; at N = 200, one filler was chosen as "most similar" 23% of the time). However, we made the executive decision to continue data collection in light of these minor violations (we also note that at N = 500, the only violated criterion was the total exclusions, with 10% of lineups and 3% of participants excluded).

C. Bayesian priors

a. Lineup similarity priors

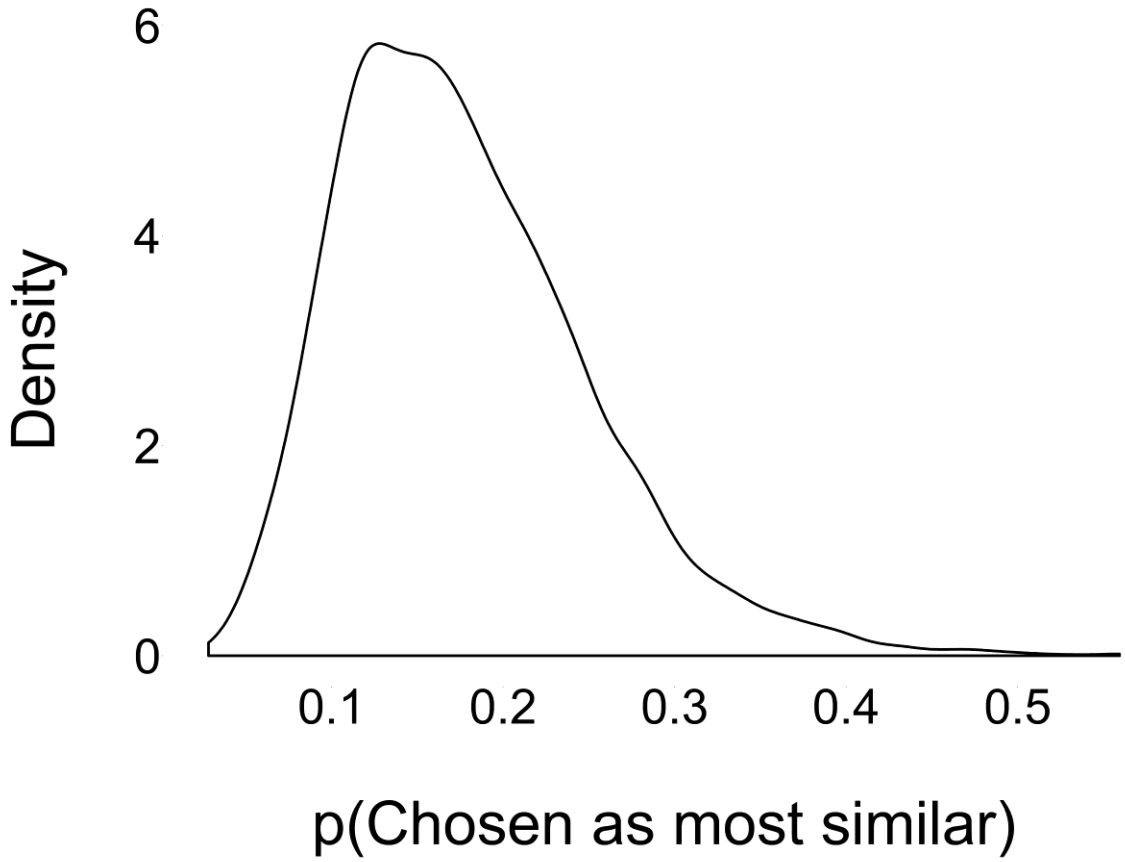


Figure C1. Prior distribution of the probability that the innocent suspect/culprit is chosen as most similar to the culprit. Density plot of 4,000 draws from the prior distribution

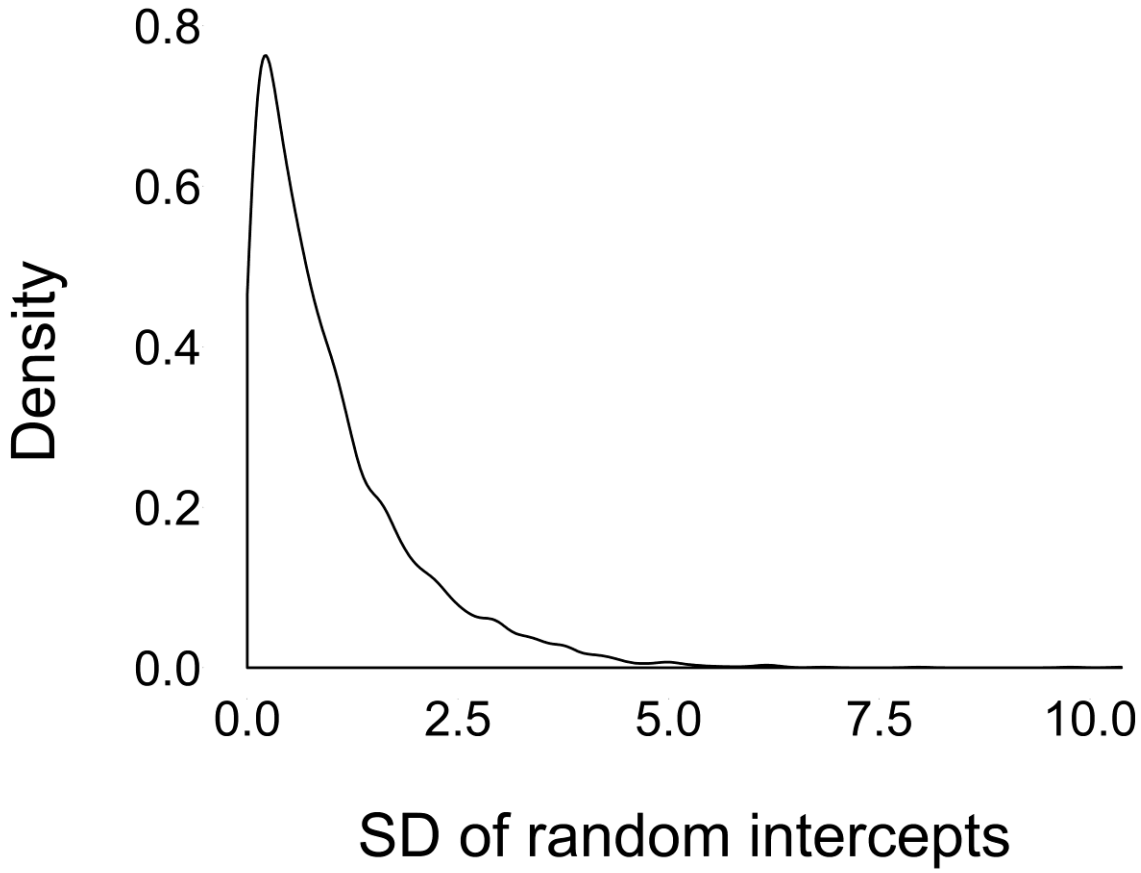


Figure C2. Prior distribution of the standard deviation of participant/lineup random intercepts. Exponential distribution places more plausibility on lower values but allows for substantial random effects variation. Density plot of 4,000 draws from the prior distribution

b. p(ID) priors

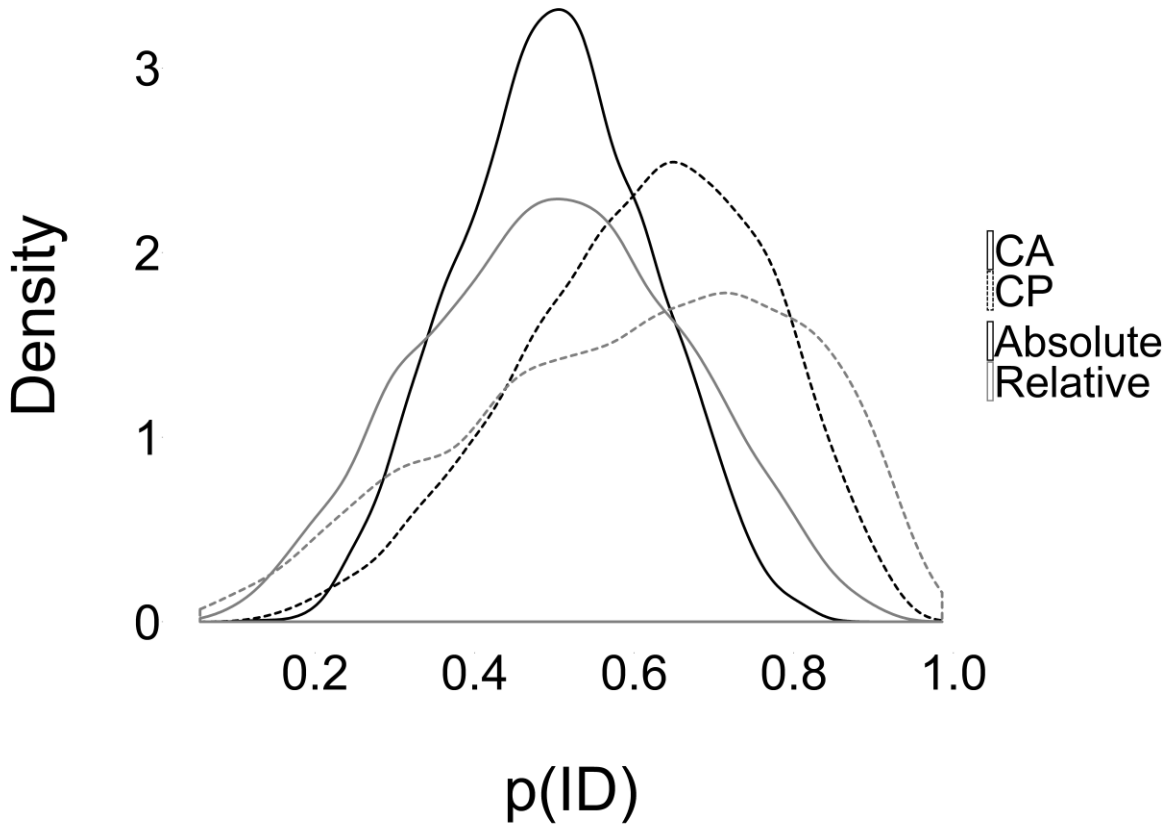


Figure C3. Prior distribution of the probability of making an identification, by Strategy and Presence*. Density plot of 4,000 draws from the prior distribution

*For Experiment 1, the priors were the same, but only included the CA distribution. Similarly, for reduced-effects models (e.g., Presence only), the priors were the same, just excluding variables not included in the model. For the fixed-effects lineup analysis, the priors were the same (but individual lineup prior distributions identical to the overall prior distributions).

c. p(Filler ID) priors

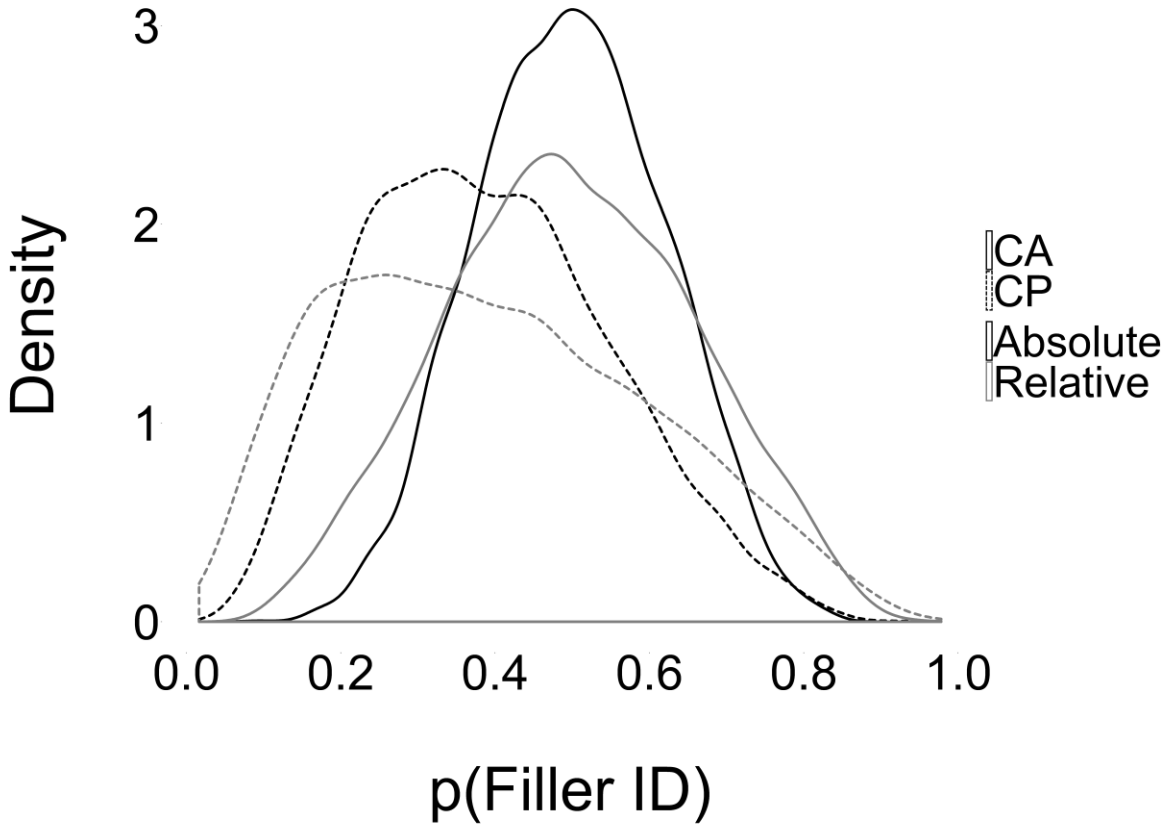


Figure C4. Prior distribution of the probability of making a filler identification, by Strategy and Presence*. Density plot of 4,000 draws from the prior distribution

*For Experiment 1, the priors were the same, but only included the CA distribution. Similarly, for reduced-effects models (e.g., Presence only), the priors were the same, just excluding variables not included in the model. For the fixed-effects lineup analysis, the priors were the same (but individual lineup prior distributions identical to the overall prior distributions).

