A Meta-Analysis of Lineup Size Effects on Eyewitness Identification

Stefana Juncu[1] and Ryan J. Fitzgerald[2]

[1]University of Portsmouth, UK

[2]Simon Fraser University, Canada

Accepted for publication in *Psychology, Public Policy, and Law*.

Data and analysis code used in this meta-analysis are available on the Open Science Framework and can be accessed at https://osf.io/w9trz/?view_only=d568397743b341799c4f4fefbb1d71c4.

Online Supplemental Material are available at

https://8b7246a0-281b-41f6-b29c-2989305cb2b5.filesusr.com/ugd/e1db49_a0a2358c48ca48d89695b9442e03d9f0.pdf

An earlier version of this research was presented in 2019 at the 13th Biennial Meeting of the Society for Applied Research in Memory and Cognition.

Corresponding author: Stefana Juncu

Email: Stefana.juncu@port.ac.uk

Phone: +44 23 9284 6318

Address: Department of Psychology, University of Portsmouth, King Henry Building, King Henry I Street, Portsmouth, UK, PO1 2DY

Abstract

The inclusion of known-innocent fillers in a lineup is fairer than presenting only the suspect to an eyewitness and offers protection from mistaken identification if the suspect is innocent. This meta-analysis addresses the question of how many fillers should be included in a lineup. Data from 17,088 participants across 14 experiments revealed a trade-off associated with increasing the number of lineup members. Innocent suspects receive better protection from larger lineups than smaller lineups, but larger lineups also make it harder for eyewitnesses to identify guilty suspects. Expected cost analyses showed that the least costly lineup size depends on the base rate of suspect guilt and the cost of incriminating an innocent suspect. If incriminating an innocent suspect is considered 10 times as costly as failing to incriminate the true perpetrator (Blackstone ratio), then larger lineups would be less costly for the majority of possible base rates. Smaller lineups would only be less costly if the base rate of suspect guilt is high, or if incriminating an innocent suspect is considered to have minimal costs. There remains much to learn about lineup size and its potential interactions with filler plausibility and the method of lineup presentation. Nevertheless, these preliminary results suggest that many jurisdictions would benefit from increasing the minimum number of fillers in their lineups.

A Meta-Analysis of Lineup Size Effects on Eyewitness Identification

Conducting a lineup is one of the critical stages of a criminal investigation. Several factors have been shown to affect the reliability of lineup identifications, and psychological research has led to procedural reforms designed to improve lineup identification in practice. These include recommendations on lineup instructions, lineup composition, and double-blind lineup administration (Wells et al., 1998, 2020). Another important consideration is the lineup's nominal size, which refers to the number of people in the lineup. We conducted a meta-analytic review to examine if and how lineup size affects eyewitness identification.

Lineup Size Policies

Policies in many jurisdictions specify the number of lineup members required to adequately protect a potentially innocent suspect (Fitzgerald et al., 2021). A required minimum for the lineup's nominal size is consistent with the reason why the London Metropolitan Police started using lineups in the mid-19th Century (Devlin, 1976). The idea was to provide a fairer option than the suggestive practice of presenting the suspect alone at the identification procedure. This practice, contemporarily known as a showup, can be suggestive and provides no procedural mechanism for knowing when a mistaken identification has occurred (Clark & Godfrey, 2009; Dysart & Lindsay, 2007; Wetmore et al., 2015). A lineup includes fillers to appear alongside the suspect, and best practice is to use fillers who are known to be innocent (Wells & Turtle, 1986). This allows for some false-positive errors to be distributed among the known innocent fillers and results in fewer innocent suspect identifications (Steblay et al., 2003).

How many people should be in a lineup? The answer to this question depends on where you are in the world (Figure 1). In the U.S. there is some variability across jurisdictions, but a nationally-representative survey of American police agencies indicates that most eyewitnesses will see a photo lineup with six lineup members (Police Executive Research Forum, 2013), which is the minimum lineup size recommended in national guidelines (Yates, 2017; Technical Working Group for Eyewitness Evidence, 1999). In the context of other common law countries, U.S. lineups are

comparatively small: South African lineups must include at least eight people (South African Police Service, 2007); English police have been using 9-member lineups for decades (Home Office, 1969, 2017); and in Canada, 10 lineup members are recommended (FPT, 2018). In a review of lineup policies, however, it was not uncommon to find recommended lineup sizes of three or four in civil law countries, and the median recommended minimum was five (Fitzgerald et al., 2021).

Policymakers tend to agree on the need for a minimum lineup size. Of the 54 countries reviewed by Fitzgerald et al. (2021), 80% had a policy on the minimum number of lineup members. This far exceeds the prevalence of the lineup reforms most strongly endorsed by the scientific community (Wells et al., 1998), such as using a blind lineup administrator (9%), obtaining an eyewitness confidence statement (13%), or warning that the perpetrator may be absent from the lineup (20%). Thus, even the policies that lacked the more established procedural safeguards tended to include a minimum lineup size, which may reflect the intuition that a certain number of lineup members is needed for the identification procedure to be fair.

Provided that the quality of fillers remains constant, increasing lineup size should decrease the risk to an innocent suspect. Fillers can attract false positive choices and draw them away from the suspect (Fitzgerald et al., 2013; Lindsay & Wells, 1980). If every filler is an effective alternative, the likelihood of an eyewitness misidentifying an innocent suspect would be inversely proportional to the lineup's nominal size, such that the chances of an identification landing on an innocent suspect would be 1/9 in a 9-member lineup, 1/6 in a 6-member lineup, and so on. This logic, combined with early research showing no cost of increasing lineup size on correct identifications (Nosworthy & Lindsay, 1990), has been invoked to support arguments for increasing the minimum lineup size in the U.S. from 6 to 9, 12, or even 20+ lineup members (e.g., Levi & Lindsay, 2001).

The challenge posed by a requirement for large lineups, however, is that it becomes harder to meet the conventional standards of lineup fairness. Lineup fairness is typically judged by the proportion of lineup members who match the eyewitness description of the perpetrator (Malpass,

1981; Wells et al., 1979). If an eyewitness reports that the perpetrator had bushy eyebrows, the perceived fairness of a lineup would be compromised by the inclusion of any fillers with thin eyebrows. In cases with common eyewitness descriptions, it should be possible to find enough suitable fillers for 9- or 12-member lineups. But eyewitnesses typically report 7-9 descriptors (Wells et al., 2020) and when these descriptors combine into a more idiosyncratic description, there may be an upper bound to the number of available fillers who would be effective alternatives to the suspect.

If a lineup size policy is too demanding and not supported with resources for effective implementation, it could preclude conducting a lineup altogether. Wells (2001) was supportive of increasing the size of U.S. lineups in principle, but warned that police are not always able to find more than five suitable fillers in practice, particularly if the suspect has distinctive features or belongs to a minority group. Wells noted that if a policy required a minimum of 20 lineup members, as some have advocated (Levi & Lindsay, 2001), it would become impossible to conduct a legal lineup for certain suspects. Lineups used to be regularly cancelled in England, and one of the reasons for cancellation was that the police could not locate enough suitable fillers (Pike et al., 2002). England has long had one of the more stringent lineup size policies, with a required minimum of nine members in all lineups, and over a third of the English police officers (N = 50) reported that they often encountered problems obtaining enough suitable lineup fillers (Pike et al., 2002). However, at the time of the cancellations, lineups in England had to be administered live. This meant that fillers had to be recruited locally and physically present for the identification procedure. England has now transitioned to a system of video lineups, which has a much lower cancellation rate and is supported by centralized filler databases with tens of thousands of images, such as that provided by the National VIPER Bureau (https://www.viper.police.uk). The lack of this type of filler database was one of the reasons that Wells opposed raising the minimum lineup size in the U.S.

To summarize, lineup size policies vary widely and there could be practical considerations for any jurisdiction planning to increase the required minimum number of lineup members. Provided that a system is in place to make a wide range of filler images available, a policy that requires lineups to be on the larger end of the spectrum could be a concrete measure for protecting innocent suspects. However, there are many questions about lineup size with implications for policy that remain unanswered. Most important, does the empirical literature support the prediction that increasing nominal lineup size reduces the risk to innocent suspects? And does lineup size affect correct identifications when the lineup contains a guilty suspect?

Does lineup size affect suspect identifications?

Eyewitness scientists once questioned whether lineup size has any effect on eyewitness identification decisions. Following mixed results in early experimental work comparing 6- and 12-member lineups (Cutler et al., 1987; Cutler et al., 1986), Nosworthy and Lindsay (1990) reported that if a lineup already had a few plausible fillers, increasing the lineup's nominal size had no effect on identification outcomes. In their first experiment, the added lineup members were all duds – people who looked nothing like the culprit – so the null effect of lineup size was unsurprising. But in their second experiment, Nosworthy and Lindsay created a pool of lineup members who all resembled the culprit and still found no effect of adding them to a lineup. They compared five lineup sizes, ranging from 4 to 20, and regardless of how many lineup members were added or whether the confederate from a staged crime was included or not – lineup size did not seem to matter. Nosworthy and Lindsay concluded that only a few good fillers were needed to protect innocent suspects and questioned the need for larger lineups.

Despite these early findings, simulation research provides insight into the conditions when lineup size would be expected to matter. Using Clark's (2003) WITNESS model, Wetmore et al. (2017) explored the potential interactions between lineup size, lineup fairness, and lineup presentation. In fair lineups, adding fillers tended to reduce identifications of both guilty and innocent suspects, producing the same type of trade-off found in previous reviews of other lineup

procedures (Clark, 2012; Palmer & Brewer, 2012). It is considered a trade-off because the benefit of increasing protection for innocent suspects in larger lineups would also come with a cost of fewer guilty suspect identifications. The simulations indicated that increasing the size of fair lineups would cause a trade-off regardless of whether the lineup was presented simultaneously or sequentially, which corresponds with findings from the only published experimental work that manipulated nominal size in both simultaneous and sequential lineups (Meissner et al., 2005).

The simulations also revealed two scenarios in which nominal lineup size would not be expected to affect suspect identifications (Wetmore et al., 2017). First, if the innocent suspect strongly resembled the culprit and the lineup was biased against that innocent suspect, such that the lineup fillers were not matched to the suspect's appearance, the WITNESS model indicated that similar rates of suspect identification would be found in 3-, 6-, and 12-member lineups. Under these parameter specifications, WITNESS predicted that suspect identifications would be common, filler identifications would be rare, and lineup size would have no influence on identification outcomes. Second, if there was only a moderate resemblance between the fillers and the suspect, increasing lineup size was predicted to result in only a small decrease in suspect identifications. This was true regardless of whether the suspect was guilty or innocent; and the pattern was also unaffected by whether the innocent suspect was highly similar to the culprit (biased lineup) or, as with the fillers, only moderately similar to the culprit (fair lineup). Thus, one insight to be gleaned from the simulations is that increasing lineup size is only likely to have an effect if the added fillers are sufficiently matched to the appearance of the suspect.