d. Confidence & Difficulty priors

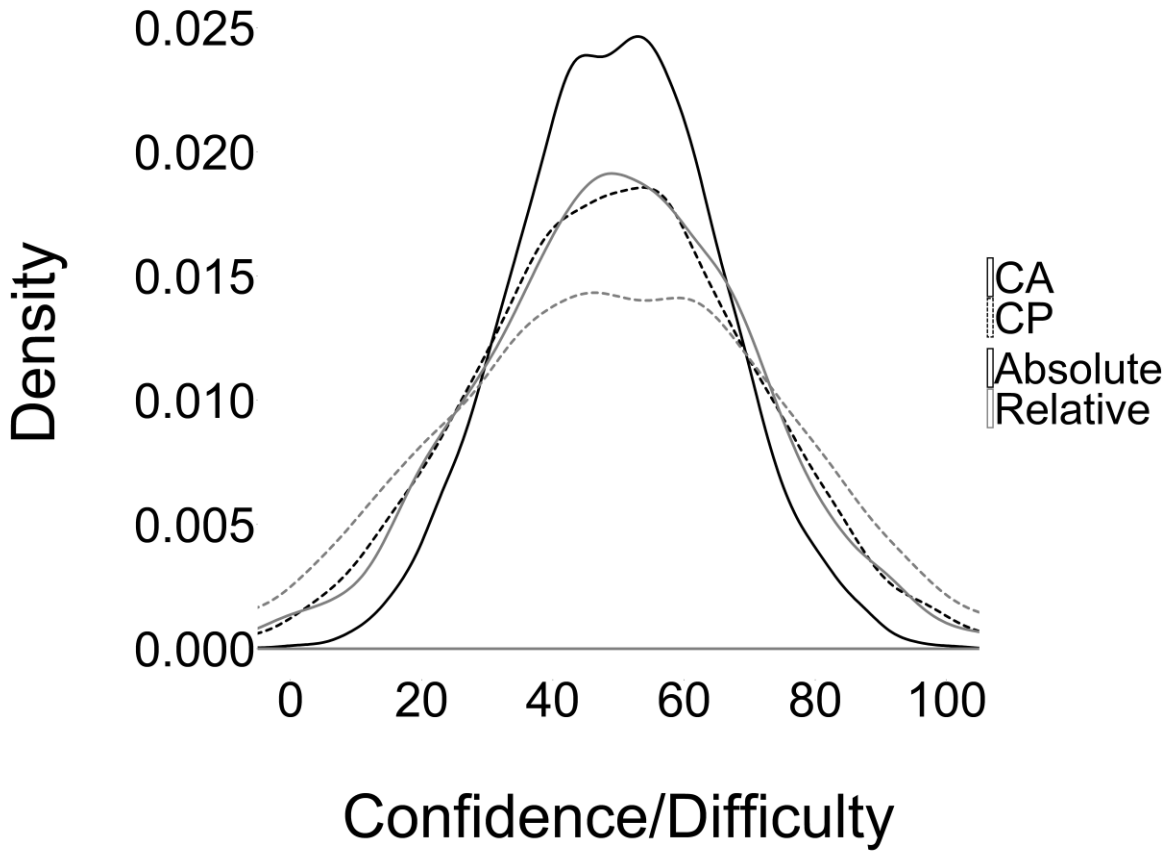


Figure C5. Prior distribution of confidence/difficulty ratings, by Strategy and Presence. Density plot of 4,000 draws from the prior distribution

e. Order effect priors

i. Experiment 1

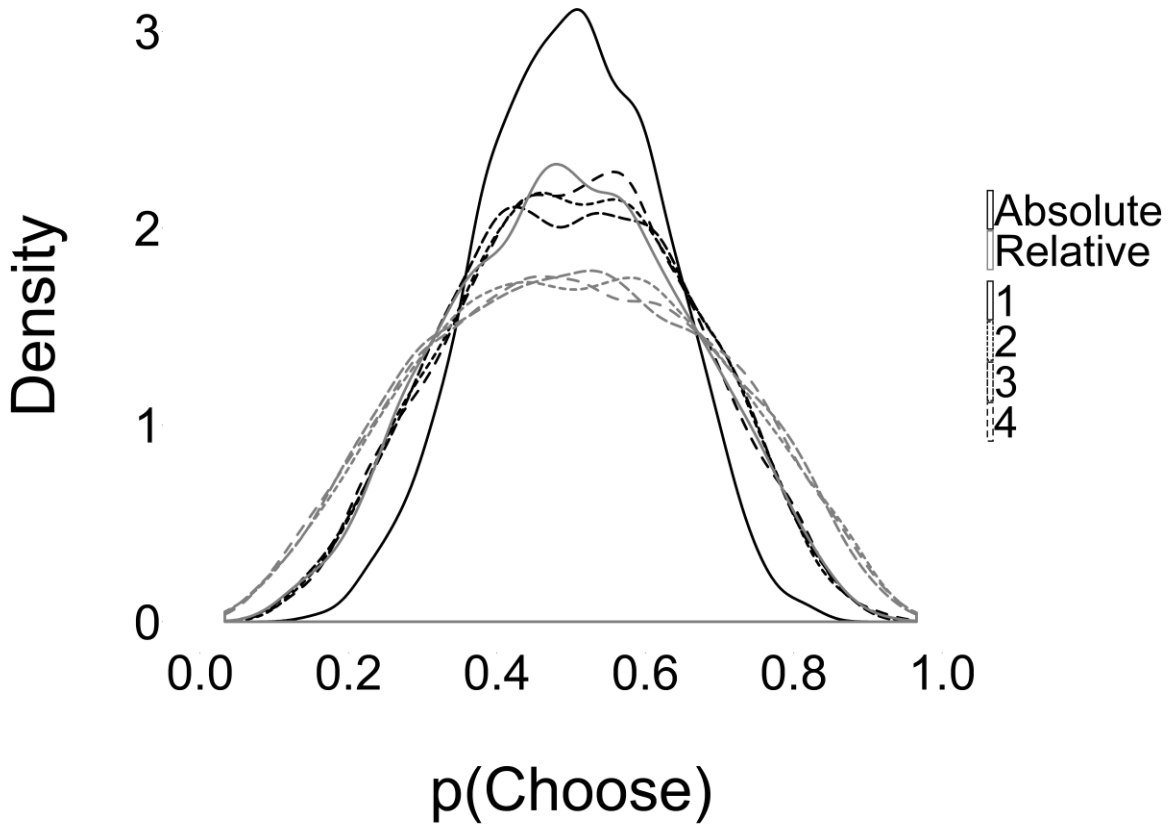


Figure C6. Experiment 1: Prior distribution of the probability of choosing, by Strategy and Lineup. Density plot of 4,000 draws from the prior distribution

ii. Experiment 2

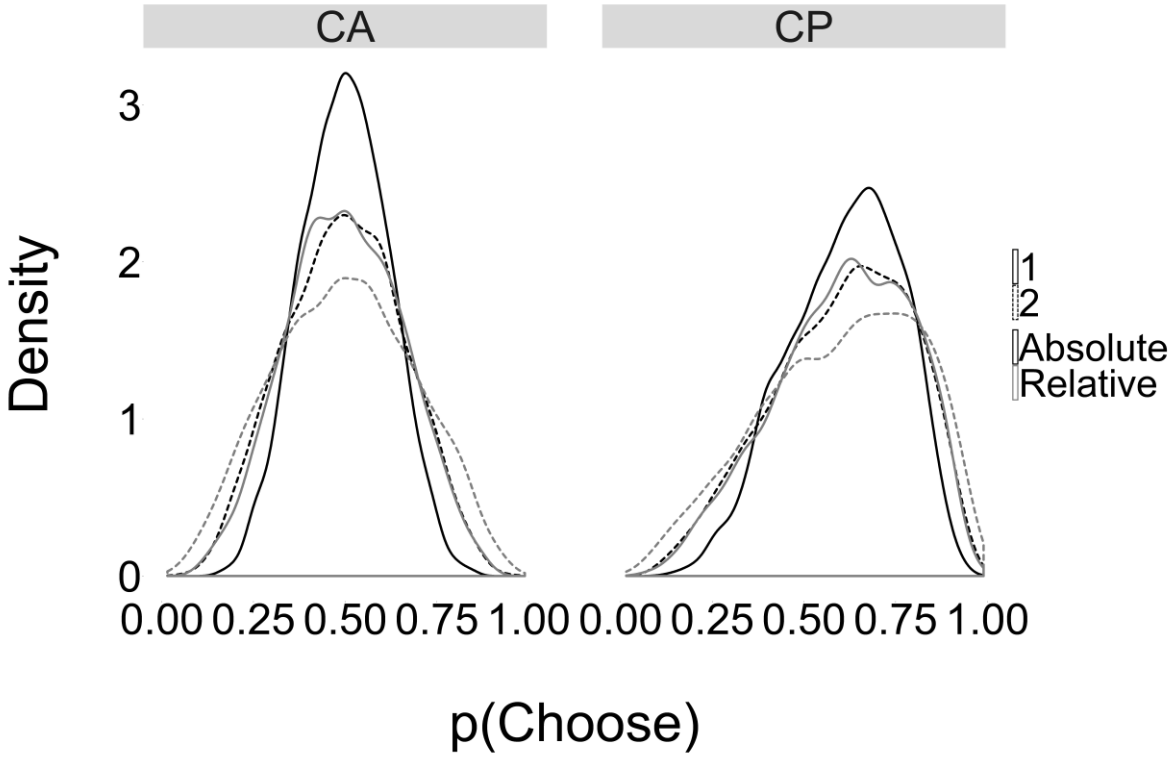


Figure C7. Experiment 2: Prior distribution of the probability of choosing, by Strategy, Presence and Lineup. Density plot of 4,000 draws from the prior distribution

D. Supplementary hypotheses and analyses

a. **H5 (Supplementary): We hypothesize no differences in decision confidence/difficulty across strategy conditions.** Though there are a priori reasons to expect differences in overall lineup accuracy as a function of strategy, we do not have a prior reason to believe that self-reported decision confidence and difficulty will differ by strategy.

Confidence. We first tested for Strategy and Presence effects[24] on lineup decision confidence in our combined dataset using the same model comparison, estimation and BF

[24] We did not include presence effects in our pre-registered models, but have included them here to provide a more complete picture of confidence ratings. The same was true for difficulty.

approaches we used for IDs. The model comparison results for confidence are shown below in Figure 29:

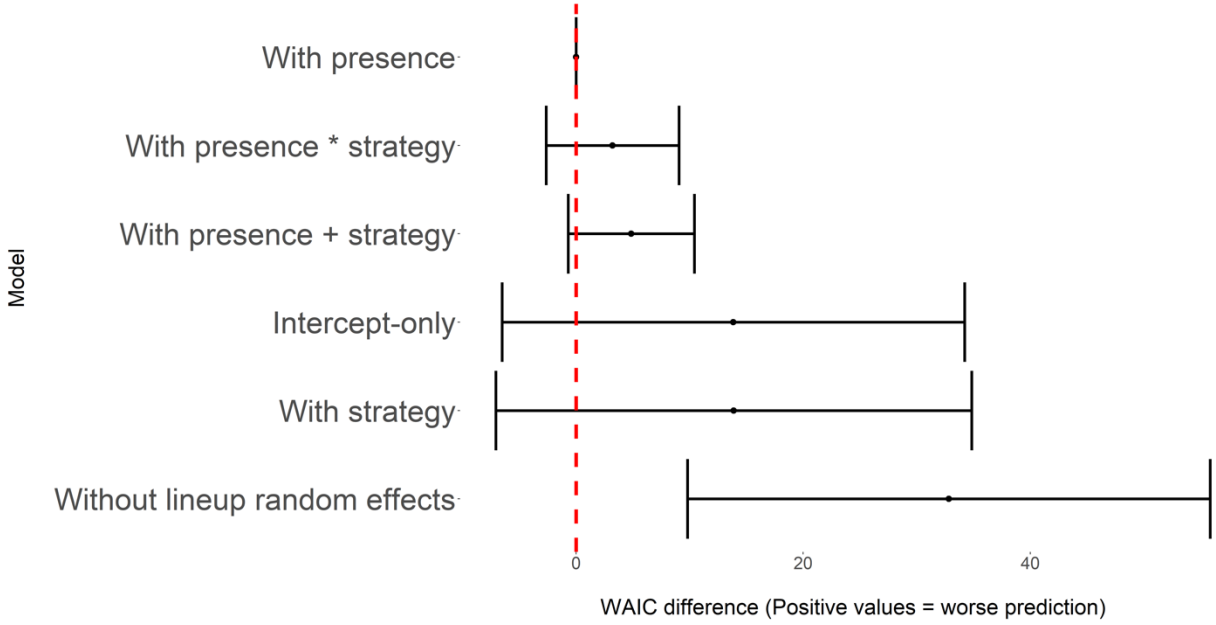


Figure D1. Experiments 1 & 2: WAIC model comparison results for predicting confidence. Models presented in descending order of predictive ability relative to the best model (1st Row). Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

The key result here is that although the best model suggests an effect of presence, removing *all* substantive effects from the model (i.e., the intercept-only) model did not result in a clear decrease in predictive accuracy, suggesting trivial effects of strategy and presence on decision confidence (but comparatively substantial by-lineup variability). The results for our estimation analysis are shown below in Figure 30:

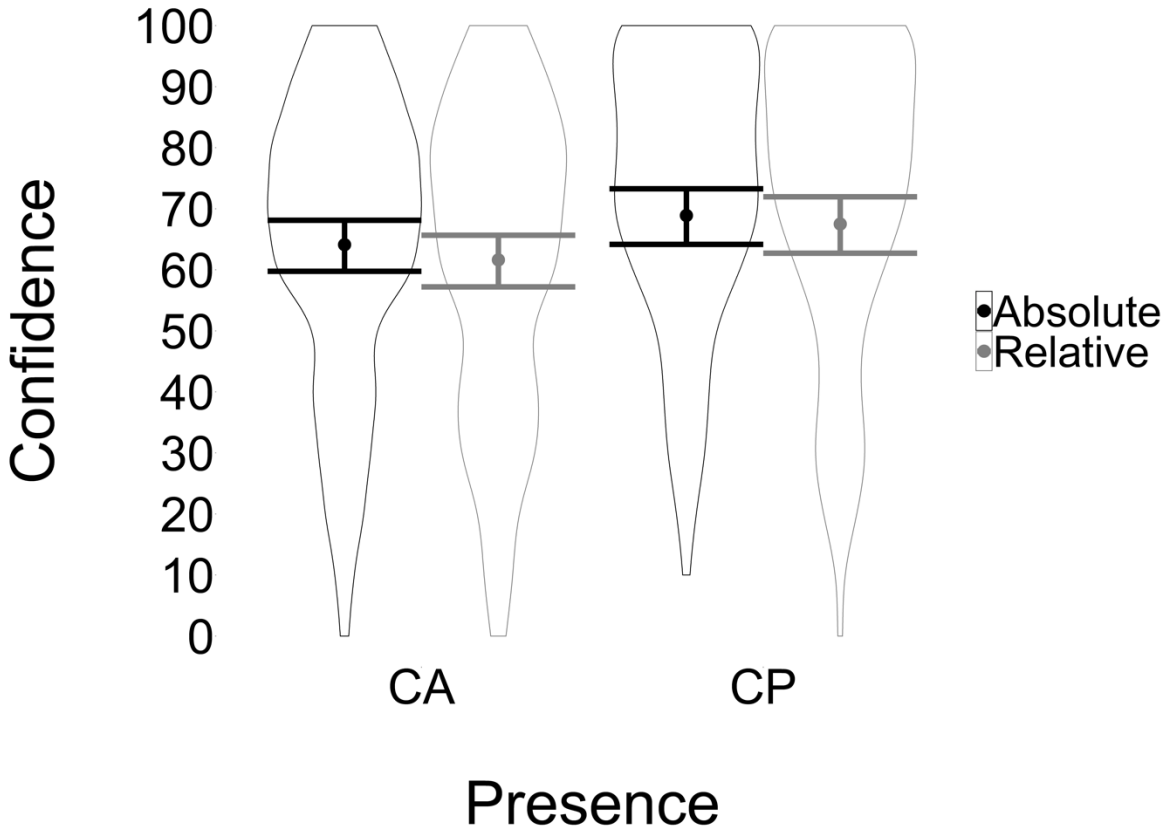


Figure D2. Experiments 1 & 2: Confidence by Strategy and Presence. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models). Violin plots represent the distribution of confidence ratings.

Contrary to our hypothesis, these results are suggestive of a slight confidence reduction with relative judgments, both for CA lineups (estimated difference = 2.48, 95% CR[-.54, 5.47]) and for CP lineups (estimated difference = 1.38, 95% CR[-2.33, 5.14]). However, these differences are small (< 5% on a scale that can range from 0% - 100%), and both

the model comparisons and BFs provided some evidence against any differences (BF against a CA difference = 2.85, BF against a CP difference = 6.82).

Difficulty. We conducted similar analyses for lineup decision difficulty. The model comparison results are shown below in Figure 31:

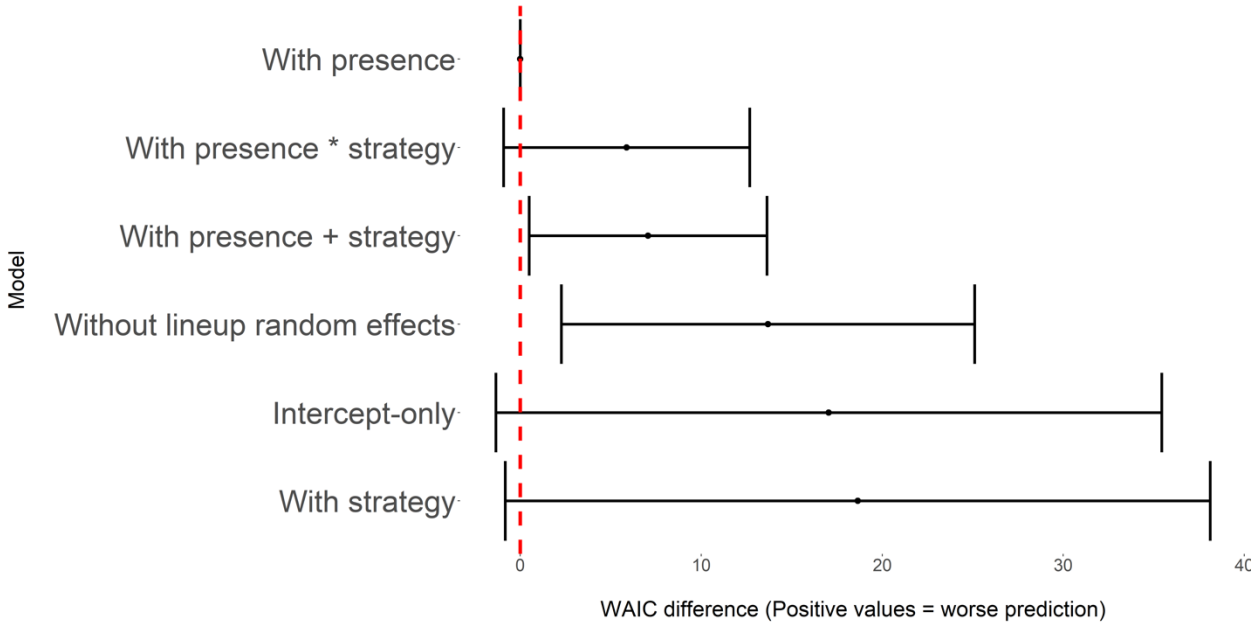


Figure D3. Experiments 1 & 2: WAIC model comparison results for predicting difficulty. Models presented in descending order of predictive ability relative to the best model (1st Row). Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

Here, the comparisons suggest a greater decrease in predictive accuracy when removing all substantive effects (but not necessarily when removing lineup random effects). Although the model comparisons point to a potential Presence effect, they are less supportive of a Strategy effect. The results of the estimation analysis are presented below in Figure 32:

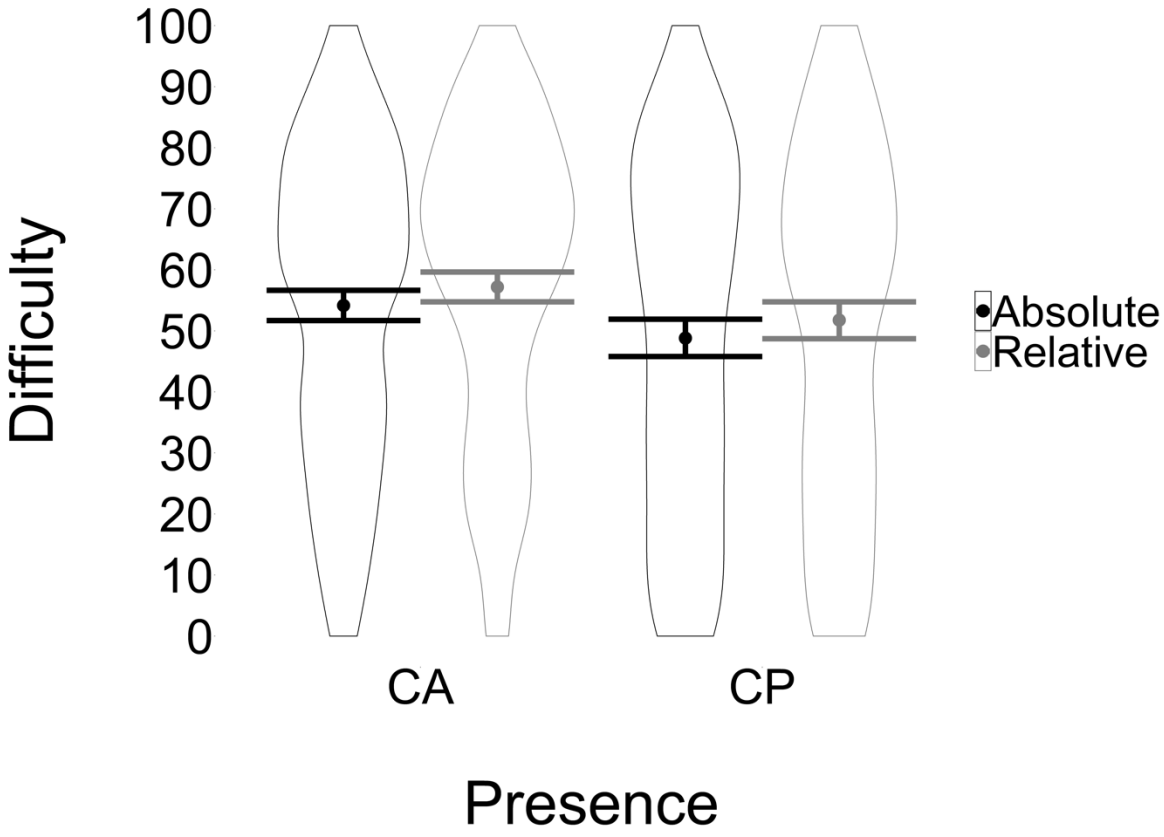


Figure D4. Experiments 1 & 2: Difficulty by Strategy and Presence. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models). Violin plots represent the distribution of confidence ratings.

Based on these results, participants may have viewed their decisions on relative rank-order lineups as slightly more difficult, both for CA lineups (estimated difference = 3.03, 95% CR[-.39, 6.57]) and for CP lineups (estimated difference = 2.95, 95% CR[-1.26,

7.18])[25]. Again, these differences were small, and our model comparisons and BFs provided some evidence against meaningful differences (BF against a CA difference = 2.14, BF against a CP difference = 3.43). It is worth noting that difficulty ratings showed more variability than confidence ratings. Perhaps participants had greater difficulty in translating subjective meta-memorial judgments of difficulty into concrete ratings. Our analyses of correct, false, and filler IDs suggested minimal effects of strategy, and our confidence and difficulty analyses continue this trend. If our manipulation did affect participants' meta-memorial judgments, those effects were small.

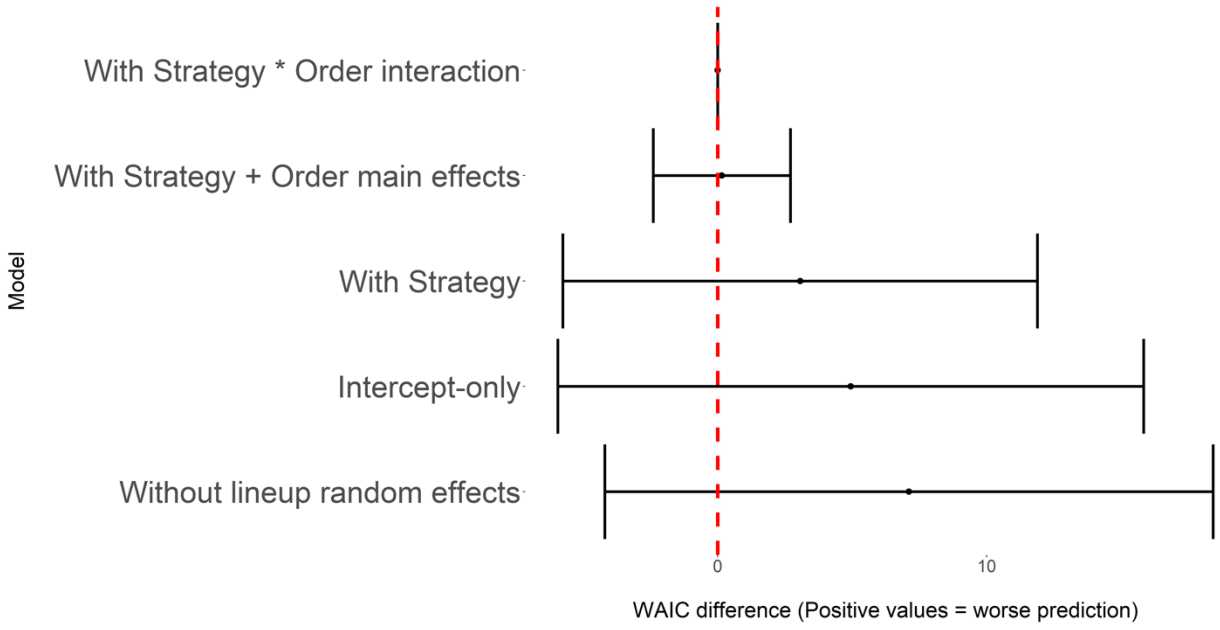
b. **H6 (Supplementary): We hypothesize no difference in choosing rates as a function of lineup number. That is, participant will not tend to choose/reject more on earlier vs. later lineups.** In all of our experiments, we presented participants with multiple lineups. Hypothesis 6 addresses a potential concern with multi-lineup designs: in the real world, it is rare that a single eyewitness will view multiple lineups for different crimes. Though prior research suggests that multi-lineup designs are as valid as single-lineup designs (Mansour, Beaudry, & Lindsay, 2017), there are remaining concerns about potential practice effects, awareness of task demands, etc. (2017). Thus, we expect no effects of lineup order but will test for them.

Although previous research suggests that multi-lineup designs are valid and that choosing does not differ substantially across lineups, only alternating Crime-Lineup-Crime-Lineup-etc. designs have been directly tested (Mansour et al., 2017). As a result, we

[25] Examining the results of Experiments 1 & 2 individually, it appears that the CA difficulty difference in the combined data is largely driven by Experiment 1. Perhaps in this experiment, participants viewed relative rank-order decisions as more difficult because they had to complete the rank-order process four times.

tested for potential order effects in both our experiments (which used Crime-Crime-Lineup-Lineup-etc. designs). Like in the Mansour and colleagues study (2017), we operationalized choosing as *any* choosing (suspect, culprit, or filler, collapsing across presence). Analyzing choosing this way affords a picture of whether general response criterion is likely to change across repeated lineups. We analyzed order effects separately by experiment because the design differences are more pronounced for lineup order (i.e., Experiment 1 had four CA lineups, Experiment 2 had only two lineups, one CA and one CP). To evaluate the contributions of order effects above and beyond our previously examined substantive effects, we added order main effects and interactions to previous full models (including Strategy * Presence effects) and compared them[26]. Otherwise, our analytic strategy followed that of prior hypotheses.

Experiment 1. Model comparison results for Experiment 1 are shown below in Figure D5:



[26] Our pre-registered analyses collapsed across Strategy and Presence. Here we include these effects in our models to provide a more complete picture of order effects.

Figure D5. Experiment 1: WAIC model comparison results for order effects. Models presented in descending order of predictive ability relative to the best model (1st Row). Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

The best models contained order effects, which suggests that order may affect choosing (though dropping these and other effects did not substantially degrade prediction). The results of the estimation analyses are shown below in Figure 34:

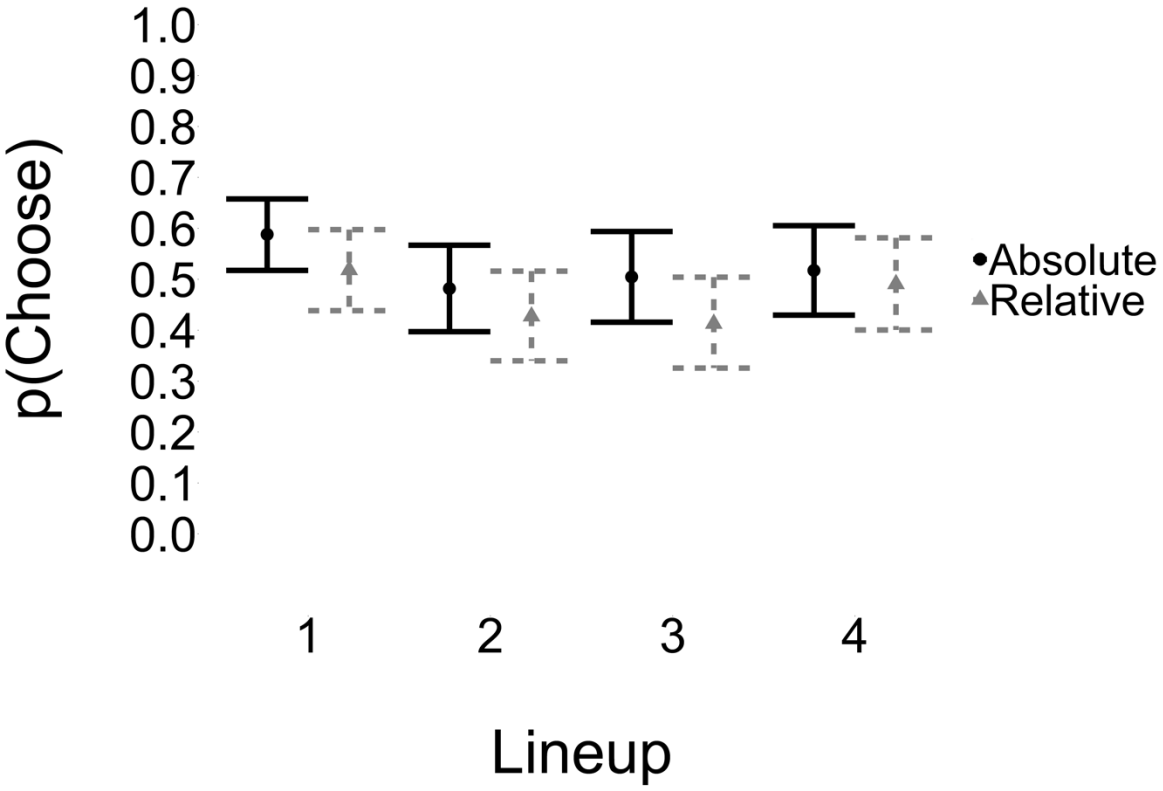


Figure D6. Experiment 1: Probability of choosing by Strategy and Lineup. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the

range in which we can be 95% sure that the true value lies (based on these data and models).

These results show that the slight decrease in false IDs that we observed in our H1 and H3 analyses is likely consistent across multiple lineups. As far as order differences go, the most pronounced difference appears to be between the 1st and 2nd lineups, after which choosing remained relatively stable. Examining this difference further, for absolute lineups choosing probability was estimated to be 11% lower on the 2nd lineup (95% CR[1%, 21%]), and for relative lineups choosing probability was estimated to be 9% lower on the 2nd lineup (95% CR[-2%, 20%]). BFs provided some evidence for the absolute lineups difference (BF in favor = 3.70), but some evidence against the relative lineups difference (BF against = 1.21). Though these differences are small, they suggest that choosing may indeed change in the early states of multi-lineup studies. Importantly though, order did not appear to moderate our small strategy differences.