Despite the null findings in early research and resulting suggestions that nominal size is not an important policy consideration, more recent findings suggest that lineup size policies would indeed have consequences on the suspect identification rate. There are a variety of potential explanations for the nonsignificant effects observed in the influential study by Nosworthy and Lindsay (1990). They did not report quantitative measures of suspect-filler similarity, so it is possible that the fillers were not matched closely enough to the suspect to have an effect. A Type

II error is another possibility, as Nosworthy and Lindsay only tested 27-32 participants per condition. In two recently published lineup size experiments, which both tested thousands of participants (Akan et al., 2020; Wooten et al., 2020), lineup size produced the exact trade-off pattern predicted in the WITNESS simulations (Wetmore et al., 2017). Therefore, regardless of whether the suspect is guilty or innocent, we predict the meta-analysis will show that the likelihood of a suspect identification is inversely related to the nominal size of the lineup.

Does lineup size affect discrimination between guilty and innocent suspects?

Criminal investigators use identification procedures to discriminate between suspects who are guilty and suspects who are innocent. Showups are regarded as an identification procedure with poorer suspect guilt discriminability than lineups because showups increase the risk of an innocent suspect misidentification and also decrease the chance of a guilty suspect identification (Clark, 2012; Neuschatz et al., 2016). However, a procedure's effect on discrimination is not always so clear-cut. When a procedural intervention causes guilty and innocent suspect identifications to shift in the same direction (i.e., both outcomes increase or decrease), *d'* or Receiver Operating Characteristic (ROC) analysis can be calculated from hit and false alarm rates to assess the net gain or loss in suspect guilt discrimination. For eyewitness identification tasks, hits refer to the correct identification rate in culprit-present lineups and false alarms typically refer to the innocent suspect identification rate in culprit-absent lineups. Filler identifications are also false alarms, but they are known errors and thus are treated the same as nonidentifications in calculations of suspect guilt discrimination. In signal detection theory, *d'* is the difference between *z*-transformed hit and false alarm rates (Macmillan, & Creelman, 1991). ROC analysis uses eyewitness confidence ratings to plot hit and false alarm rates on a curve (Mickes et al., 2012). Identification responses in the primary literature are not always reported separately at different levels of confidence, and we are unable to fully incorporate ROC curves in our meta-analysis of lineup size. However, we are able to assess suspect guilt discriminability in the meta-analysis using *d'* (Mickes et al., 2014), and in this section we review findings from the limited number of ROC analyses of lineup size.

The relation between lineup size and discriminability may be informed by theoretical explanations of the lineup advantage over showups. According to the diagnostic feature detection hypothesis (Wixted & Mickes, 2014), presenting a suspect with fillers in a simultaneous lineup helps the eyewitness determine which facial features are shared across the lineup members. The diagnostic feature detection hypothesis assumes these shared features are non-diagnostic and predicts that simultaneous lineups allow eyewitnesses to discount the shared features and focus their attention on features that are diagnostic of the suspect's guilt. When fillers are from a Gaussian distribution, and therefore not all highly similar to the culprit, this hypothesis predicts that lineup outcomes will improve as lineup size increases (Wixted et al., 2018). Namely, as more fillers are added, there is a higher chance that eyewitness will recognize the non-diagnostic features that can be ignored and increase their focus on the features that are diagnostic.

Filler siphoning is another explanation for the lineup advantage over showups (Wells et al., 2015). Fillers tend to attract a portion of the mistaken identifications and draw choices away from the suspect. Increasing the size of lineup could increase this siphoning effect, but suspect guilt discriminability would only be affected if the extent of the filler siphoning is moderated by suspect guilt. Smith et al. (2017) proposed a differential filler siphoning account of the lineup advantage over showups, arguing that innocent suspects would tend to match the eyewitness' memory less than guilty suspects and, therefore, more identifications should be siphoned away from innocent suspects than from guilty ones. Accordingly, when lineup size increases and there are more fillers to attract identifications, increased siphoning would be predicted for both culprit-absent and culprit-present lineups; however, because the increase in siphoning would be more pronounced in culprit-absent lineups, suspect guilt discriminability would improve as lineup size increases.

Discriminability has been assessed in a small number of lineup size studies. Using a repeated-measures lineup paradigm, Meissner et al. (2005) computed a nonparametric measure of suspect guilt discrimination (*A'*) and found that larger lineups performed better than smaller lineups. In subsequent research, however, ROC analysis revealed no effect of lineup size on suspect

guilt discriminability (Akan et al., 2020; Wooten et al., 2020). One potential limitation of these ROC studies is that discriminability was estimated via partial Area Under the Curve (pAUC). To compare pAUCs, common practice is to truncate the curve of the procedure with the highest false alarm rate. Although truncation is necessary for a fair comparison of pAUCs, it results in information loss and could affect interpretation of the results (Smith et al., 2019). Another potential limitation of conventional ROC analysis of lineup data is that only suspect identifications are taken into account and all other outcomes are ignored.

Smith et al. (2020) proposed the use of full ROC curves, which incorporate data from all identification outcomes and provide a measure of investigator discriminability. With this approach, it is proposed that investigators could use any identification outcome and the associated eyewitness confidence rating to inform their assessment of suspect guilt. For instance, a full ROC might reveal that a low confidence filler identification is highly diagnostic of innocence. A firm lineup rejection could be similarly diagnostic of innocence. Although the data required for this approach has not typically been reported in lineup size studies, for all studies in which the data was available we plotted full ROCs. In each of the plots, which are reported in Online Supplemental Materials, the curves associated with different lineup procedures intersect. When this happens, Smith and colleagues (2020) recommend conducting expected utility analyses as a follow-up. We used the same data to plot Confidence Accuracy Characteristic curves (Mickes, 2015) which can also be found in the Online Supplemental Materials.

What are the expected costs of lineup size?

If a change in lineup size would result in a trade-off, a clear policy recommendation is unlikely to emerge from a simple analysis of identification outcomes. Even if certain lineup sizes were found to improve suspect guilt discriminability, this would have to be interpreted in the context of the prior probability of guilt and the costs associated with each identification outcome (Clark, 2012; Wells et al., 2015). After observing no effect of lineup size on suspect guilt discriminability, Wooten et al. (2020) suggested that 3-member lineups may perform similarly with

12-member lineups; however, this interpretation assumes a 50% likelihood that the suspect is guilty and that misidentifying an innocent suspect identification would be no more costly than failing to identify a guilty suspect. A different interpretation could emerge by considering different suspect guilt probabilities and the costs associated with different identification outcomes. Such considerations can be formally modelled by calculating (dis)utility values for different identification outcomes based on social and policy considerations (e.g., Clark, 2012; Lampinen et al., 2019; Malpass, 2006). One implementation of this type of utility analysis is the expected cost model (Yang et al., 2019), which informs policy by determining which identification procedure yields the smallest expected cost.

The expected cost model provides a nuanced interpretation of the trade-off between guilty and innocent suspect identifications by incorporating three input parameters: the conditional probability of an identification outcome, the prior probability that the suspect is guilty, and the cost associated with each identification outcome (Yang et al., 2019). Conditional probabilities are outcome response rates, such as suspect identifications, filler identifications, and lineup rejections. The prior probability of guilt, also known as the guilty base rate, is the probability that the suspect is guilty before conducting the lineup. Although a 50% base rate is common in experimental studies, the prior probability of guilt in the real world is hard to assess and would likely vary across jurisdictions. One factor that could feed into the guilty base rate is the amount of incriminating evidence against a suspect. Wixted et al. (2016) used lineup outcomes from the Houston Police Department to model the guilty base rate and estimated that only a third of lineups included a guilty suspect. One explanation for such a low base rate is that U.S. guidelines do not specify a minimal amount of evidence that would be required to conduct a lineup. Archival data from Northern California indicate that in 40% of the reviewed cases, there was no evidence against the suspect before conducting a lineup (Behrman & Richards, 2005). To increase the guilty base rate, Wells et al. (2020) recommend only conducting a lineup if there is reasonable suspicion of guilt. This would

correspond with Danish law, which states that lineups may only be conducted if the accused is reasonably suspected of a serious offence (Fitzgerald et al., 2021).

The cost of an identification outcome is the discrepancy between the identification outcomes and the goals of the police investigation (Yang et al., 2019). If an identification procedure successfully achieves the goal of identifying the culprit, the associated cost of this outcome is zero. Table 1 shows that for all identification outcomes that do not achieve this goal, a cost is incurred for failing (*f*) to incriminate the culprit. All other costs are defined in relation to *f* (cost = 1). If an innocent suspect is identified from a culprit-absent lineup, there is also a cost for wrongfully incriminating an innocent suspect, which has historically been conceptualized as a ratio (*r*) to reflect societal beliefs about the increased cost of incriminating an innocent person in relation to the cost of failing to incriminate the perpetrator. For instance, Blackstone (1769, page 352) famously opined, "it is better that ten guilty persons escape, than that one innocent suffer." If this principle is applied, $r = 10$. Whenever an innocent suspect is incriminated, there is the additional cost of failing to incriminate the culprit (Steblay et al., 2011; Wells et al., 2012). Therefore, the total cost of an innocent suspect identification is calculated by summing the costs of *f* and *r*. For example, applying Blackstone's ratio would result in a total cost of 11 (1 + 10).

Table 1

Costs associated with each lineup outcome

Lineup	**Suspect ID**	**Filler ID**		**Lineup Rejection**
		Equal Cost (Filler = Rejection)	**Separate Cost** (Filler > Rejection)	
Culprit Present	0	f	$f + i$	f
Culprit Absent	$f + r$	f	$f + i$	f

Note. 0 = no cost, f = cost of failing to incriminate the culprit, r = the ratio between the cost of incriminating an innocent suspect and the cost of failing to incriminate the culprit, i = the cost of impeached witness. Equal Cost Analysis does not account for filler identifications impeaching witness credibility and thus assumes equal costs for filler identifications and lineup rejections. Separate Cost Analysis assumes that filler identification are twice as costly as rejections, due to the effect that filler identifications can have on the witness' credibility.

Yang et al. (2019) considered two ways to assign costs for filler identifications. One approach is to assign an equal cost for filler identifications and lineup rejections, as both outcomes incur the cost of failing to incriminate the culprit (*f*). This is referred to as Equal Cost Analysis. However, filler identifications are arguably more costly for the criminal justice system than lineup rejections because an eyewitness who identifies a filler cannot be tested with another lineup if a new suspect becomes the focus of investigation (Steblay et al., 2011; Wells et al., 2012). Thus, the second approach is to assign an additional, separate cost for filler identifications for impeaching (*i*) the credibility of the witness. Yang and colleagues refer to this approach as Separate Cost Analysis and set $i = 1$, which causes filler identifications ($f + i = 1 + 1 = 2$) to become twice as costly as lineup rejections ($f = 1$).