Experiment 2. Model comparisons for Experiment 2 (which included only two lineups) are shown below in Figure D7:

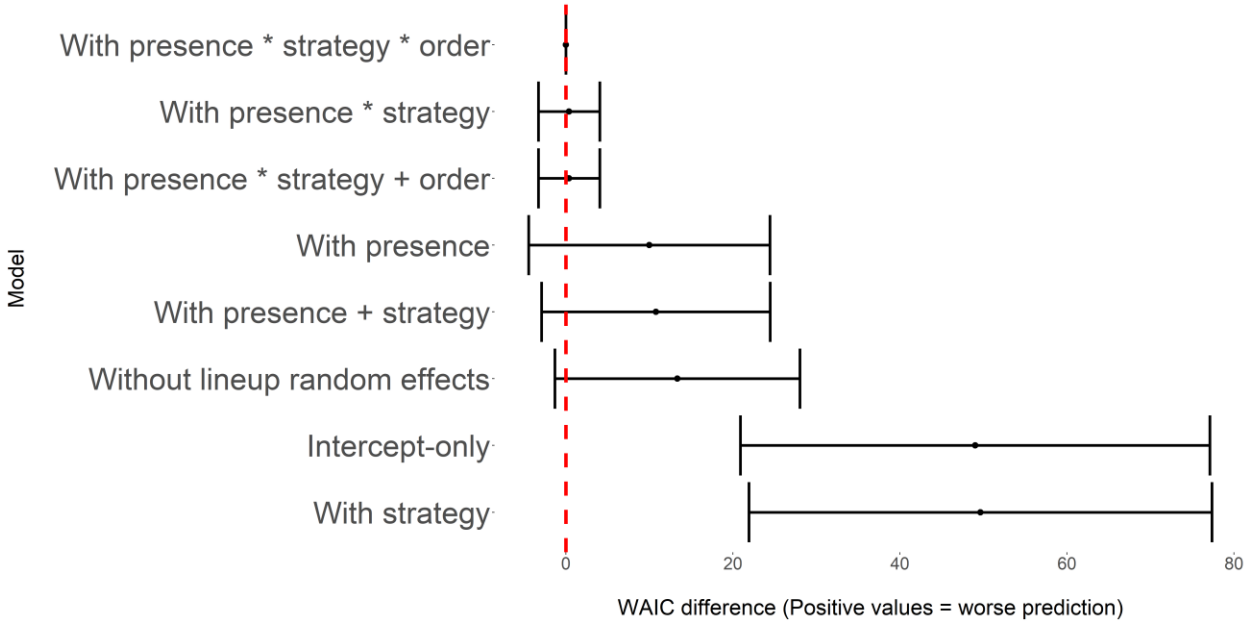


Figure D7. Experiment 2: WAIC model comparison results for order effects. Models presented in descending order of predictive ability relative to the best model (1st Row). Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

These model comparisons suggest that order may be important in predicting choosing, with presence somewhat less important, and strategy even less so. To understand these effects further, we conducted estimation analyses, shown below in Figure 36:

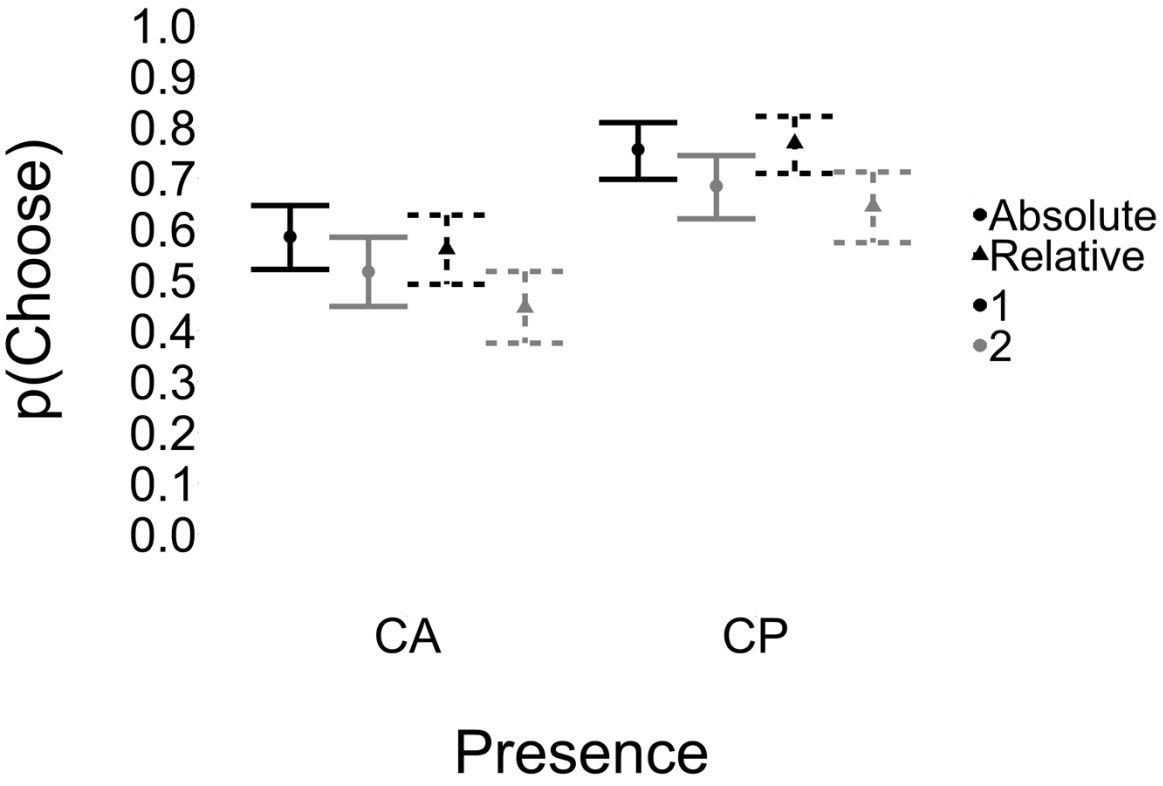


Figure D8: Experiment 1: Probability of choosing by Strategy, Presence, and Lineup (1 vs. 2). Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models).

Again we see some indication that choosing decreases after an initial lineup across strategies and culprit-presence. Due to the larger sample size of Experiment 2, these differences are more evidence than in Experiment 1. For absolute CA lineups, the estimated decrease in choosing from the 1st to 2nd lineup was 7% (95% CR[-1%, 15%]), and the BF in support of this difference was 1.47. For absolute CP lineups, the estimated

decrease was 7% (95% CR[0%, 15%]), with a supporting BF of 2.94. For relative CA lineups, the estimated decrease was 11% (95% CR[3%, 20%]), with a supporting BF of 11.11. Finally, for relative CP lineups, the estimated decrease was the largest at 12% (95% CR[5%, 20%]), with a supporting BF of 50. Again, our small strategy differences did not appear to be moderated by order.

In both experiments, we found evidence for a small decrease in choosing after the 1st lineup. These differences were small, but consistent for both absolute and relative rank-order lineups. However, the differences were similar in magnitude to (if not larger than) our previously observed strategy differences, providing additional evidence for minimal strategy effects relative to other lineup factors (e.g., culprit presence, introduction of multiple lineups).

a. Exploratory: Experiments 1 & 2: Possible gender effects

Though we did not pre-register any hypotheses regarding gender, as discussed in our limitations section, it is possible that lineup judgments were affected by ingroup/outgroup status in relation to the culprits (who were all male). Thus, we conducted a version of our main analysis (p(ID) ~ Strategy * Presence) with self-identified gender as an additional interacting factor. The results of the estimation analysis are shown below in Figure D9.

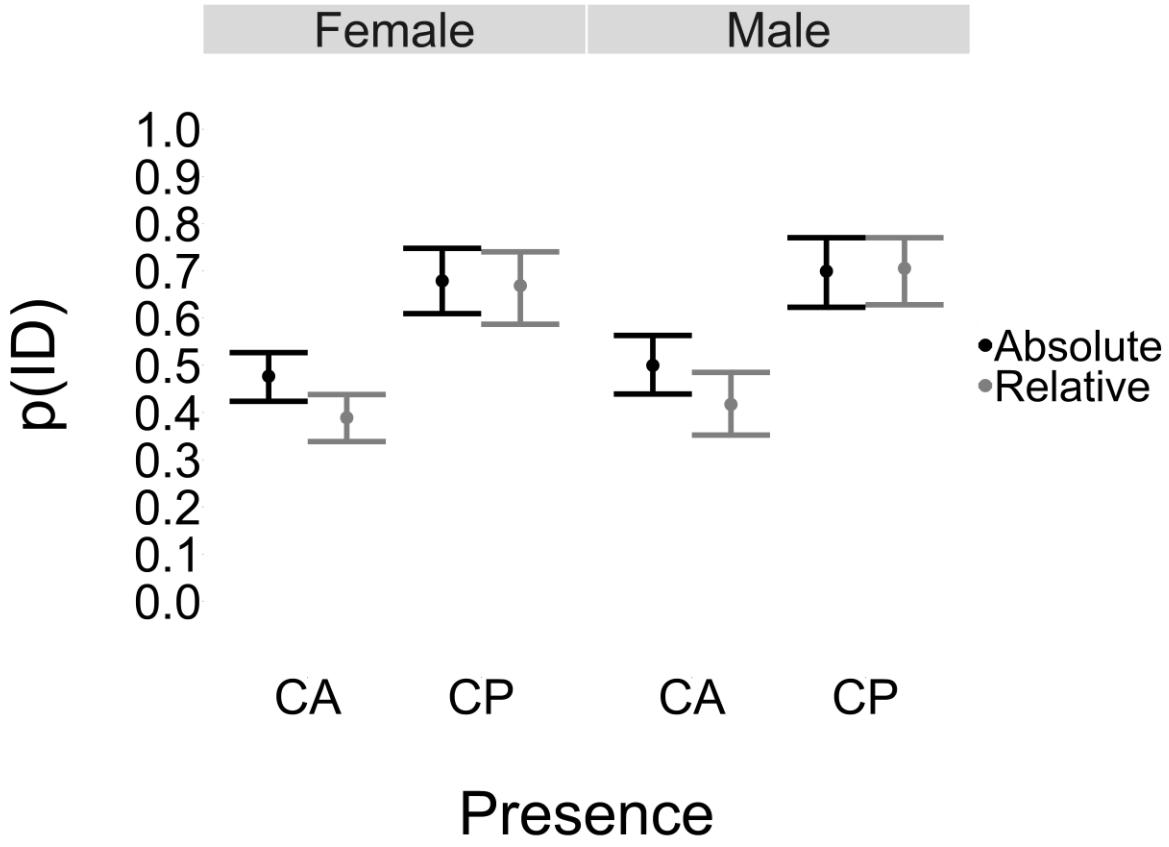


Figure D9. Experiments 1 & 2: Probability of making a culprit or suspect identification by Strategy, Presence, and self-identified Gender. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models).

As we can see, the patterns were quite similar for males and females[27]. Bayes factors revealed some evidence against a gender difference for Absolute CA lineups ($BF_{01} = 11.36$), and evidence against and moderation of this (lack of) difference by Strategy or Presence ($1.54 < BF_{01} < 4.28$).

[27] Because only 6 participants identified as something other than male or female, we did not include them as a separate category for analysis. However, it would be interesting to collect data in a larger nonbinary sample to examine response patterns, as these individuals are likely to perceive male *and* female culprits as outgroup members.

These analyses, though exploratory, suggest that gender did not contribute to substantial ingroup/outgroup biases in judgment.

E. Experiment 1-specific results

a. Lineup member similarity

i. Carjacking CA lineup

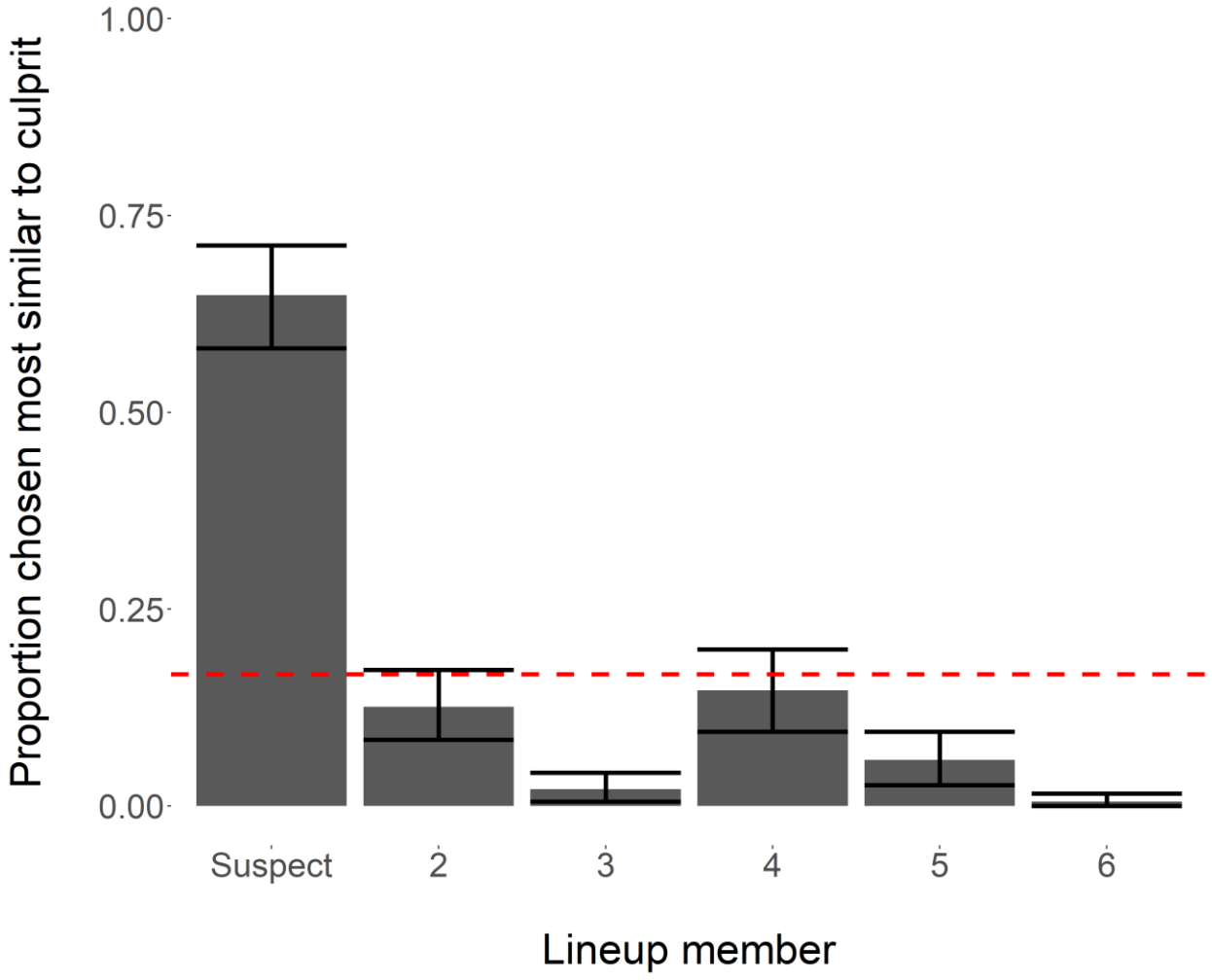


Figure E1. Carjacking culprit-absent lineup: Lineup member choosing. Errors bars = Bootstrapped 95% CRs. Dashed red line = 16.7%.

BF > 100 in favor of suspect choosing > 16.7%.

ii. Theft CA lineup

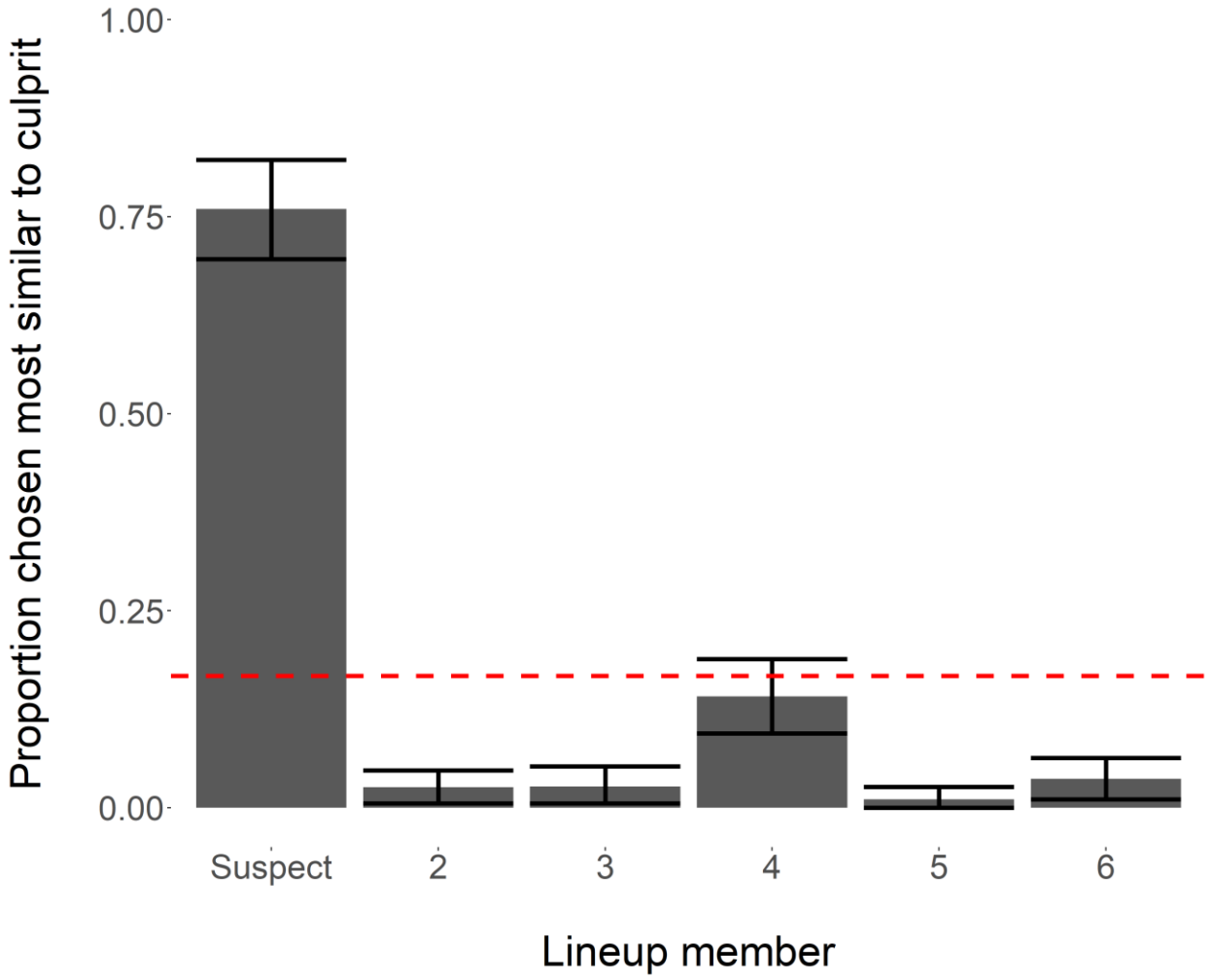


Figure E2. Theft culprit-absent lineup: Lineup member choosing. Errors bars = Bootstrapped 95% CRs. Dashed red line = 16.7%.

BF > 100 in favor of suspect choosing > 16.7%.

b. False IDs

i. WAIC model comparison results

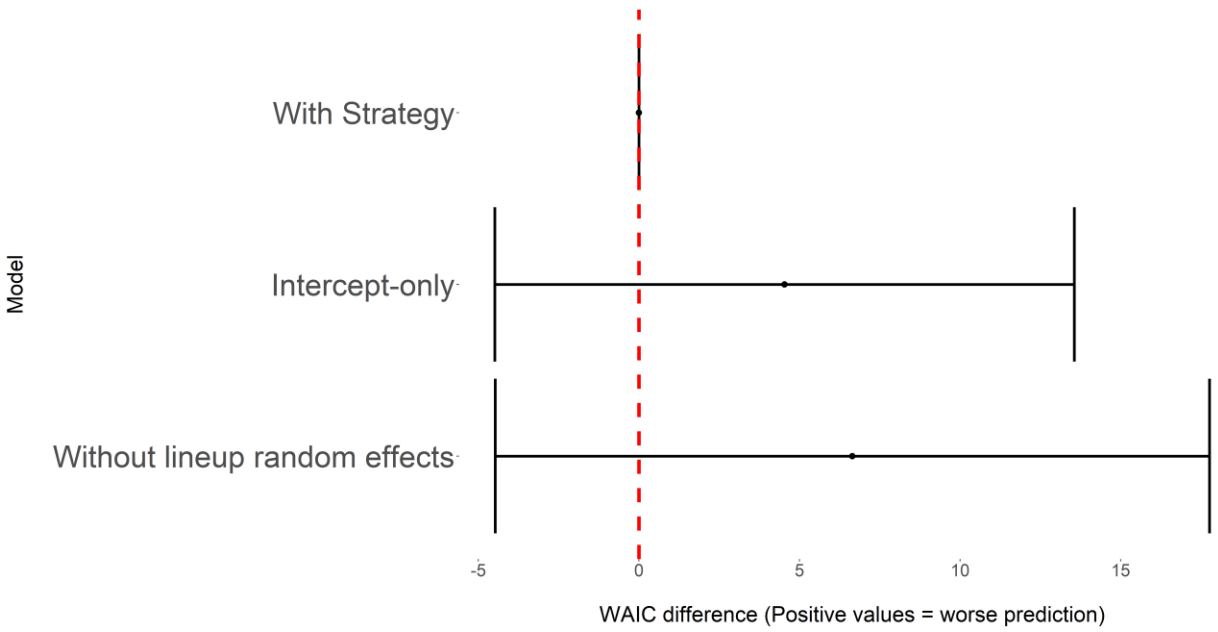


Figure E3. Experiment 1: WAIC model comparison results for predicting p(ID). Models presented in descending order of predictive ability relative to the best model (1st Row). Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

ii. Estimation results

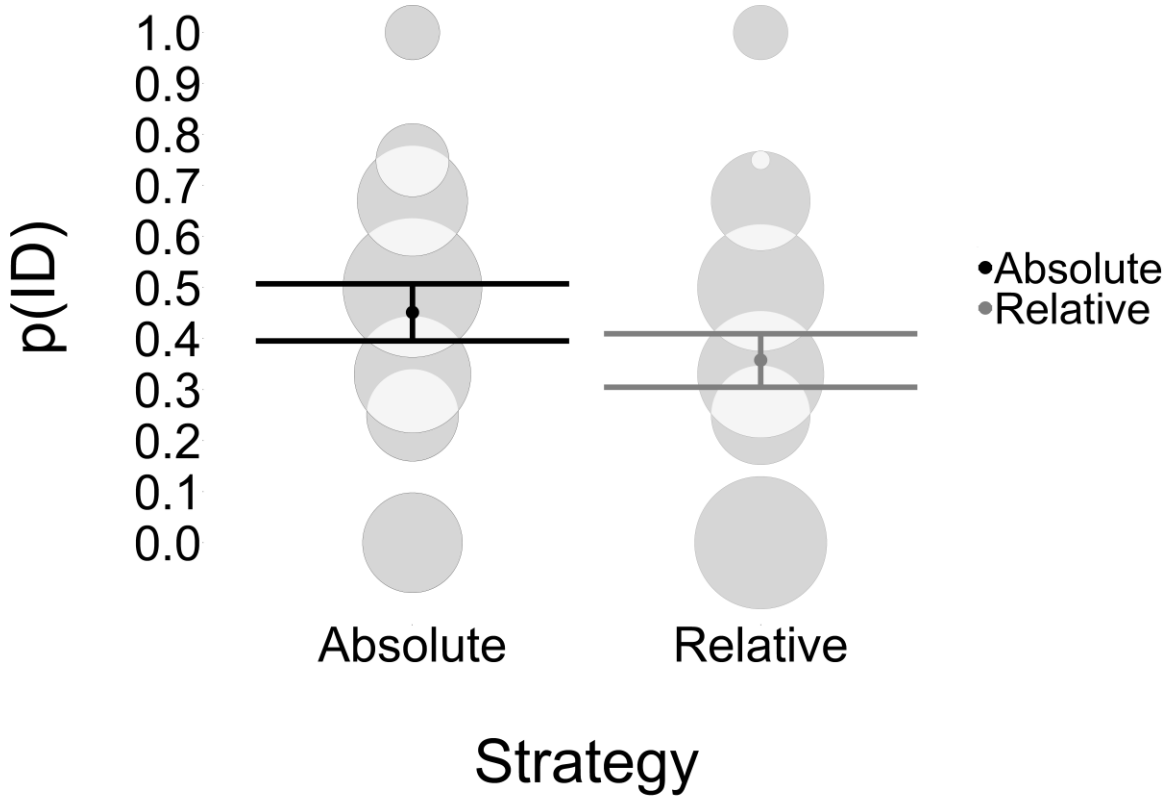


Figure E4. Experiment 1: Probability of making a false ID by Strategy. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models). Shaded circles represent the relative proportion of participants who gave responses at each probability.

Estimated absolute – relative difference = 9% [95% CR: 2%, 17%].

iii. BFs

BF = 4.76 in favor of an absolute versus relative difference in false IDs.

c. Filler IDs

i. WAIC model comparison results

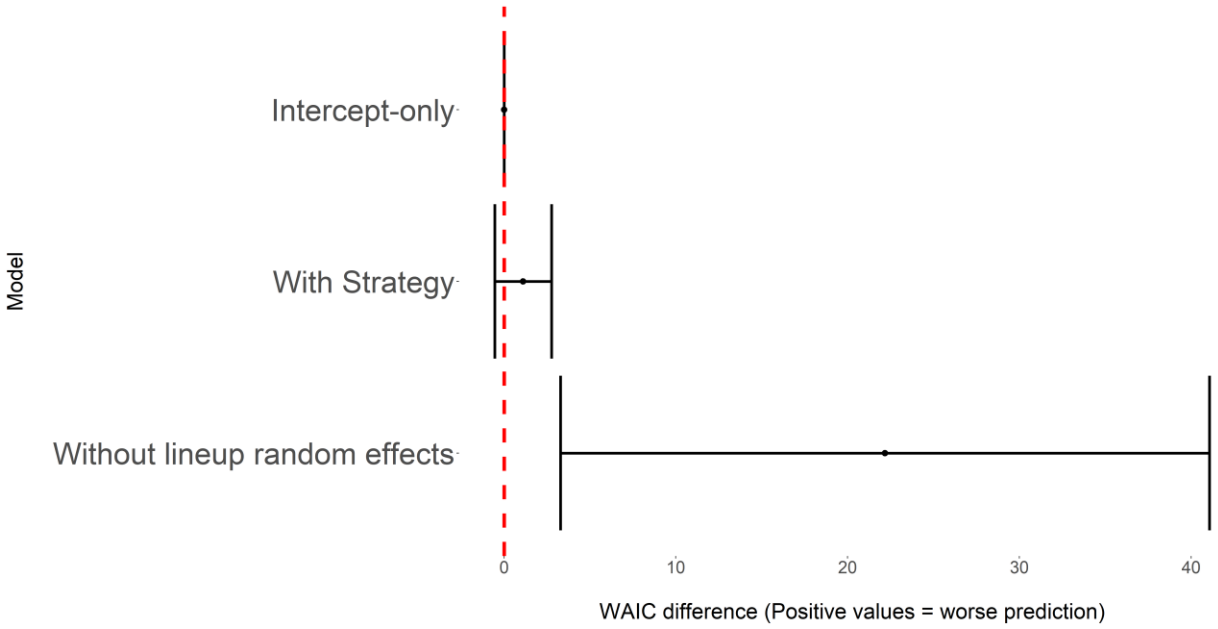


Figure E5. Experiment 1: WAIC model comparison results for predicting p(filler ID). Models presented in descending order of predictive ability relative to the best model (1st Row). Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

ii. Estimation results

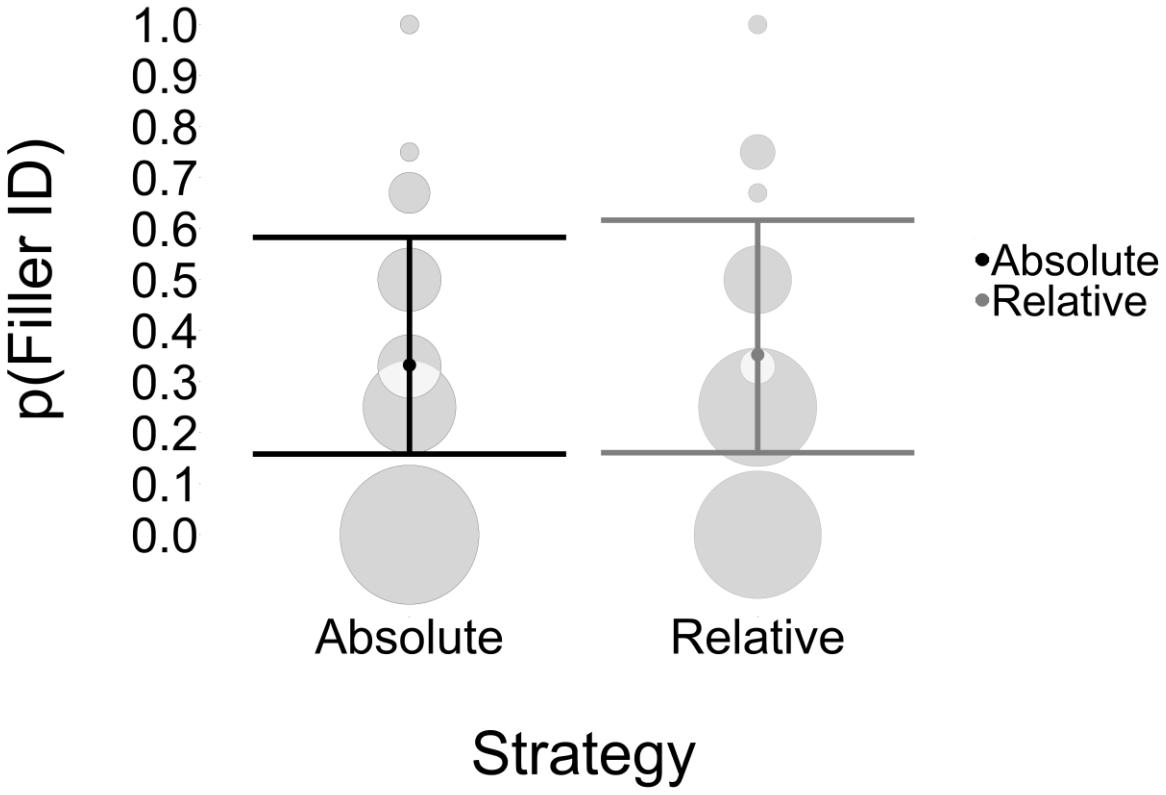


Figure E6. Experiment 1: Probability of making a filler ID by Strategy. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models). Shaded circles represent the relative proportion of participants who gave responses at each probability.