The overall expected cost of a procedure (*c*) is the average of the costs of all identification outcomes (Yang et al., 2019). A procedure is considered superior if its expected cost is lower than that of an alternative procedure across the full range of plausible cost ratios and guilty base rates. Given that lineup size is anticipated to result in a tradeoff, a more likely scenario is that the expected costs of different lineup sizes will intersect. When this happens, policy guidance may need to be contingent upon assumptions about the prior probability of guilt and societal views toward protecting the innocent (Yang et al., 2019). For instance, if current guidelines for conducting a lineup in the U.S. are assumed (i.e., no corroborative evidence is needed), the expected cost analysis might support a recommendation for a relatively large lineup. Alternatively, if Danish guidelines for conducting a lineup are assumed (i.e., reasonable suspicion is needed), the expected costs analysis might support a recommendation for a smaller lineup. These scenarios could be upended by specifying an exceptionally high ratio for the cost of protecting innocent suspects in relation to losing identifications of guilty suspects (*r*), which is value-driven and in theory has no upper bounds. Indeed, Blackstone's *r* has not been universally endorsed, and arguments have been made for $r = 1$, $r = 100$, and even $r = 1000$; however, in case law and academic literature $r = 10$ is the

most common (Clark, 2012; Volokh, 1997). In this meta-analysis, we explore a variety of cost ratios with a particular focus on the change in expected costs when *r* is increased from 1 to 10.

Potential Moderators of Lineup Size Effects

Lineup size effects could be moderated by a variety of factors, many of which are beyond the scope of this meta-analysis. Notwithstanding the recent uptick in interest (Akan et al., 2020; Seale-Carlisle et al., 2019; Wooten et al., 2020), lineup size has long been a neglected area of study and the literature has not yet matured to the stage for a comprehensive evaluation of all potential moderators. Although these shortcomings restricted the scope of the meta-analysis, we applied methods to minimize the influence of potentially important variables that could not be subjected to a formal moderator analysis.

Lineup size might have a more substantial impact if the lineup is presented sequentially. Sequential presentation forces the eyewitness to consider each lineup member individually. Thus, compared with simultaneous lineups, witnesses would be less free to skim a sequential lineup and focus their attention on the most plausible candidates. The size of sequential lineups could be especially consequential if the suspect is placed in a later position and the procedure is set to terminate whenever an identification is made. Unfortunately, these possibilities cannot be tested via meta-analysis at this time. We know of only three lineup size studies that included sequential lineups: one that does not include simultaneous lineups, one that is unpublished and another that does not report identification response rates separately for the sequential and simultaneous conditions (Table 2). Given that we lacked the data needed for a meaningful moderator test, we report a separate meta-analysis with all sequential lineup data excluded as a sensitivity analysis to assess whether the sequential data were having an influence on the summary effects.

Lineup size effects are also likely to be moderated by the similarity between the fillers, the innocent suspect, and the culprit. Consistent with Wooten and colleagues' (2020) findings, Wetmore et al. (2017) showed that when the innocent suspect and the fillers were modelled to be equally similar to the culprit, increasing lineup size resulted in a lower rate of suspect

identifications and a higher rate of filler identifications. When the innocent suspect was a better match to the culprit than the fillers, the similarity of fillers also mattered. If the fillers were bad matches for the culprit, lineup size did not affect suspect identifications. However, if they were good matches, suspect identifications decreased and choosing increased as more fillers were added to a lineup. These types of similarity manipulations would be ideal candidates for moderator analyses, but within-study examinations of these variables in the lineup size literature are either rare or non-existent, and a between-study moderator analysis would depend on a tradition of measuring the similarity of the lineup members in lineup size experiments. Regretfully, no such tradition exists in the literature we reviewed and filler similarity could not be included as a moderator variable.

Methodological characteristics could also moderate lineup size effects. For instance, a lineup should only have one suspect, but researchers varied in whether they designated one of the culprit-absent lineup members to be the innocent suspect. When no culprit-absent lineup member is designated, the innocent suspect identification rate is typically estimated by dividing the overall choosing rate from culprit-absent lineups by the number of people in the lineup. This approach imposes an upper limit on the innocent suspect identification rate (i.e., the inverse of the lineup's nominal size) and could underestimate the risk to innocent suspects. However, less than 1/3 of the studies included a designated innocent suspect (Table 2), so to maintain consistency we applied the nominal size correction to estimate the innocent suspect identification rate for all studies. Another methodological issue is that in some studies, the identity of the lineup members was not counterbalanced across the lineup size manipulation (Table 2). In other words, the number of lineup members was confounded with the identity of the lineup members, such that the larger lineups included fillers the smaller lineups did not. The absence of counterbalancing could lead to misleading results. For example, if everyone in the smaller lineup is a good match to the culprit and the fillers added to the larger lineup are not good matches, the study would say more about the

Table 2

Study Characteristics

Study	Published	*N*	Lineup Sizes	Lineups per Subject	Stimulus Counterbalancing	Designated Innocent Suspect	Lineup Presentation
Akan et al. (2020) Exp 1	Yes	4401	2, 4, 6, 8	1	Yes	No	Simultaneous
Bailey (2011)	No	109	6, 8	1	No	No	Simultaneous
Brewer et al. (2006)	Yes	196	4, 8, 12	2	Yes	No	Simultaneous
Cole (1985) Exp 3	No	118	6, 9	1	No	No	Simultaneous
Cole (1985) Exp 4	No	48	6, 9	1	No	No	Simultaneous
Juncu & Fitzgerald (in prep)	No	1211	4, 6, 8	1	Yes	Yes	Simultaneous
Meissner et al. (2005) Exp 3	Yes	260	2, 4, 6, 8, 10, 12	16	Yes	No	Simultaneous, Sequential
Nosworthy & Lindsay (1990) Exp 1	Yes	192	4, 7, 10	1	No	Yes	Simultaneous
Nosworthy & Lindsay (1990) Exp 2	Yes	270	4, 8, 12	1	No	Yes	Simultaneous
Pozzulo et al. (2010)	Yes	89	6, 12	1	No	No	Simultaneous
Seale-Carlisle et al. (2019)	Yes	1938	6, 9	1	Yes	No	Sequential
Steblay & Baumann (2010)	No	324	6, 12	1	No	Yes	Simultaneous, Sequential
Wagenaar & Veefkind (2011)	Yes	548	2, 6, 10	1	Yes	No	Simultaneous
Wooten et al. (2020)	Yes	10433	3, 6, 9, 12	1	No	No	Simultaneous

quality of the fillers than it would about lineup size. Counterbalancing holds filler quality constant and better isolates the effect of lineup size. Therefore, we report a sensitivity analysis that includes only the counterbalanced studies.

Research Questions

We aim to address the following research questions:

(1) Does lineup size affect the suspect identification rate? If so, is the effect moderated by the guilt of the suspect?

(2) Does lineup size affect discrimination between guilty and innocent suspects?

(3) What are the expected costs of increasing/decreasing lineup size?

Method

Literature Search

Search procedures. One of the authors had a repository of studies comparing different lineup sizes, which comprised our initial sample. We also used the Google Scholar and PsychINFO databases to search for any additional studies using various combinations of the following terms: *lineup size, accuracy, identification, nominal size.* Next, we checked the reference sections of the located studies as well as citation records to locate additional studies. Finally, we emailed 55 researchers who have previously published articles examining eyewitness identification and requested unpublished data or manuscripts that would meet our inclusion criteria. Of the researchers contacted, 31 replied and three sent unpublished data that met the inclusion criteria. A flowchart of the search procedure is depicted in Figure 2, using the PRISMA format (Preferred Reporting Items for Systematic Reviews and Meta-Analyses: Moher et al., 2009).

Inclusion criteria. In order to be included in the final sample, an experimental study needed to meet the following criteria: (a) used a between-subjects manipulation of two or more lineup sizes, which had a minimum of lineup size of 2 and a maximum lineup size of 15; (b)

tested identification using a lineup containing a single previously-encountered person or zero previously-encountered persons (lineups containing multiple culprits or lineups containing a culprit and a bystander were beyond the scope of this meta-analysis); (c) asked participants to make a categorical lineup decision (i.e., they identified one person from the lineup or they rejected the lineup); (d) reported the results for culprit-present lineups and culprit-absent lineups separately; (e) reported sufficient information to compute an odds ratio.

Final dataset. The search ended in April 2020. A total of 14 experiments (64% published) from 12 manuscripts met the inclusion criteria. Manuscript year ranged from 1985 to 2020. In total, data from 8,797 participants were extracted for culprit-present lineups and data from 8,291 participants were extracted for culprit-absent lineups. Several studies included comparisons of more than two lineup sizes and therefore multiple comparisons were included. If a study had subgroups and (a) the subgroups had fewer than 15 participants each and (b) the subgroups were not the product of a lineup size or culprit-presence manipulation, then we collapsed the data for subgroups to produce a more stable effect size estimate. Otherwise, we computed separate effect sizes for all subgroups.

Meta-Analytic Procedure

Outcome measures. The first author and a research assistant independently reviewed and extracted the data from each article that met the inclusion criteria. Separate analyses were computed for culprit-present lineup outcomes (guilty suspect identifications, filler identifications, and rejections) and culprit-absent lineup outcomes (innocent suspect identifications, filler identifications, and rejections). For culprit-absent lineups, not all researchers had a designated innocent suspect. In order to be consistent, for all studies we estimated suspect identifications from culprit-absent lineups as the proportion of all false positive identifications divided by the lineup's nominal size. Researchers occasionally provided a 'not sure' option. For ease of comparison among studies that did or did not provide

this option, all 'not sure' outcomes were treated as lineup rejections. In studies with culprit-absent and culprit-present conditions, two additional outcomes were calculated, choosing and suspect guilt discriminability. Discriminability was calculated at the group level using the formula $d' = z\text{H} - z\text{FA}$ (Macmillan & Creelman, 1991; Mickes et al., 2014), where H is the rate of hits (guilty suspect identifications) and FA is the rate of false alarms (innocent suspect identifications). In this context, *d'* does not refer to underlying discriminability (i.e., the ability of participants to discriminate between innocent and guilty lineup members; Wells et al., 2015). Instead, it refers to empirical discriminability or the effectiveness of the lineup procedure in discriminating between guilty and innocent suspects over the long run (Wixted & Mickes, 2015, 2018).

We categorised the lineup sizes into "smaller" and "larger" groups. The categorization was within-study and thus reflected relative differences in lineup sizes. For instance, a 6-member lineup was categorized as smaller if compared with a 9-member lineup and categorized as larger if compared with a 3-member lineup. In studies with more than two lineup sizes, we calculated effect sizes for every comparison between smaller and larger lineups (e.g., 3 vs 6, 6 vs 9, 3 vs 9). For effect size calculations, identification outcomes were treated as binary (e.g., the guilty suspect was identified or not identified) and log odds ratios were calculated. These effects were weighted and meta-analytically summarized on the log scale, but are reported as odds ratios for ease of interpretation.

Robust variance estimation. Tanner-Smith and Tipton (2014) argue that most meta-analysts have previously ignored effect size dependencies. They discuss several types of dependencies present in complex data structures such as ours. Three main types of dependencies can arise. The first type is hierarchical dependence, which occurs when multiple studies are run by the same research group. For example, Cole (1985) contributed two experiments to this meta-analysis. The second type is correlated effects, which occur when the

same comparison group is used to estimate multiple effect sizes within one experiment. For example, many experiments had three lineup size conditions, resulting in three effect sizes (A vs B, A vs C, B vs C). The third type of dependency occurs when multiple effect sizes are nested within a study that looks at separate groups of participants using the same procedure or stimuli. For example, Juncu and Fitzgerald (in prep.) tested participants using the same procedure and the same culprits but with fair or unfair lineups.

Although researchers often ignore these dependencies, data processing and selection techniques can be used to deal with them. Some examples include randomly selecting one effect per study or using the average of multiple effect sizes within each study. However, these techniques result in a loss of potentially valuable information. An alternative is using robust variance estimation (Hedges et al., 2010), which simultaneously addresses multiple types of dependency by adjusting the standard error of each effect size and does not require knowledge of the underlying covariance structure among effect sizes.

Weighting method. We used Comprehensive meta-analysis (Version 3; Borenstein et al., 2005) to compute log odds ratios for each comparison. The extracted data was weighted and meta-analyzed in the R package *robumeta*, which uses robust variance estimation. When using robust variance estimation, Tanner-Smith and Tipton (2014) recommend using the weights based on the most prevalent type of dependency. In our final dataset, correlated effects (i.e., the same group is compared with multiple comparison groups) were more frequent than hierarchical effects. Therefore, we used the correlated effect weighting method. We set rho to the default setting (.80). We also conducted sensitivity analyses which showed that changes in rho would not substantially affect our results.

Outliers. We used Comprehensive Meta-Analysis to look for outliers using the random-effects model. Outliers were defined as effect sizes with standardized residuals larger than 1.96. We removed the biggest outlier, one at a time, until all standardized residuals were

below 1.96. Following Higgins' (2008) suggestion, we ran analyses both with and without outliers.

Publication bias. It is always possible that some studies that report small or non-significant effects are systematically under-represented in meta-analyses due to publication bias. For each outcome variable, we conducted publication bias analyses on the aggregated study level effect sizes using the random-effects model in Comprehensive Meta-Analysis. First, we assessed publication bias via visual analyses of funnel plots symmetry and then via trim-and-fill procedure. Although visual inspection depends on subjective judgement to identify bias in the form of plot asymmetry, trim-and-fill applies statistical modeling to identify studies that make the funnel plot asymmetric and imputes placeholders where missing studies should be to make the plot symmetric (Zelinsky & Shadish, 2018). The trim-and-fill method also calculates a new, adjusted summary effect size (Duval & Tweedie, 2000). Arguably, some identification outcomes will be considered more or less important when trying to publish studies about lineup size effects. Although all funnel plots can be found in the Online Supplemental Materials, we will only discuss publication bias analyses for suspect identifications, suspect-guilt discriminability, and choosing.

Sensitivity analyses. Different experimental methodologies were used in the primary studies and these differences could influence the effect of lineup size on identification decisions. As a sensitivity analysis, we repeated the meta-analysis after applying additional methodological exclusion criteria. First, we excluded studies that did not counterbalance fillers across each lineup size. Second, we excluded subgroups that used sequential lineup presentation. Third, we excluded a study that presented numerous lineups to each subject, which might be considered conceptually distinct from studies that present only one or two lineups per subject. These sensitivity analyses indicate whether the effect of lineup size varies with different inclusion/exclusion criteria.

Moderator Analyses

We assessed the impact of moderator variables using metaregression with robust variance estimation.

Lineup size difference. The first covariate we looked at was the size difference between the smaller and the larger lineups. For example, although some researchers compared lineup sizes of 6 versus 9 (a size difference of 3; Cole, 1985), others compared lineups of 6 versus 12 (a size difference of 6; Steblay & Baumann, 2010). We hypothesized that the effects of lineup size will be larger when the size difference is larger. In cases when the size difference varies both within (e.g., Wooten et al., 2020) and between studies, robust variance estimation allows for a covariate's impact to be parsed into between-study effects and within-study effects (Uttal et al., 2013). If a study compared a lineup size of 3 with lineup sizes of 6 and 9, for example, and the differences between 3 and 9 are greater than the differences between 3 and 6 or 6 and 9, that would constitute a within-study effect. Conversely, if the studies that only compared lineup sizes of 6 and 12 yielded larger effects than studies that only compared lineup sizes of 6 and 9, this would be considered a between-study effect. When performed with log odds ratio as the outcome variable, the metaregression produces a coefficient that can be interpreted as a ratio of log odds ratios (Higgins & Green, 2011).

6-member lineups vs small/large lineups. For the main analysis, we computed effect sizes using all comparisons between "smaller" and "larger" lineups within a study. This incorporated as much of the data as possible and tests for relative differences in lineup sizes, but it also limits what can be learned about particular lineup sizes. Therefore, we conducted a moderator analysis with 6-member lineups as reference group to be compared with lineups that were "small" (size = 2-4) or "large" (size = 8-12). This reference group was chosen because six is a mid-range lineup size and was also the most frequently used size in the primary studies. Six is also the most common size of photo lineups in the U.S.

Expected Cost Analysis

Yang and colleagues' (2019) expected cost model incorporates conditional probabilities, costs of each identification outcome, and the prior probability of guilt to estimate the expected cost of a procedure. To compare smaller and larger lineups, conditional probabilities were obtained from weighted summary mean rates associated with each lineup outcome (i.e., suspect identification, filler identification, or lineup rejection). We considered the full range of possible base rates [0-1], and following from Yang et al. (2019) we assessed the impact of different cost ratios (r = 1, 5, 10, 100) to reflect different perspectives on the cost of incriminating an innocent in relation to failing to incriminate the guilty. In addition to performing an Equal Cost Analysis in which filler identifications and rejections are assumed to have equal costs, we performed a Separate Cost Analysis in which filler identifications were assumed to be twice as costly lineup rejections (Table 1).

Results

Within each study, lineup size subgroups were categorized in relation to each other as "smaller" or "larger". Table 3 presents the number of outlying effect sizes removed (Outliers), the number of experiments (*m*), the number of effect sizes after removing outliers (*k*), the weighted means for the two groups compared, the effect size (ES) and 95% confidence intervals (*LL, UL*), the significance test (*t, df, p)*, and the heterogeneity indices (Tau^2, I^2). All outcomes have been analyzed and reported separately for culprit-present and culprit-absent lineups, except discriminability and choosing which are computed using data from both lineup types. The odds ratios were computed such that a value above 1 would indicate that as lineup size increased, the likelihood of a given outcome decreased. In text, all effect sizes reported are accompanied by 95% confidence intervals [*LL*, *UL*]. Forest plots are reported along with the main effect analyses to display the distribution of effect sizes in primary studies. Although we used disaggregated subgroup data in our analyses, individual effect sizes from each experiment

were averaged for the forest plots (Figures 3-5). All outcomes in Table 3 were computed with outliers removed. The exclusion of outliers did not change the significance of any analyses. Analyses of all effect sizes before excluding outliers can be found in the Online Supplemental Materials (Table S1- analysis called "Entire sample"). Publication status (i.e., published or not) did not moderate the effect of lineup size on any of the identification outcomes (Table S2 in Online Supplemental Materials).

Main Effects of Lineup Size

Culprit present lineups. The odds of a hit (guilty suspect identification) were significantly greater when using smaller lineups than when using larger lineups, *OR* = 1.44 [1.32, 1.58]. Trim and Fill analysis suggested no missing studies and therefore no publication bias was detected for this outcome. The effect on hits coincided with lower odds of picking a filler from smaller than from larger lineups, *OR* = 0.65 [0.57, 0.76]. The odds of a rejection were 1.10 [1.03, 1.17] greater from smaller lineups relative to larger lineups (Figure 3).

Culprit absent lineups. The odds of an innocent suspect identification from culprit-absent lineups were significantly greater when using smaller compared with larger lineups, *OR* = 1.82 [1.62, 2.05]. No publication bias was detected for this outcome. Using smaller lineups resulted in a decrease in filler identifications from culprit absent lineups, *OR* = 0.73 [0.69, 0.78], and an increase in lineup rejections, *OR* = 1.15 [1.08, 1.23], in comparison with larger lineups (Figure 4).

Choosing. Choosing represents the overall rate at which lineup members were selected, collapsed across culprit-present and culprit-absent lineups. The odds of choosing a lineup member were significantly lower from smaller lineups than from larger lineups (Figure 5), *OR* = 0.91 [0.87, 0.95]. The trim and fill method suggested that one study is missing and that the point estimate would decrease by 0.01.

Table 3

Main Effects of Lineup Size on Identification Outcomes

					Lineup Size (Weighted Means)		Effect Size & 95% CIs			Test of the Null			Heterogeneity	
Culprit	**Outcome**	**Outliers**	***m***	***k***	**Smaller**	**Larger**	***ES***	***LL***	***UL***	***t***	***df***	***p***	***Tau²***	***I²***
Present	Suspect	8	14	96	.42	.34	1.44	1.32	1.58	8.99	11	.001	0.00	0.10
	Filler	13	12	73	.29	.37	0.65	0.57	0.76	6.47	10	.001	0.02	17.56
	Rejection	5	12	81	.31	.29	1.10	1.03	1.17	3.44	10	.006	0.00	0.00
Absent	Suspect	2	13	97	.12	.07	1.82	1.62	2.05	11.30	11	<.001	0.00	0.00
	Filler	10	13	89	.43	.51	0.73	0.69	0.78	11.00	11	<.001	0.00	0.00
	Rejection	6	13	93	.45	.42	1.15	1.08	1.23	4.90	11	<.001	0.00	0.00
Both	Choosing	6	11	75	.61	.63	0.91	0.87	0.95	4.87	10	<.001	0.00	0.00
	Discriminability	3	13	96	0.99	1.06	-0.03	-0.07	0.01	1.74	11	.109	0.00	0.00

Note. *m* = the number of experiments included, *k* = the number of effect sizes included after removing outliers, *ES* = Effect Size; *CI* = Confidence Interval; *LL* = Lower Limit; *UL* = Upper Limit. For discriminability the weighted means are *d'* scores and the effect size is Hedges' *g*. For all other outcomes, the weighted means are proportions and the effect sizes are odds ratios.