Estimated relative – absolute difference = 2% [95% CR: -6%, 11%].

iii. BFs

BF = 2.24 against an absolute versus relative difference in CA filler IDs.

d. Confidence

i. WAIC model comparison results

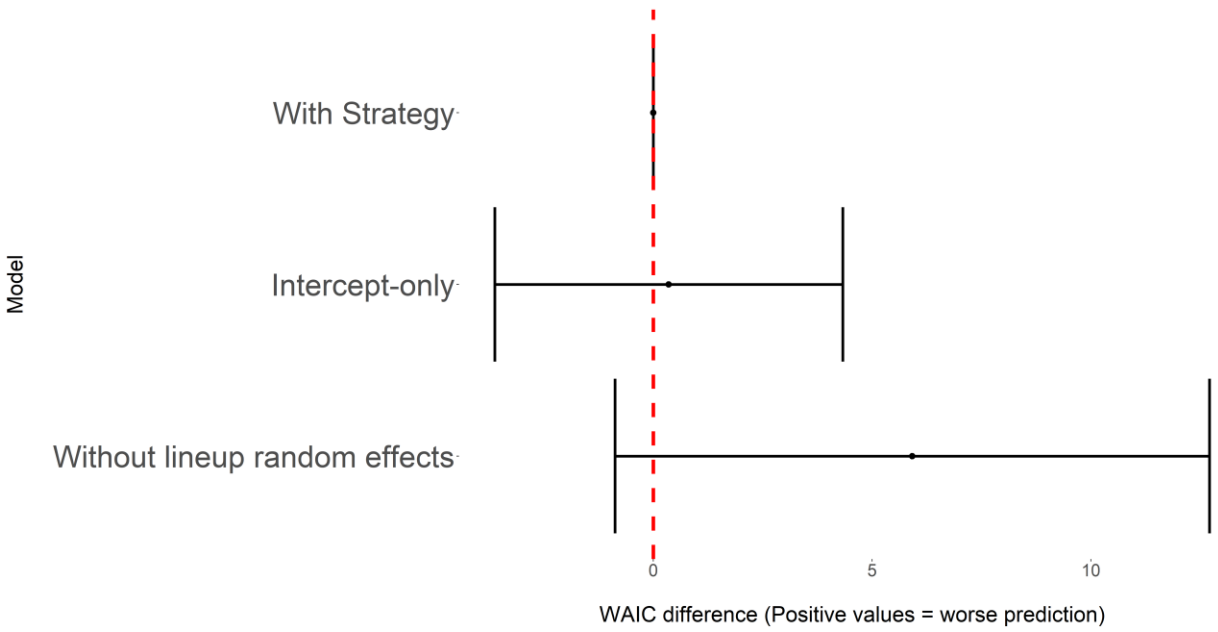


Figure E7. Experiment 1: WAIC model comparison results for predicting confidence. Models presented in descending order of predictive ability relative to the best model (1st Row). Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

ii. Estimation results

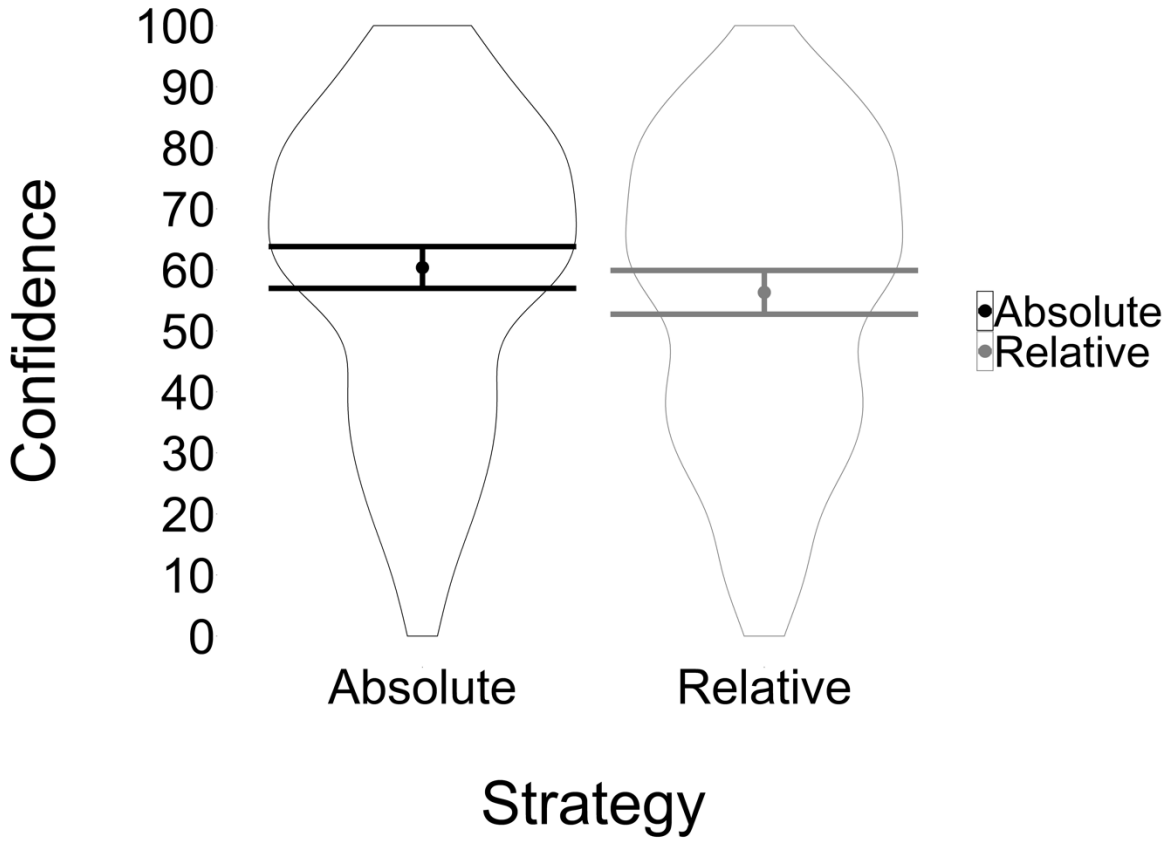


Figure E8. Experiment 1: Confidence by Strategy. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models). Violin plots represent the distribution of confidence ratings.

Estimated absolute – relative difference = 4.07 (95% CR[-.82, 8.95]).

iii. BFs

BF = 1.64 against an absolute versus relative difference in confidence.

e. Difficulty

i. WAIC model comparison results

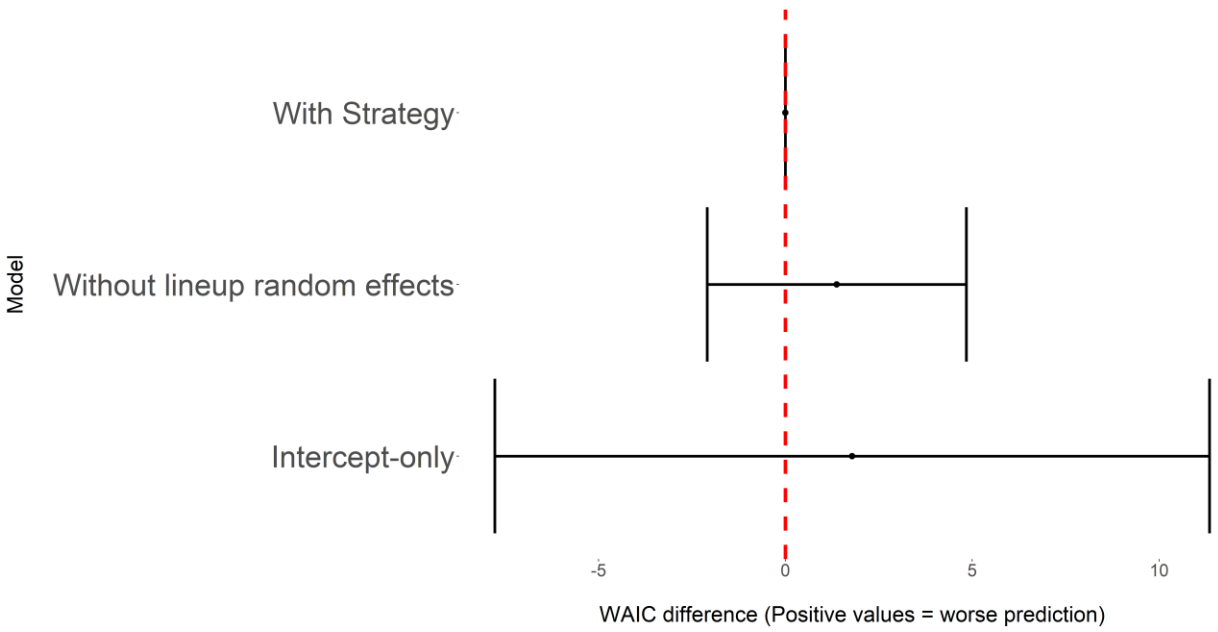


Figure E9. Experiment 1: WAIC model comparison results for predicting difficulty. Models presented in descending order of predictive ability relative to the best model (1st Row). Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

ii. Estimation results

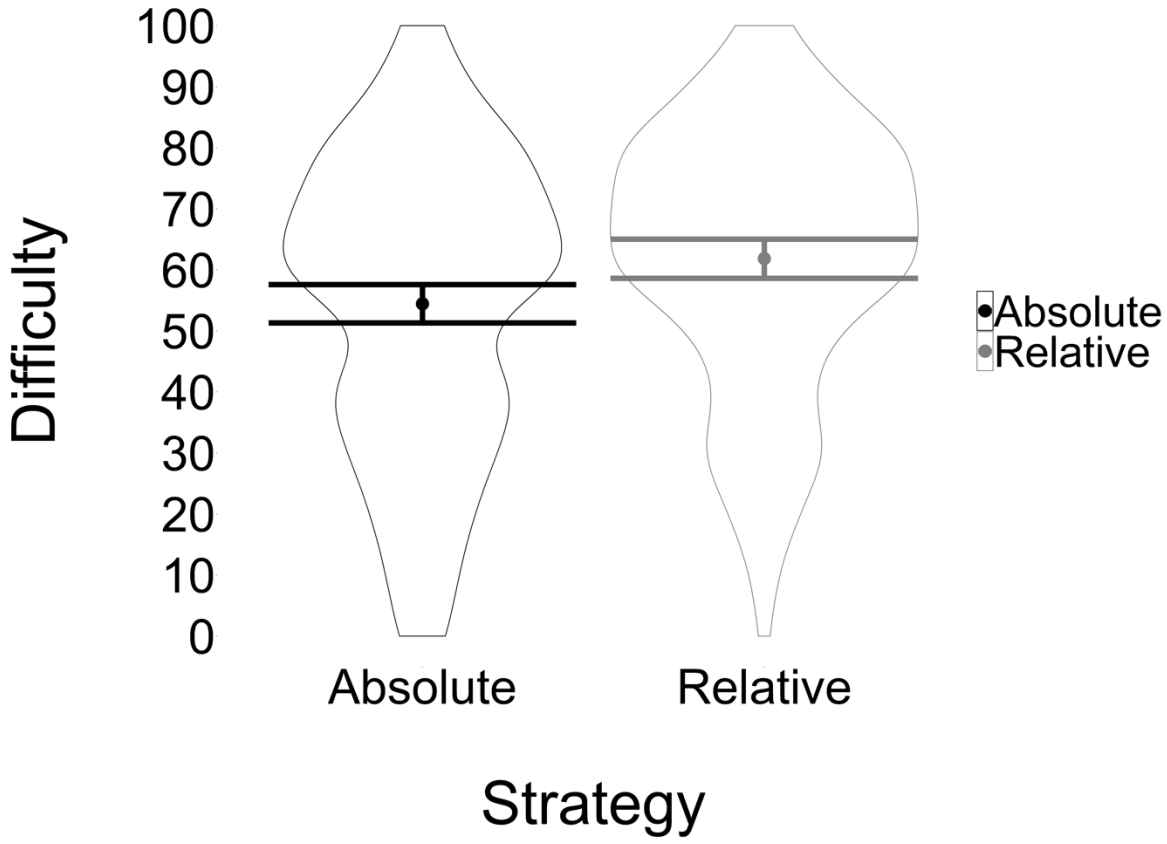


Figure E10. Experiment 1: Difficulty by Strategy. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models). Violin plots represent the distribution of difficulty ratings.

Estimated relative – absolute difference = 7.40 (95% CR[2.91, 11.89]).

iii. BFs

BF = 25 in favor of an absolute versus relative difference in confidence.

F. Experiment 2-specific results

a. Lineup member similarity

i. Carjacking CA lineup

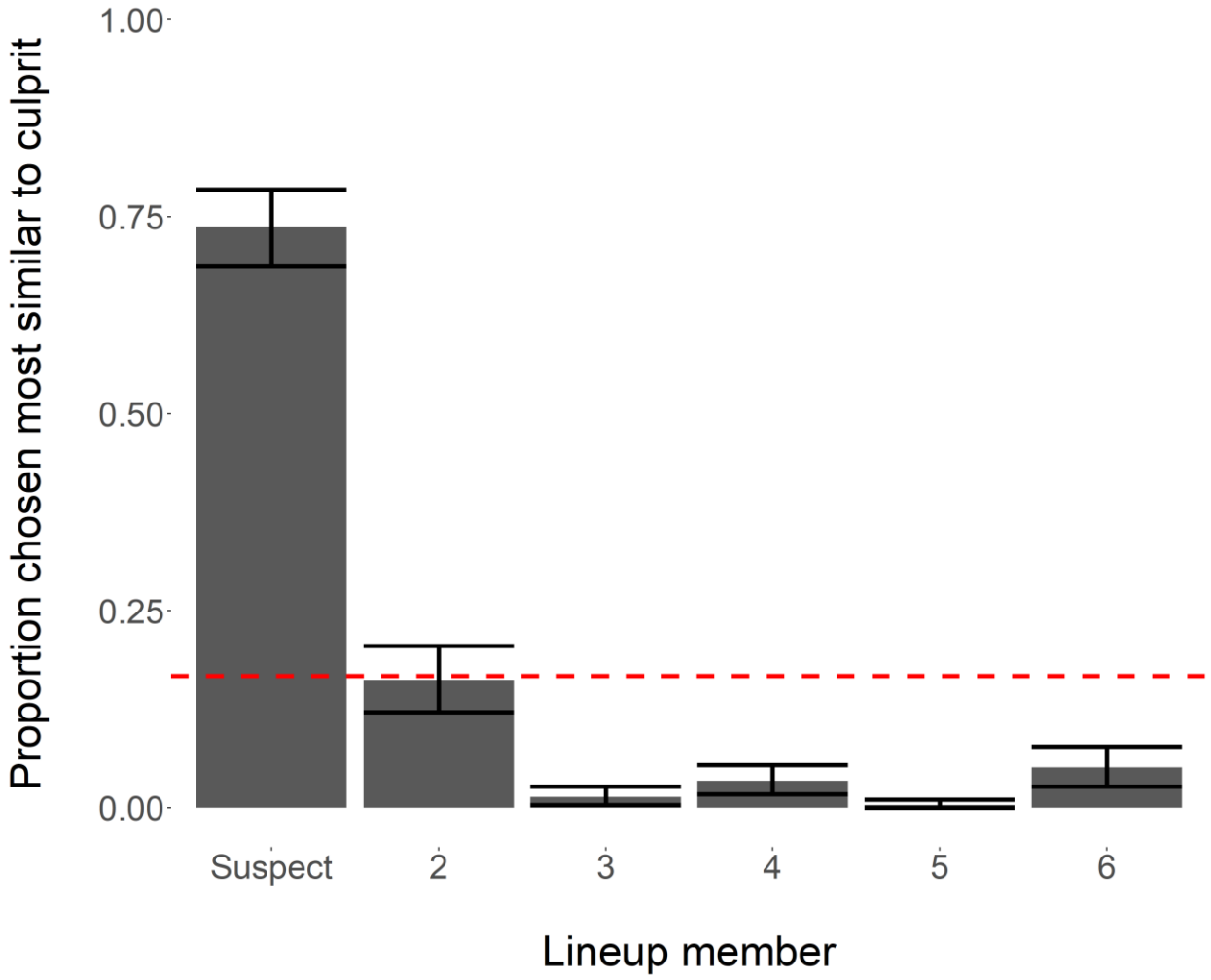


Figure F1. Carjacking culprit-absent lineup: Lineup member choosing. Errors bars = Bootstrapped 95% CRs. Dashed red line = 16.7%.