Discriminability. Lineup size did not significantly affect discriminability (Figure 5), Hedges' g = -0.03 [-0.07, 0.01]. No publication bias was detected for this outcome.

Sensitivity Analyses

For sensitivity analyses, we repeated the meta-analysis after applying new exclusion criteria. The first sensitivity analysis was performed after removing studies that did not report counterbalancing lineup members (Bailey, 2011; Cole, 1985; Nosworthy & Lindsay, 1990; Steblay & Baumann, 2010; Wooten et al., 2020). The second was performed after removing conditions with sequential lineup presentation (Meissner et al., 2005; Seale-Carlisle et al., 2019; Steblay & Baumann, 2010). The third was performed after removing the study that presented numerous lineups to each participant in a repeated measures paradigm (Meissner et al., 2005). For most identification outcomes, the exclusions had minimal impact on the effect sizes and significance tests. However, there were two important differences. First, although the odds of choosing from smaller lineups were significantly lower than from larger lineups with the full sample (OR = 0.88 [0.81, 0.96], p = .009), when only simultaneous lineups were included choosing was not significantly affected by lineup size (OR = 0.90 [0.77, 1.05], p = .16). Second, contrary to the full sample, in which lineup size had no significant effects on discriminability, discriminability was significantly improved by increasing lineup size after removing the sequential lineup conditions, Hedges' g = -0.04 [-0.06, -0.03], p < .001. We report the sensitivity analyses in full as Online Supplemental Materials (Table S1).

Moderator Analyses

6-member vs small/large lineups. Moderator analyses for comparisons of 6-member lineups with small and large lineups are reported in Table 4. The effect of lineup size on identifications of the guilty suspect from culprit-present lineups was not moderated by whether 6-member lineups were compared with small or large lineups. In both cases, increasing lineup size led to a comparable reduction in guilty suspect identifications. For

Table 4

Moderator effects of comparing 6-member lineups with small lineups (size = 2-4) and large lineups (size = 8-12)

					Lineup Size (Weighted Means)			Effect Size & 95% CIs			Test of Null			Test of Moderator		
Culprit	**Outcome**	**Comparison**	***m***	***k***	**Small**	**6**	**Large**	***ES***	***LL***	***UL***	***t***	***df***	***p***	***t***	***df***	***p***
Present	Suspect	6 vs Small	5	16	.49	.41		1.39	1.25	1.54	8.76	4	.001	0.05	8	.962
		6 vs Large	10	33		.39	.32	1.38	1.21	1.27	5.54	9	<.001			
	Filler	6 vs Small	4	14	.18	.29		0.49	0.39	0.62	9.82	3	.002	3.37	7	.012
		6 vs Large	9	30		.34	.42	0.72	0.61	0.86	4.29	8	.003			
	Rejection	6 vs Small	4	14	.35	.32		1.16	1.10	1.22	8.83	3	.003	1.86	7	.105
		6 vs Large	9	30		.29	.28	1.06	0.97	1.15	1.50	8	.172			
Absent	Suspect	6 vs Small	5	16	.16	.09		1.87	1.71	2.04	19.80	4	<.001	4.00	8	.004
		6 vs Large	10	30		.10	.06	1.58	1.49	1.68	17.60	9	<.001			
	Filler	6 vs Small	5	16	.32	.45		0.51	0.30	0.87	3.54	4	.024	2.86	8	.021
		6 vs Large	10	30		.45	.52	0.75	0.69	0.82	7.53	9	<.001			
	Rejection	6 vs Small	5	16	.54	.46		1.36	0.97	1.91	2.52	4	.065	2.92	8	.019
		6 vs Large	10	30		.42	.42	1.04	0.90	1.19	0.58	9	.574			
Both	Choosing	6 vs Small	4	14	.56	.62		0.79	0.66	1.06	4.15	3	.025	4.00	7	.005
		6 vs Large	9	27		.64	.64	0.97	0.88	1.07	0.54	8	.604			
	Discriminability	6 vs Small	5	16	1.00	1.13		-0.04	-0.06	-0.03	7.41	4	.002	3.03	8	.016
		6 vs Large	10	30		1.00	1.04	-0.01	-0.03	0.01	1.35	9	.211			

Note. *m* = number of studies, *k* = number of effect sizes, ES = Effect Size; *CI* = Confidence Interval; *LL* = Lower Limit; *UL* = Upper Limit. For discriminability the weighted means are *d'* scores and the effect size is Hedges' *g*. For all other outcomes, the weighted means are proportions and the effect sizes are odds ratio.

culprit-absent lineups, increasing lineup size also reduced innocent suspect identifications regardless of whether 6-member lineups were compared with small or large lineups; however, for innocent suspect identifications, the effect size was larger in the comparison with small lineups (*OR* = 1.97 [1.71, 2.04]) than in the comparison with large lineups (*OR* = 1.58 [1.49, 1.68]), resulting in a significant moderator effect, $t(8) = 4.00$, $p = .004$. This analysis also yielded significant moderator effects in choosing, $t(7) = 4.00$, $p = .005$, and discriminability, $t(8) = 3.03$, $p = .016$. The 6-member lineups had higher choosing rates than the small lineups, *OR* = 0.79 [0.66, 1.06], but further adding fillers to a 6-member had no effect on choosing, *OR* = 0.97 [0.88, 1.07]. Discriminability was also significantly higher in 6-member lineups than in the small lineups, Hedges' g = -0.04 [-0.06, -0.03], but did not significantly improve when increased past six, Hedges' g = -0.01 [-0.03, 0.01].

Lineup size difference. We used metaregression to assess whether the difference between two lineup sizes moderates the effect of lineup size. For this covariate, we separated between- and within-study effects. Table 5 presents the metaregression coefficients, 95% CIs, and significance tests. The only significant associations were for within-study effects. Therefore, these effects were significant only when the experimental procedure was controlled. When between-study confounds were present, the association was not significant.

The size of the increase in hits for smaller compared with larger lineups was positively associated with the difference in lineup size, and the same pattern was observed for innocent suspect identifications. For example, the difference in suspect identifications was less pronounced when comparing a lineup size of 4 with a lineup size of 6 than when comparing a lineup size of 4 with a lineup size of 8. The association between the lineup size difference and effect size magnitude was significant regardless of suspect guilt, but the association was stronger for innocent suspect identifications than for guilty suspect identifications. Lineup size difference also significantly moderated the effect sizes for choosing and discriminability.

Table 5.

Regression Coefficients for Moderator Analysis of Lineup Size Difference

			Effect Size & 95% *CI*s			Test of the Null		
Culprit	**Lineup Choice**	**Effect Type**	**ES**	***LL***	***UL***	***t***	***df***	***p***
Present	Suspect	Within-study	0.08	0.05	0.12	5.84	9	<.001
		Between-study	0.01	-0.14	0.14	0.02	9	.987
	Filler	Within-study	-0.07	-0.13	-0.01	2.54	8	.034
		Between-study	0.05	-0.20	0.30	0.48	8	.653
	Rejection	Within-study	0.03	0.01	0.06	2.77	8	.024
		Between-study	-0.01	-0.14	0.14	0.02	8	.988
Absent	Suspect	Within-study	0.12	0.06	0.18	4.73	9	<.001
		Between-study	0.03	-0.07	0.13	0.72	9	.488
	Filler	Within-study	-0.06	-0.10	-0.02	-3.33	9	.009
		Between-study	-0.07	-0.20	0.07	-1.13	9	.288
	Rejection	Within-study	0.01	-0.02	0.04	0.83	9	.426
		Between-study	0.05	-0.09	0.18	0.78	9	.455
Both	Choosing	Within-study	-0.02	-0.03	-0.01	-2.57	8	.033
		Between-study	0.01	-0.07	0.08	0.25	8	.809
	Discriminability	Within-study	-0.01	-0.02	-0.01	-4.89	9	<.001
		Between-study	-0.01	-0.07	0.05	-0.37	9	.722

Note. *CI* = Confidence Interval; *LL* = Lower Limit; *UL* = Upper Limit.

Expected Cost Analysis

We computed expected cost analyses for larger and smaller lineups using the formulas from Yang and colleagues (2019), with the weighted means in Table 3 as conditional probabilities. Figure 6 depicts the Equal Cost Analysis (cost of filler identification = cost of lineup rejection). If incriminating an innocent suspect is assigned the same cost as failing to incriminate a guilty suspect (r = 1), smaller lineups are superior for all guilty base rates above 44%. However, if incriminating an innocent suspect is considered 10 times more costly (r = 10; i.e., the Blackstone ratio), smaller lineups are only superior for guilty base rates of 90-100%. Expected costs estimated with additional cost ratios are reported in Online Supplimental Materials.

Figure 7 depicts the Separate Cost Analysis (cost of filler identification > cost of lineup rejection). Increases of lineup size result in higher filler identification rates and filler identifications are more costly in the Separate Cost Analysis, so larger lineups perform less favourably with this approach. If r = 1, smaller lineups are less costly across the entire range of possible base rates. If the Blackstone ratio is applied, the cost functions for smaller and larger lineups intersect at a lower probability of guilt than was determined via Equal Cost Analysis, but larger lineups are still superior for the majority of possible guilty base rates. Namely, in Separate Cost Analysis with r = 10, larger lineups result in a lower cost than smaller lineups for all base rates below 75%. Overall, these analyses suggest that smaller lineups would only be less costly if the prior probability of guilt is high or the cost of mistakenly identifying an innocent suspect is low.

Discussion

The meta-analysis revealed that as lineup size increases the likelihood of suspect identifications decreases, and that this effect is consistent across culprit-present and culprit-absent lineups. Increasing lineup size also resulted in a small increase in overall choosing.

Although our main analyses revealed no associations between lineup size and suspect guilt discriminability, moderator analyses showed that increasing a 2-4 member lineup to a 6-member lineup improved discriminability. Smaller lineups were also estimated to be more costly, except when the base rate of suspect guilt was high or the cost of incriminating an innocent suspect was low.

Regardless of guilt, suspect identifications decrease when larger lineups are used, resulting in a trade-off between protecting the innocent and prosecuting the guilty. This trade-off has been found in both experimental and simulation studies on lineup size (Wetmore et al., 2017; Wooten et al., 2020), as well as in experiments with other lineup procedures (Clark, 2012). The effect of lineup size on suspect identifications is consistent with the filler siphoning account (Wells et al., 2015), which states that adding fillers to the lineup reduces the likelihood of a suspect identification by increasing the probability that one of the fillers will best correspond with the witness' memory of the culprit. This account has been extended to the differential filler siphoning theory, which predicts greater siphoning from innocent suspects than from guilty ones, which would increase suspect guilt discriminability (Smith et al., 2017). Although no main effects on discriminability were detected to support this theory in the meta-analysis on the full corpus of studies, the increase in discriminability for 6-member lineups over 2-4 member lineups could be explained by differential filler siphoning.

Why suspect guilt discriminability did not continue to improve when increasing from a 6-member lineup to an 8-12 member lineup is not entirely clear, but may be indicative of diminishing returns associated with increasing lineup size. When adding lineup members, each new filler provides less protection for an innocent suspect than the fillers added prior (Wells et al., 2006). For instance, if a mistaken identification occurs and the lineup is fair, the risk to an innocent suspect is 33.3% in 3-member lineups, 16.7% in 6-member lineups, and 11.1% in a

9-member lineups. Therefore, increasing lineup size from 3 to 6 reduces the absolute risk to the innocent suspect by 16.7%, whereas increasing from 6 to 9 only reduces the risk by 5.6%.

The lower discriminability for small lineups could also be explained by the diagnostic feature detection hypothesis (Wixted & Mickes, 2014; Wixted et al., 2018). According to this account, witnesses benefit from comparing lineup members because it allows them to discriminate between unique features of a lineup member and features shared with other lineup members. If the lineup is small, there may be fewer shared features among the lineup members and this could reduce the effectiveness of the feature detection process. Thus, it is possible that increasing the lineup size refines the search for diagnostic features. Akan et al. (2020), however, note that increasing lineup size would only increase suspect guilt discriminability if the additional fillers highlight additional non-diagnostic features. From this perspective, detection of diagnostic features would depend more upon who is in the lineup than upon how many people are included. Given that only minimal information was reported about the appearance of the fillers in the primary studies, we were unable to assess whether the added fillers would be likely to possess additional features that were nondiagnostic and shared with other lineup fillers. Although this precluded us from directly testing the diagnostic feature detection hypothesis, it is noteworthy that increasing lineup size increased suspect guilt discriminability in the sensitivity analysis that included only simultaneous lineups, which are theorized to facilitate the detection of diagnostic features. However, sequential presentation was uncommon in the lineup size literature, and we were unable to conduct moderator analyses to determine whether the boost in discriminability for larger lineups is contingent upon simultaneous presentation.

Expected Costs

Clark (2012) found a trade-off like the one in this meta-analysis for many lineup procedures and argued that, in addition to identification outcome probabilities, researchers

should consider the utilities associated with each outcome and the prior probability of suspect guilt. If we were to focus exclusively on suspect identification rates or suspect guilt discriminability, we would be making an implicit assumption that the cost of incriminating an innocent suspect is equivalent to the cost of failing to incriminate the culprit (r = 1). Our interpretation would also be contingent upon a 50% likelihood that the suspect is guilty. However, criminal justice systems often prioritize protecting the innocent from false incrimination and guilty base rates are likely to vary across jurisdictions that apply different investigative procedures. To provide a more nuanced exploration of these parameters, we used the expected cost model proposed by Yang et al. (2019). In these analyses, the least costly lineup size depended on both the base rate of guilt and the relative costs assigned to each lineup outcome.

When the cost of incriminating an innocent suspect was set to be no greater than the cost of letting the perpetrator go free (r = 1), the expected cost analysis tended to favour smaller lineups. If r = 1 in the Separate Cost Analysis, smaller lineups were less costly across the full range of base rates (Figure 7). The Separate Cost Analysis assumes filler identifications negatively impact the credibility of witnesses, which results in a disproportionate increase to the expected cost of larger lineups due to their higher conditional probability of filler identifications, relative to smaller lineups (Table 3). However, assuming r = 1 in the Separate Cost Analysis results in a rather peculiar cost structure. Under these specifications, the cost assigned for an innocent suspect identification is effectively the same as the cost assigned to a filler identification (see Table 1). Unlike innocent suspect identifications, which increase the risk of wrongful conviction, filler identifications only have the potential to impeach the eyewitness' credibility. Therefore, this cost structure is at odds with legal systems that prioritize protecting innocents over prosecuting the guilty.

In the Equal Cost Analysis, no consideration is given to the potential cost of a filler identification on witnesses credibility. This reduces the cost assigned to filler identifications, which in turn reduces the expected cost of larger lineups. If $r = 1$ in Equal Cost Analysis, the expected costs of smaller and larger lineups intersect at a base rate of approximately 50%, such that smaller lineups would be less costly in a jurisdiction that mostly investigated guilty suspects and larger lineups would be less costly in a jurisdiction that mostly investigated innocent suspects (Figure 6). In this scenario, the cost of a filler identification (failing to incriminate the perpetrator, $f = 1$) is lower than the total cost of an innocent suspect identification ($f + r = 1 + 1 = 2$). However, when compared with the prevailing view of legal commentators, most of whom argue $r = 10$ (Volokh, 1997), this approach still substantially underestimates the cost of incriminating an innocent suspect.

When the cost of incriminating an innocent suspect is increased to correspond with the Blackstone ratio ($r = 10$), the expected cost is lower for larger lineups than for smaller lineups across the majority of guilty base rates. In the Equal Cost Analysis, larger lineups were the less costly option unless the guilty suspect base rate was 90% or higher. Even in the Separate Cost Analysis, which disproportionately penalizes larger lineups for increasing the conditional probability of a filler identification, the superior option was still larger lineups for all guilty base rates below 75%. Thus, for legal systems designed to achieve the Blackstone ratio, the expected cost analysis indicates that larger lineups are preferred unless at least 75% of suspects are guilty.

The recommendations we can make based on this analysis, however, must be considered in light of its limitations. The results of the expected cost analyses are dependent on the conditional probabilities of lineup identification outcomes from experimental data. Although using weighted means based on meta-analyzed data is better than using single study data (Yang et al., 2019), the conditional probabilities are still the product of an experimental

methodology and their relevance to policy could be jeopardized by the use of methods that do not generalize outside of the laboratory context.

Methodological Considerations

In most of the lineup size studies we reviewed, there was no designated innocent suspect in culprit-absent lineups (Table 2). When no culprit-absent suspect is designated, the false alarm rate is typically estimated using a nominal size correction: the overall choosing rate from culprit-absent lineups is divided by the number of people in the lineup. With this approach, the estimated rate of innocent suspect identifications cannot exceed the inverse of the lineup's nominal size. Thus, the nominal size correction could underestimate the risk to innocent suspects who match the appearance of the culprit more than the average filler. Indeed, in their meta-analysis, Clark et al. (2008) showed that estimating suspect identifications via nominal size correction is likely to underestimated the risk to innocent suspects. Wixted and Wells (2017) note that the nominal size correction might nevertheless give a good approximation for a typical innocent suspect because it would also overestimate the risk to innocent suspects who match the culprit less than the average filler. This logic only holds if the resemblance between an innocent suspect and the actual perpetrator is a random occurrence, which ultimately depends on how the person became a suspect (Wells & Penrod, 2011), but this point is moot for the present purposes. There were so few studies with designated innocent suspects that we had to use the nominal size correction to maintain consistency when calculating the innocent suspect identification rate across studies.

The nominal size correction's distorting effect could be especially pronounced in lineup size comparisons. For most experimental manipulations the same lineup size correction would be applied to each condition. For instance, if 6-member lineups are used in a comparison between biased and unbiased lineup instructions, then the overall false alarm rate for the culprit-absent lineups in both conditions would be divided by 6. But in lineup size research

each condition has its own nominal size, so an overall false alarm rate of 30% would be divided by 3 in a 3-member lineup (10%), by 6 in a 6-member lineup (5%), and so on. If the lineups are all perfectly fair, and identification choices distribute equally across the lineups irrespective of their size, then the nominal size correction would be justified. Conversely, if all of the identifications were concentrated on two of the lineup members, again irrespective of lineup size, applying different corrections for different lineup sizes is harder to justify. At the very least, counterbalancing would be needed to equate lineup fairness across each lineup size as much as possible.

Even with counterbalancing, we recommend designating an innocent suspect in lineup size research. Estimating the false alarm rate via a nominal size correction may be justifiable if studying general impairment variables (Wells & Olson, 2001), which are factors that reduce identification accuracy but do not disproportionately affect any particular lineup member (e.g., delay between witnessed event and lineup). But lineup size is not a general impairment variable. The meta-analysis clearly showed that lineup size affects the likelihood of a suspect identification, which makes it a suspect bias variable (Wells & Olson, 2001). For these types of variables, estimation only adds noise to the false alarm rate. Consider the example of another suspect bias variable, clothing bias. If one lineup member wears the same bright red shirt worn by the culprit and all other lineup members wear blue shirts, the biasing effect of the red shirt would be distorted if estimation was used and identifications of the blue shirted lineup members were incorporated into the false alarm rate. The best way to clearly measure these kinds of suspect biases is to designate an innocent suspect in the lineups.

A related methodological consideration is that most studies use culprit-matched fillers instead of suspect-matched fillers. Most policies specify that fillers should be matched to the appearance of the suspect (Fitzgerald et al., 2021), which means that fillers in culprit-absent lineups would be matched to the innocent suspect. However, in the lineup size literature, the

typical practice has been to match fillers to the culprit in culprit-present lineups and then use the same culprit-matched fillers in the culprit-absent lineups. In addition to the questionable ecological validity of matching culprit-absent lineup fillers to the culprit, whose appearance would not be known if the culprit is absent from the lineup, Clark and Tunnicliff (2001) reviewed 25 studies and found that using the same fillers in culprit-present and culprit-absent lineups results in underestimations of innocent suspect identifications and overestimations of filler identifications. The same-fillers method could have a disproportionate effect on innocent suspect identifications in smaller lineups, which have less room for error (Wooten et al., 2020). To better simulate suspect-matched lineups in experiments, Oriet and Fitzgerald (2018) proposed the single-lineup paradigm. Instead of using one culprit during encoding and two suspects in the lineups, in this paradigm witnesses are randomly assigned to see one of two visually similar (i.e., matching the same description) culprits during encoding and then see the same lineup with the same suspect at test. With this approach, all lineups contain the same suspect-matched fillers and the guilt of the suspect depends on the culprit that each participant was assigned to observe. This procedure is especially useful when examining lineup composition manipulations such as lineup size. To better assess the effect of lineup size, future studies should use suspect-matched lineups via the single-lineup paradigm.