BF > 100 in favor of suspect choosing > 16.7%.

ii. Theft CA lineup

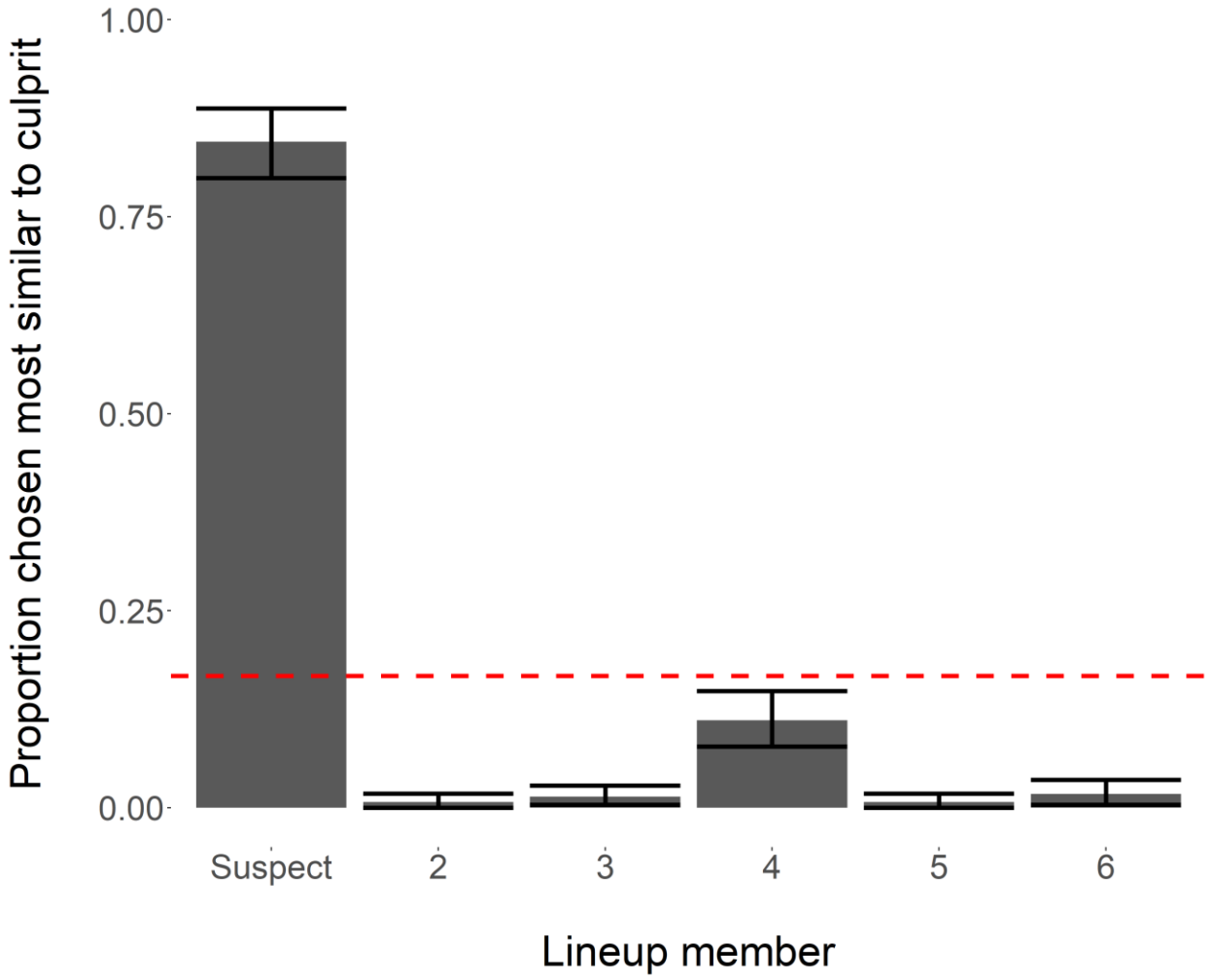


Figure F2. Theft culprit-absent lineup: Lineup member choosing. Errors bars = Bootstrapped 95% CRs. Dashed red line = 16.7%.

BF > 100 in favor of suspect choosing > 16.7%.

a. False IDs

iii. WAIC model comparison results

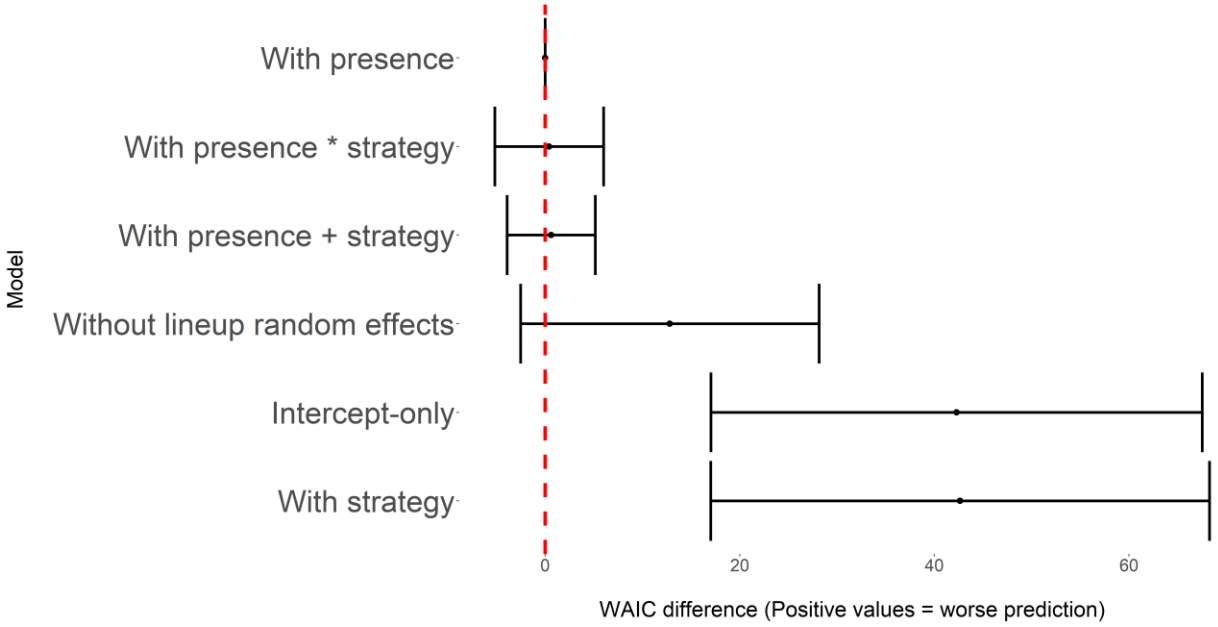


Figure F3. Experiment 2: WAIC model comparison results for predicting p(ID). Models presented in descending order of predictive ability relative to the best model (1st Row). Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

iv. Estimation results

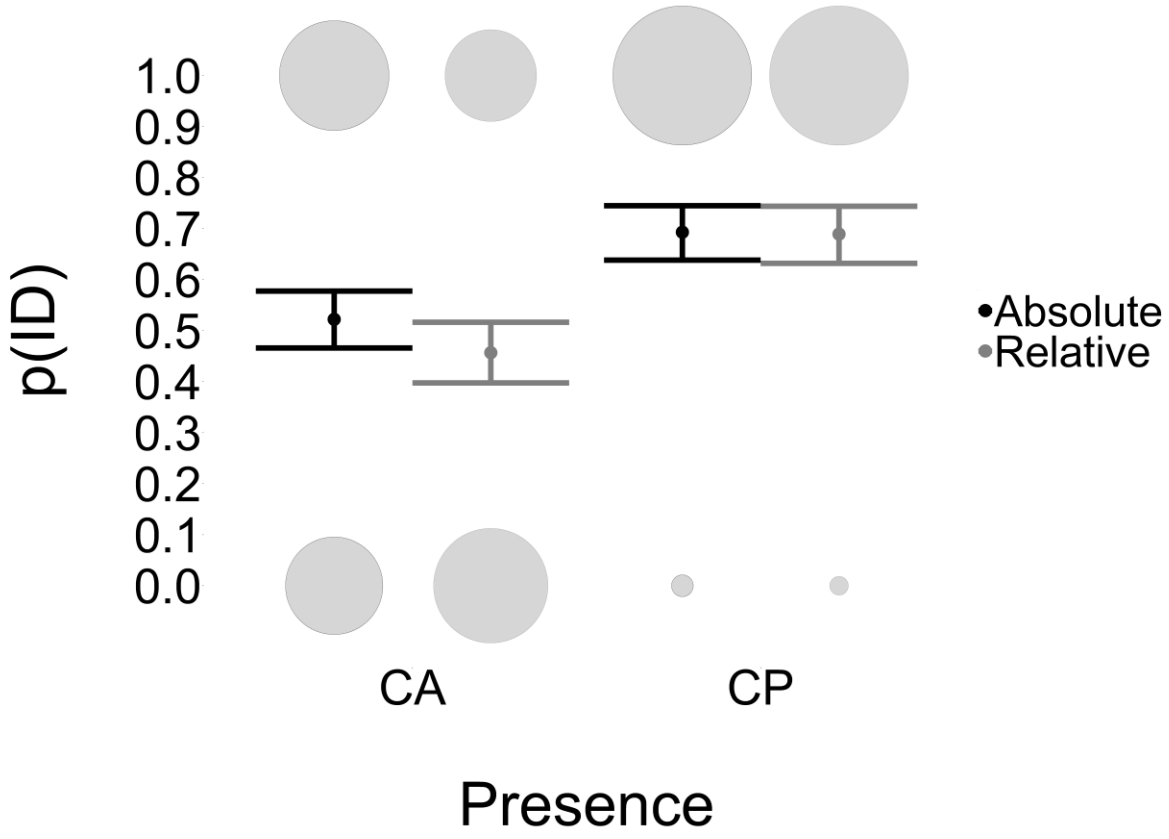


Figure F4. Experiment 2: Probability of making an ID by Strategy and Presence. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models). Shaded circles represent the relative proportion of participants who gave responses at each probability.

Estimated CA absolute – relative difference = 7% [95% CR: -1%, 14%]. Estimated CP absolute – relative difference = 0% [95% CR: -7%, 7%].

v. BFs

BF = 1.25 in favor of an absolute versus relative difference in false IDs. BF = 3.85 against an absolute versus relative difference in correct IDs.

b. Filler IDs

vi. WAIC model comparison results

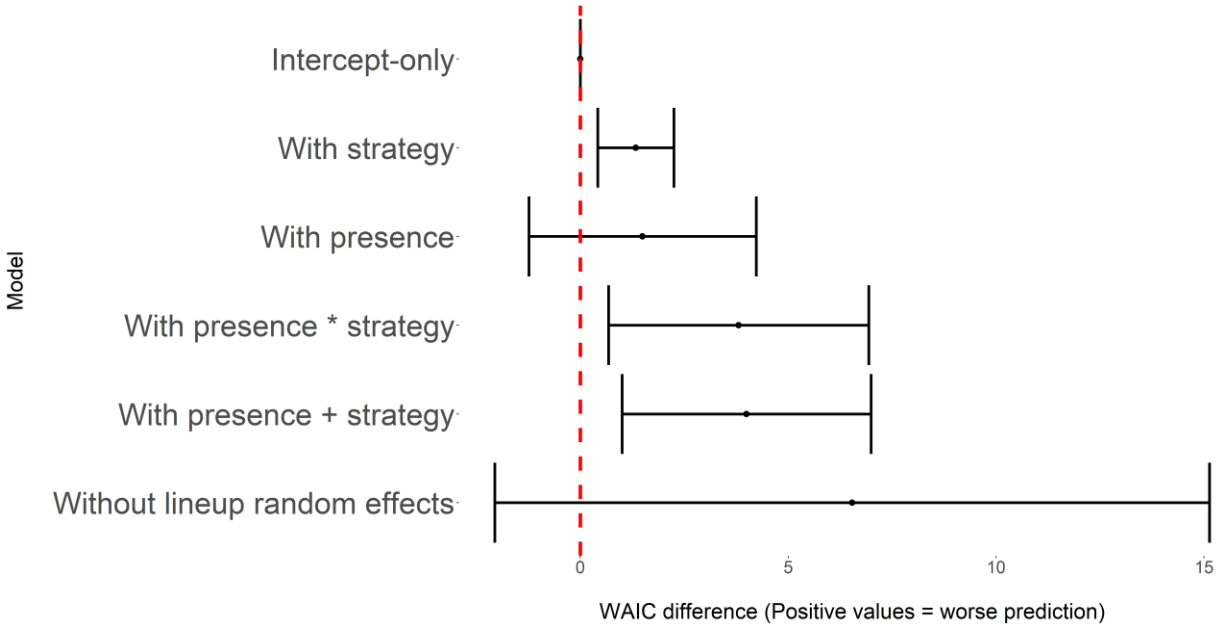


Figure F5. Experiment 2: WAIC model comparison results for predicting p(filler ID). Models presented in descending order of predictive ability relative to the best model (1st Row). Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

vii. Estimation results

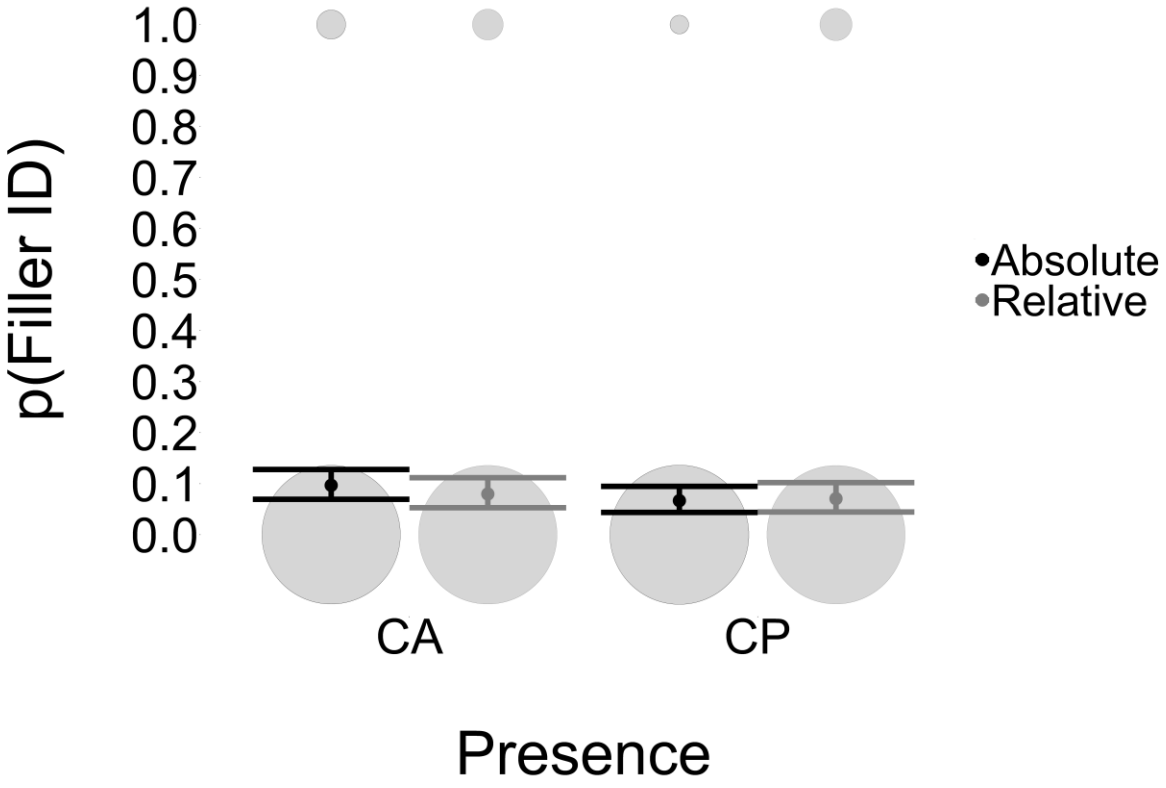


Figure F6. Experiment 2: Probability of making a filler ID by Strategy and Presence. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models). Shaded circles represent the relative proportion of participants who gave responses at each probability.