Directions for Future Research

There are many additional questions relevant to understanding lineup size, which is likely to be moderated by factors such as the quality of the fillers, whether the lineup is presented simultaneously or sequentially, and the confidence of the eyewitness. Although the literature on lineup size is still developing and not yet sufficient to incorporate these factors into the meta-analysis, we hope this section will motivate future lines of research.

Filler similarity. More research is needed to understand the relationship between the filler quantity and quality. The WITNESS simulations indicated that the number of lineup

members would be less influential if the fillers are poor matches to the suspect (Wetmore et al., 2017). There is limited experimental data that can speak to this, and we were unable to include filler quality as a moderator variable in the meta-analysis. However, Clark (2012) found that filler similarity produced a trade-off similar to that observed in the current meta-analysis, and when Yang and colleagues (2019) estimated the expected costs of lower and higher similarity fillers, they produced figures that look strikingly similar to ones reported here for smaller and larger lineups, respectively. Thus, increasing lineup size may affect lineup outcomes through a similar mechanism as increasing filler similarity. For instance, increasing lineup size may indirectly increase the number of fillers who resemble the culprit. Lucas et al. (2020) recently showed that successive increases in the number of fillers who strongly resembled the culprit led to successive increases in filler identifications and successive decreases in suspect identifications. It may be that successive increases in lineup size increase the probability that there will be at least one filler who matches the eyewitness' memory enough to draw them away from the suspect.

Presentation mode. Another important variable to consider when examining the effect of lineup size is presentation mode. Only three of the included studies manipulated lineup size in sequential lineups. Steblay and Baumann (2010) found that increasing the size of sequential lineups from 6 to 12 results in a decrease in diagnosticity. This effect was not found when simultaneous lineups were used. By contrast, Seale-Carlisle et al. (2019) observed higher, although non-significantly higher, suspect guilt discriminability for 9-member sequential lineups compared with 6-member sequential lineups. Meissner et al. (2005) reported a decrease in diagnosticity as more fillers are added to a lineup, however they did not detect an interaction with the mode of lineup presentation. These conflicting findings may have been a consequence of how the sequential lineups were operationalized in the experiments, as the methodologies were all quite different and there are many sequential lineup features that could interact with

lineup size, such as the pre-lineup instructions, the suspect's position, and the stopping rule (Horry et al., 2020; Steblay et al., 2011).

Confidence. Eyewitness confidence is highly influential for the investigative process as well as for courtroom assessments of eyewitness reliability (Lampinen et al., 2019). In Wooten and colleagues' (2020) study, when only highly confident witnesses were considered (confidence above 91%), participants who chose a suspect from a 9- or 12-member lineup were significantly more accurate than participants who chose a suspect from a 3- member lineup. Therefore, further research should be conducted to examine effects of lineup size on the relation between confidence and accuracy.

Policy

Lineup size is discussed in guidelines all over the world, and the current findings suggest that many of these policies would be improved by increasing the minimum number of lineup members. In a review of lineup policies (Fitzgerald et al., 2021), the most frequently recommended minimum lineup size was three. Field data with eyewitnesses in real criminal investigations indicate that approximately 1/3 of the people identified at a lineup procedure are known-innocent fillers (Wells et al., 2020). Given this propensity of eyewitnesses for mistaken identification, an innocent suspect faces a considerable risk in a lineup that includes only two fillers. If small lineups clearly benefited eyewitness performance, perhaps the risk could be justified. However, we found no such benefits in this meta-analysis. Small lineups increased identifications of guilty and innocent suspects alike, and 2-4 member lineups had poorer suspect guilt discriminability compared with 6-member lineups. Although the difference in discriminability was modest, it supports the principle that a lineup with only three members poses an undue risk to potentially innocent suspects.

The question of whether lineups should be even larger than six is more complicated. Lineups with 8-12 members did not improve suspect guilt discriminability over 6-member

lineups. Adding fillers to a 6-member led to fewer identifications of both guilty and innocent suspects. Making policy in this context is challenging, particularly when the base rate of suspect guilt is unknown. Wixted et al. (2016) estimated that only a third of lineups from a Houston field experiment included a guilty suspect. Further evidence of a low base rate in the U.S. comes from sexual assault cases referred to the FBI for DNA testing. Neufeld and Scheck (1996) reviewed FBI data from ~10,000 cases and noted that the DNA tests excluded the primary suspect in 25% of the ~8000 cases with a conclusive result. In other words, the base rate of guilt was 75%. Although this a much higher estimate than in the Houston field study, Neufeld and Scheck explain that sexual assault cases were normally referred to the FBI for DNA testing *after* the suspect was identified by an eyewitness. Thus, 75% would be more akin to the posterior probability of guilt, given that an eyewitness has identified the suspect. Presuming that an identification of the suspect would increase the likelihood of guilt, the prior probability that a suspect is guilty (i.e., before conducting the lineup) would be much lower.

Although there is substantial uncertainty in these estimates of the guilty base rate in U.S. lineups, they nonetheless raise the prospect that larger lineups would be beneficial. If the U.S. aims to strike a balance consistent with the Blackstone ratio, the expected cost analyses indicate that larger lineups would be less costly for all guilty base rates < 75%. Even if the exact base rate is unknown, the estimates from the Houston field study and the FBI sexual assault cases suggest it would indeed be far lower than 75%. Our findings suggest that when it is common for lineup suspects to be innocent, the best policy would be to err with lineups on the larger side. But this might be less crucial if the U.S. adopts the recent recommendation from Wells et al. (2020) to require reasonable suspicion before putting a suspect in a lineup.

If the required minimum lineup size is increased, the new policy would need to be supported with resources to be effective. Without a database of high quality filler images, a stringent minimum size policy that requires large lineups could encourage the inclusion of

implausible fillers and defeat the purpose of the policy. Even increasing the minimum size from three could be a problem for some jurisdictions, particularly given that a lax minimum lineup size often coincides with a policy preference for live lineups (Fitzgerald et al., 2021). Therefore, in some cases, increasing lineup size might be contingent upon a change from live to nonlive procedures. Live lineups do not improve eyewitness accuracy (Rubínová et al., 2020), and they cause an array of practical issues (Fitzgerald et al., 2018), so policymakers should not let a preference for live identification procedures deter them from increasing the minimum lineup size. Resources are nonetheless needed in jurisdictions that do allow photo lineups. Steblay and Wells (2020) found that 6-member photo lineups frequently included implausible fillers and are suspect-biased. Thus, without a large database of filler images, requiring large lineups could backfire. The ideal resource would be something like the VIPER database in the UK, which contains tens of thousands of filler images that are quality assured for standardization in background, quality, and general formatting.

Conclusion

The current meta-analysis revealed that increasing lineup size results in a decrease in both correct and mistaken suspect identifications. Previous researchers suggested that because increasing lineup size does not improve suspect guilt discriminability, smaller lineups might be preferred. However, the expected cost analysis showed that this depends on the costs associated with different identification outcomes, as well as on the prior probability of guilt. Smaller lineups were superior only if high guilty base rates were assumed, or if innocent suspect identifications and filler identifications were presumed to not be costly. Increasing small lineups (2-4 members) to up to 6 members increased suspect guilt discriminability, which supports a policy of including at least 5 fillers with the lineup. Additional research is needed to further refine our understanding of lineup size effects.

References

*denotes the study was included in the meta-analysis

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

*Bailey, M. (2011). *Minimizing suspect misidentification: Using eight photos instead of six* (Unpublished Master's thesis). Central Connecticut State University, New Britain Connecticut.

Behrman, B. W., & Richards, R. E. (2005). Suspect/foil identification in actual crimes and in the laboratory: A reality monitoring analysis. *Law and Human Behavior, 29*(3), 279–301. https://doi.org/10.1007/s10979-005-3617-y

Blackstone, W. (1769). Commentaries on the Laws of England, Vol. II, Book IV. New York, NY: Duyckinck, Long, Collins & Hannay, and Collins & Co.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2005). Comprehensive Meta Analysis: Version 2 [Computer software]. Englewood, NJ: Biostat.

*Brewer, N., Caon, A., Todd, C., & Weber, N. (2006). Eyewitness identification accuracy and response latency. *Law and Human Behavior, 30*(1), 31–50. https://doi.org/10.1007/s10979-006-9002-7

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

Clark, S. E. (2012). Costs and Benefits of Eyewitness Identification Reform. *Perspectives on Psychological Science, 7*(3), 238–259. https://doi.org/10.1177/1745691612439584

Clark, S. E., & Godfrey, R. D. (2009). Eyewitness identification evidence and innocence risk. *Psychonomic Bulletin & Review, 16*(1), 22–42. https://doi.org/10.3758/PBR.16.1.22

Clark, S. E., Howell, R. T., & Davey, S. L. (2008). Regularities in eyewitness identification. *Law and Human Behavior*, *32*(3), 187–218. https://doi.org/10.1007/s10979-006-9082-4

Clark, S. E., & Tunnicliff, J. L. (2001). Selecting lineup foils in eyewitness identification experiments: Experimental control and real-world simulation. *Law and Human Behavior*, *25*(3), 199–216. https://doi.org/10.1023/A:1010753809988

*Cole, S. P. (1985). *The effects of lineup size, similarity of lineup members and target distinctiveness upon eyewitness identification.* [Unpublished doctoral dissertation]. Emory University, Atlanta, GA.

Cutler, B. L., Penrod, S. D., O'Rourke, T. E. & Martens, T. K. (1986). Unconfounding the effects of context cues on eyewitness identification accuracy. *Social Behavior, 1*, 113–134.

Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987). Improving the reliability of eyewitness identification: Putting context into context. *Journal of Applied Psychology*, *72*(4), 629–637. https://doi.org/10.1037/0021-9010.72.4.629

Devlin, Lord P. (1976). *Report to the Secretary of State for the Home Department of the Departmental Committee on Evidence of Identification in Criminal Cases*. London: Her Majesty's Stationery Office.

Duval, S., & Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American statistical association*, 95(449), 89-98.

Dysart, J. E., & Lindsay, R. C. L. (2007). Show-up identifications: Suggestive technique or reliable method? In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology, Vol. 2. Memory for people* (p. 137–153). Lawrence Erlbaum Associates Publishers.