Estimated CA absolute – relative difference = 2% [95% CR: -2%, 5%]. Estimated CP absolute – relative difference = 0% [95% CR: -4%, 3%].

viii. BFs

BF = 1.47 against an absolute versus relative difference in CA filler IDs. BF = 2.45 against an absolute versus relative difference in correct IDs.

c. Discriminability

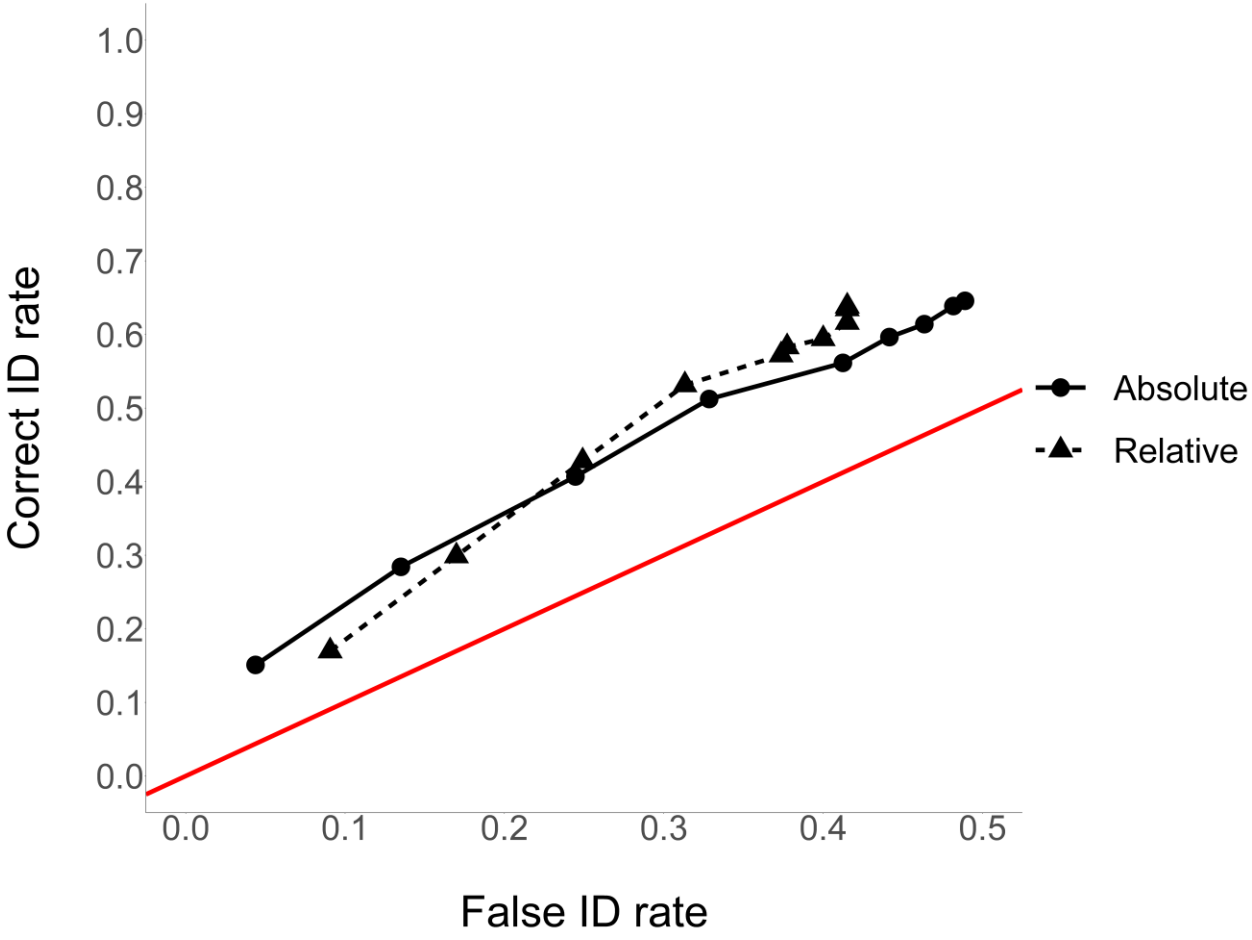


Figure F7. Experiment 2: ROC curves for absolute and relative judgment strategies.

Bootstrap estimate of pAUC difference (relative - absolute) = 1 (95% CR[-.05, .06]). Bootstrapped pAUC p = .81. Bootstrapped estimate of DPP difference = .035 (95% CR[.01, .06]).

d. Confidence

ix. WAIC model comparison results

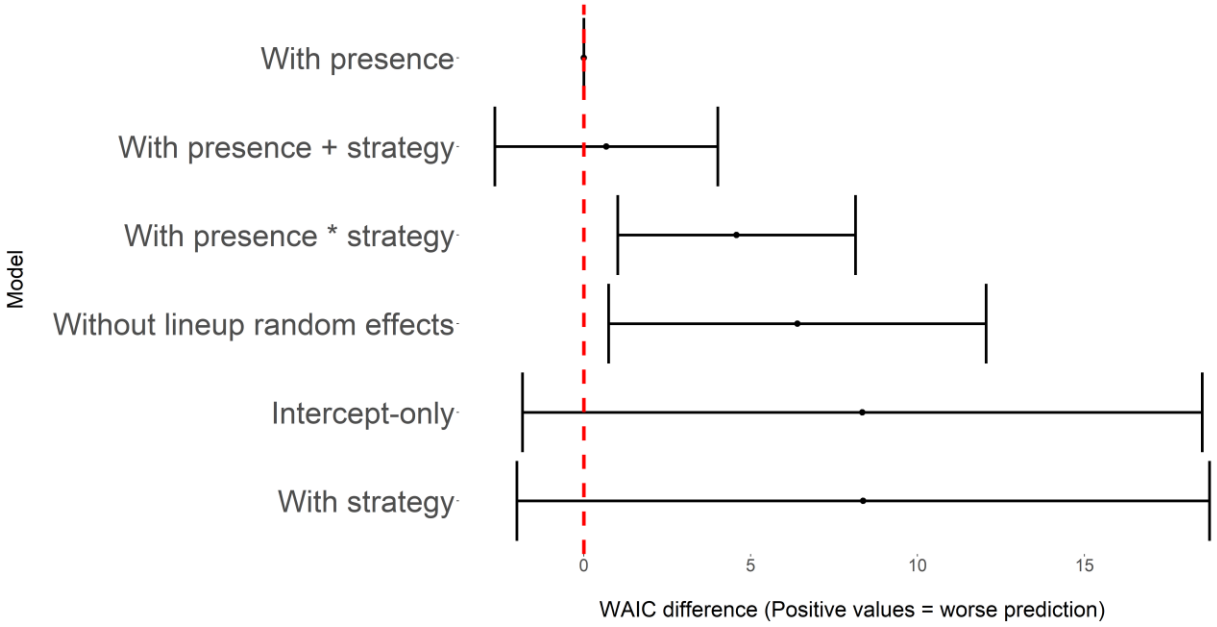


Figure F8. Experiment 2: WAIC model comparison results for predicting confidence. Models presented in descending order of predictive ability relative to the best model (1st Row). Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

x. Estimation results

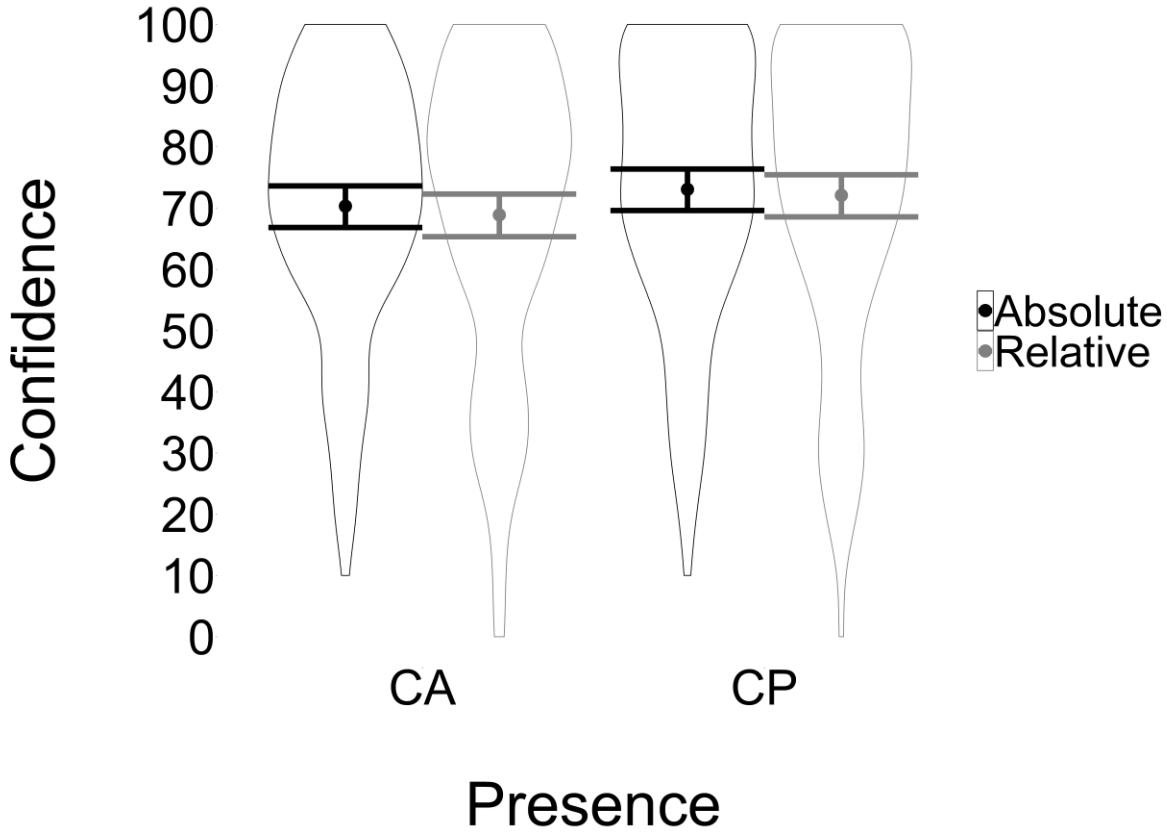


Figure F9. Experiment 2: Confidence by Strategy and Presence. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models). Violin plots represent the distribution of confidence ratings.

Estimated absolute – relative difference for CA lineups = 1.42 (95% CR[-2.29, 5.16]). Estimated absolute – relative difference for CP lineups = .96 (95% CR[-2.73, 4.65]).

xi. BFs

BF = 5.93 against an absolute versus relative difference in confidence for CA lineups. BF = 7.74 against an absolute versus relative difference in confidence for CP lineups.

e. Difficulty

xii. WAIC model comparison results

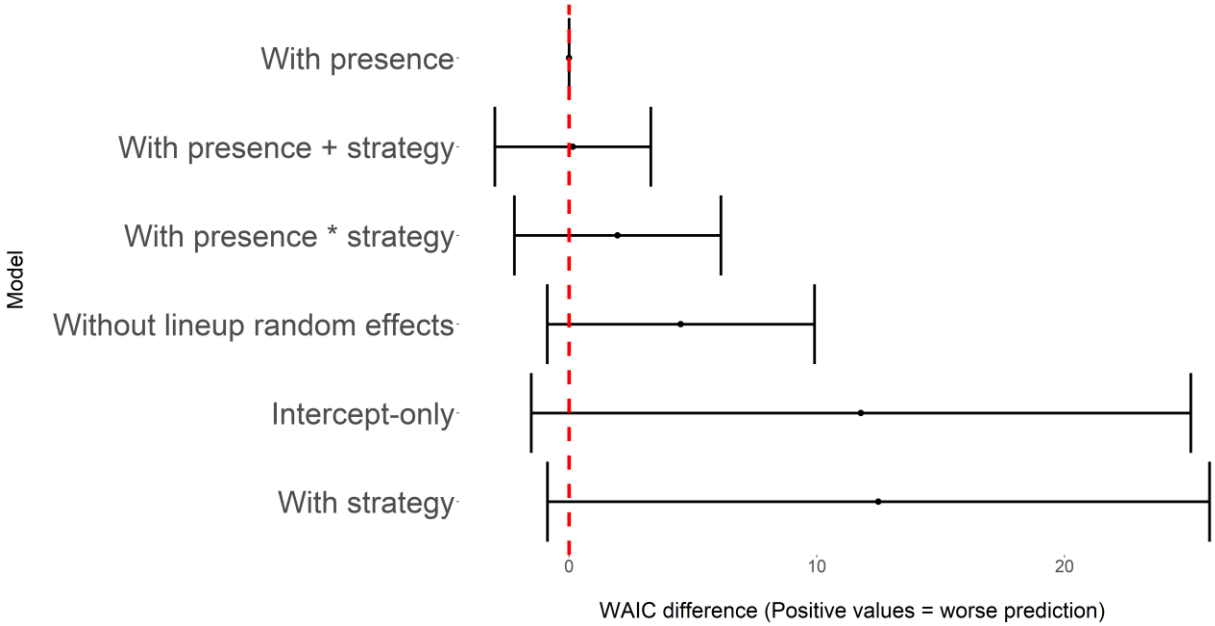


Figure F10. Experiment 2: WAIC model comparison results for predicting difficulty. Models presented in descending order of predictive ability relative to the best model (1st Row). Points represent the estimated WAIC difference from the best model, error bars represent 95% CRs on the WAIC difference.

xiii. Estimation results

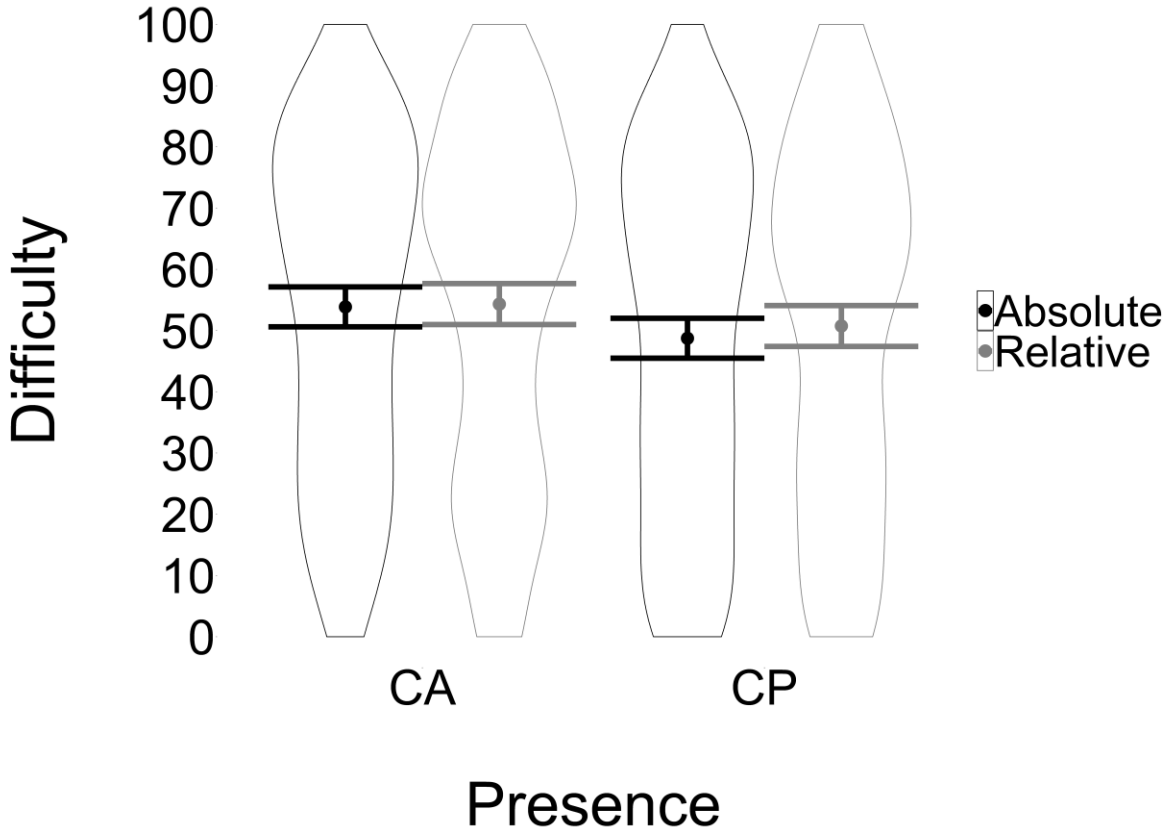


Figure F11. Experiment 2: Difficulty by Strategy and Presence. Points represent the most plausible posterior value, and errors bars (95% CRs) represent the range in which we can be 95% sure that the true value lies (based on these data and models). Violin plots represent the distribution of difficulty ratings.

Estimated relative – absolute difference for CA lineups = .47 (95% CR[-4.02, 4.87]). Estimated relative – absolute difference for CP lineups = 2.03 (95% CR[-2.46, 6.74]).

xiv. BFs

BF = 6.82 against an absolute versus relative difference in difficulty for CA lineups. BF = 4.82 against an absolute versus relative difference in difficulty for CP lineups.