FPT Federal/Provincial/Territorial Heads of Prosecutions Subcommittee on the Prevention of Wrongful Convictions. (2018). *Innocence at stake: The need for continued vigilance to prevent wrongful convictions in Canada.* Public Prosecution Service of Canada. https://www.ppsc-sppc.gc.ca/eng/pub/is-ip/index.html

Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis. *Psychology, Public Policy, and Law, 19,* 151-164. http://dx.doi.org/10.1037/a0030618

Fitzgerald, R. J., Price, H. L., & Valentine, T. (2018). Eyewitness identification: Live, photo, and video lineups. *Psychology, Public Policy, and Law*, *24*, 307-325. http://dx.doi.org/10.1037/law0000164

Fitzgerald, R. J., Rubínová, E., & Juncu, S. (2021). Eyewitness identification around the world. In A. M. Smith, M. P. Toglia, & J. M. Lampinen (Eds.), *Methods, measures, and theories in eyewitness identification tasks* (pp. 294-322). Taylor and Francis. https://doi.org/10.4324/9781003138105-16

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Higgins, J. P. T. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology*, *37*(5), 1158–1160. https://doi.org/10.1093/ije/dyn204

Higgins, J. P. T, & Green, S. (2011). Cochrane handbook for systematic reviews of interventions (Version 5.1.0). The Cochrane Collaboration.

Home Office. (1969). Identification Parades. Home Office Circular 9/1969. In P. A. Devlin (Ed.), *Report to the Secretary of State for the Home Department on the Departmental*

Committee on Evidence of Identification in Criminal Cases (pp. 158–161). London, UK: Her Majesty's Stationery Office.

Home Office. (2017). *Police and Criminal Evidence Act 1984 (PACE) Code D*. Retrieved from https://www.gov.uk/government/publications/pace-code-d-2017

Horry, R., Fitzgerald, R. J., & Mansour, J. K. (2020). "Only your first yes will count": The impact of prelineup instructions on sequential lineup decisions. *Journal of Experimental Psychology: Applied.* Advance online publication. https://doi.org/10.1037/xap0000337

*Juncu, S., & Fitzgerald, R. J. (in prep). *The effect of lineup size and similarity on eyewitness identification decisions*. Unpublished manuscript.

Lampinen, J. M., Smith, A. M., & Wells, G. L. (2019). Four utilities in eyewitness identification practice: Dissociations between receiver operating characteristic (ROC) analysis and expected utility analysis. *Law and Human Behavior, 43*(1), 26–44. https://doi.org/10.1037/lhb0000309

Levi, A. M., & Lindsay, R. C. L. (2001). Lineup and photo spread procedures: Issues concerning policy recommendations. *Psychology, Public Policy, and Law*, *7*(4), 776–790. https://doi.org/10.1037/1076-8971.7.4.776

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

Lucas, C. A., Brewer, N., & Palmer, M. A. (2020). Eyewitness identification: The complex issue of suspect-filler similarity. *Psychology, Public Policy, and Law*. Advance online publication. http://dx.doi.org/10.1037/law0000243

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: a user's guide*. Cambridge UK: Cambridge, UP.

Malpass, R. S. (1981). Effective size and defendant bias in eyewitness identification lineups. *Law and Human Behavior, 5*, 299-309. https://doi.org/10.1007/BF01044945

Malpass, R. S. (2006). A policy evaluation of simultaneous and sequential lineups. *Psychology, Public Policy, and Law*, *12*(4), 394–418. https://doi.org/10.1037/1076-8971.12.4.394

*Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition*, *33*(5), 783–792. https://doi.org/10.3758/BF03193074

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

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

Mickes, L., Moreland, M. B., Clark, S. E., & Wixted, J. T. (2014). Missing the information needed to perform ROC analysis? Then compute d′, not the diagnosticity ratio. *Journal of Applied Research in Memory and Cognition*, *3*(2), 58–62. https://doi.org/10.1016/j.jarmac.2014.04.007

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine*, 6(7), e1000097.

Neufeld, P., & Scheck, B. C. (1996). Convicted by juries, exonerated by science: Case studies in the use of DNA evidence to establish innocence after trial. Commentary. Crime

Magazine: An Encyclopedia of Crime. Retrieved from http://www.crimemagazine.com/dna.htm

Neuschatz, J. S., Wetmore, S. A., Key, K. N., Cash, D. K., Gronlund, S. D., & Goodsell, C. A. (2016). A comprehensive evaluation of showups. In B. Bornstein & M. K. Miller (Eds.), Advances in psychology and law (pp. 43–69). Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-29406-3_2

*Nosworthy, G. J., & Lindsay, R. C. (1990). Does nominal lineup size matter? *Journal of Applied Psychology*, *75*(3), 358–361. https://doi.org/10.1037/0021-9010.75.3.358

Oriet, C., & Fitzgerald, R. J. (2018). The single lineup paradigm: A new way to manipulate target presence in eyewitness identification experiments. *Law and Human Behavior*, *42*(1), 1–12. https://doi.org/10.1037/lhb0000272

Palmer, M. A., & Brewer, N. (2012). Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law and Human Behavior*, 36(3), 247.

Pike, G., Brace, N., & Kynan, S. (2002). The visual identification of suspects: procedures and practice. London: Home Office, Policing and Reducing Crime Unit, Research, Development and Statistics Directorate

Police Executive Research Forum [PERF]. (2013). *A national survey of eyewitness identification processes in law enforcement agencies.* Washington, DC: Author. Retrieved from http://policeforum.org/library/eyewitness-identification/NIJEyewitnessReport.pdf

*Pozzulo, J. D., Dempsey, J. L., & Wells, K. (2010). Does lineup size matter with child witnesses. *Journal of Police and Criminal Psychology*, *25*(1), 22–26. https://doi.org/10.1007/s11896-009-9055-x

Rubínová, E., Fitzgerald, R. J., Juncu, S., Ribbers, E., Hope, L., & Sauer, J. D. (2020). Live presentation for eyewitness identification is not superior to photo or video presentation. *Journal of Applied Research in Memory and Cognition*. Advance online publication.

*Seale-Carlisle, T. M., Wetmore, S. A., Flowe, H. D., & Mickes, L. (2019). Designing police lineups to maximize memory performance. *Journal of Experimental Psychology: Applied, 25*(3), 410–430. https://doi.org/10.1037/xap0000222

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.

Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2017). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior*, *41*(2), 127–145. https://doi.org/10.1037/lhb0000219

Smith, A. M., Yang, Y., & Wells, G. L. (2020). Distinguishing between investigator discriminability and eyewitness discriminability: A method for creating full receiver operating characteristic curves of lineup identification performance. *Perspectives on Psychological Science*, *15*, 589-607. https://doi.org/10.1177/1745691620902426

South African Police Service. (2007). *National instruction: Identification parades*. Pretoria: Commissioner of the South African Police Service

*Steblay, N. & Baumann, A. (2010). *Lineup size and identification accuracy* [Unpublished manuscript]. Psychology Department, Augsburg University.

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, *27*(5), 523–540. https://doi.org/10.1023/A:1025438223608

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*, 99–139. https://doi.org/10.1037/a0021650

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

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*(1), 13–30. https://doi.org/10.1002/jrsm.1091

Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement*. US Department of Justice, Office of Justice Programs, National Institute of Justice.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352–402. https://doi.org/10.1037/a0028446

Volokh, A. (1997) *N Guilty Men.* University of Pennsylvania Law Review, Vol. 146, No. 2, Available at SSRN: https://ssrn.com/abstract=11412

*Wagenaar, W. A., & Veefkind, N. (1992). *Comparison of one-person and many-person lineups: A warning against unsafe practices.* In F. Lösel, D. Bender, & T. Bliesener (Eds.), *Psychology and law: International perspectives* (p. 275–285). Walter De Gruyter.

Wells, G. L. (2001). Police lineups: Data, theory, and policy. *Psychology, Public Policy, and Law, 7*(4), 791–801. https://doi.org/10.1037/1076-8971.7.4.791

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior*, *44*(1), 3–36. https://doi.org/10.1037/lhb0000359

Wells, G. L., Leippe, M. R., & Ostrom, T. M. (1979). Guidelines for empirically assessing the fairness of a lineup. *Law and Human Behavior*, *3*(4), 285–293. https://doi.org/10.1007/BF01039807

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence. *Psychological Science in the Public Interest*, *7*(2), 45–75. https://doi.org/10.1111/j.1529-1006.2006.00027.x

Wells, G. L., & Olson, E. A. (2001). The other-race effect in eyewitness identification: What do we do about it? *Psychology, Public Policy, and Law*, *7*(1), 230–246. https://doi.org/10.1037/1076-8971.7.1.230

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. 237–256). John Wiley and Sons.

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

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, *22*(6), 603–647. https://doi.org/10.1023/A:1025750605807

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

Wells, G. L., & Turtle, J. W. (1986). Eyewitness identification: The importance of lineup models. *Psychological Bulletin*, *99*(3), 320–329. https://doi.org/10.1037/0033-2909.99.3.320

Wells, G. L., Yang, Y., & Smalarz, L. (2015). Eyewitness identification: Bayesian information gain, base-rate effect equivalency curves, and reasonable suspicion. *Law and Human Behavior*, *39*(2), 99–122. https://doi.org/10.1037/lhb0000125

Wetmore, S. A., McAdoo, R. M., Gronlund, S. D., & Neuschatz, J. S. (2017). The impact of fillers on lineup performance. *Cognitive Research: Principles and Implications*, *2*(1), 48. https://doi.org/10.1186/s41235-017-0084-1

Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition*, *4*(1), 8–14. https://doi.org/10.1016/j.jarmac.2014.07.003

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

Wixted, J. T., & Mickes, L. (2015). ROC analysis measures objective discriminability for any eyewitness identification procedure. *Journal of Applied Research in Memory and Cognition*, *4*, 329-334. https://doi.org/10.1016/j.jarmac.2015.08.007

Wixted, J. T., & Mickes, L. (2018). Theoretical vs. empirical discriminability: the application of ROC methods to eyewitness identification. *Cognitive Research: Principles and Implications*, *3*, 9. https://doi.org/10.1186/s41235-018-0093-8

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences*, *113*(2), 304–309. https://doi.org/10.1073/pnas.1516814112

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

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

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

Yang, Y., Smalarz, L., Moody, S. A., Cabell, J. J., & Copp, C. J. (2019). An expected cost model of eyewitness identification. *Law and Human Behavior*, *43*(3), 205–219. https://doi.org/10.1037/lhb0000331

Yates, S. Q. (2017). Memorandum for heads of department law enforcement components, all department prosecutors. Washington, DC: Office of the Deputy Attorney General, U.S. Department of Justice. Retrieve from https://www.justice.gov/archives/opa/press-release/file/923201/download

Zelinsky, N. A., & Shadish, W. (2018). A demonstration of how to do a meta-analysis that combines single-case designs with between-groups experiments: The effects of choice making on challenging behaviors performed by people with disabilities. *Developmental Neurorehabilitation*, 21(4), 266-278.

Figure 1

Policies on Recommended Minimum Number of Lineup Members

Note. Data refer to the sample of policies reviewed by Fitzgerald, Rubínová, & Juncu (2021).

Figure 2

Literature Search and Study Selection

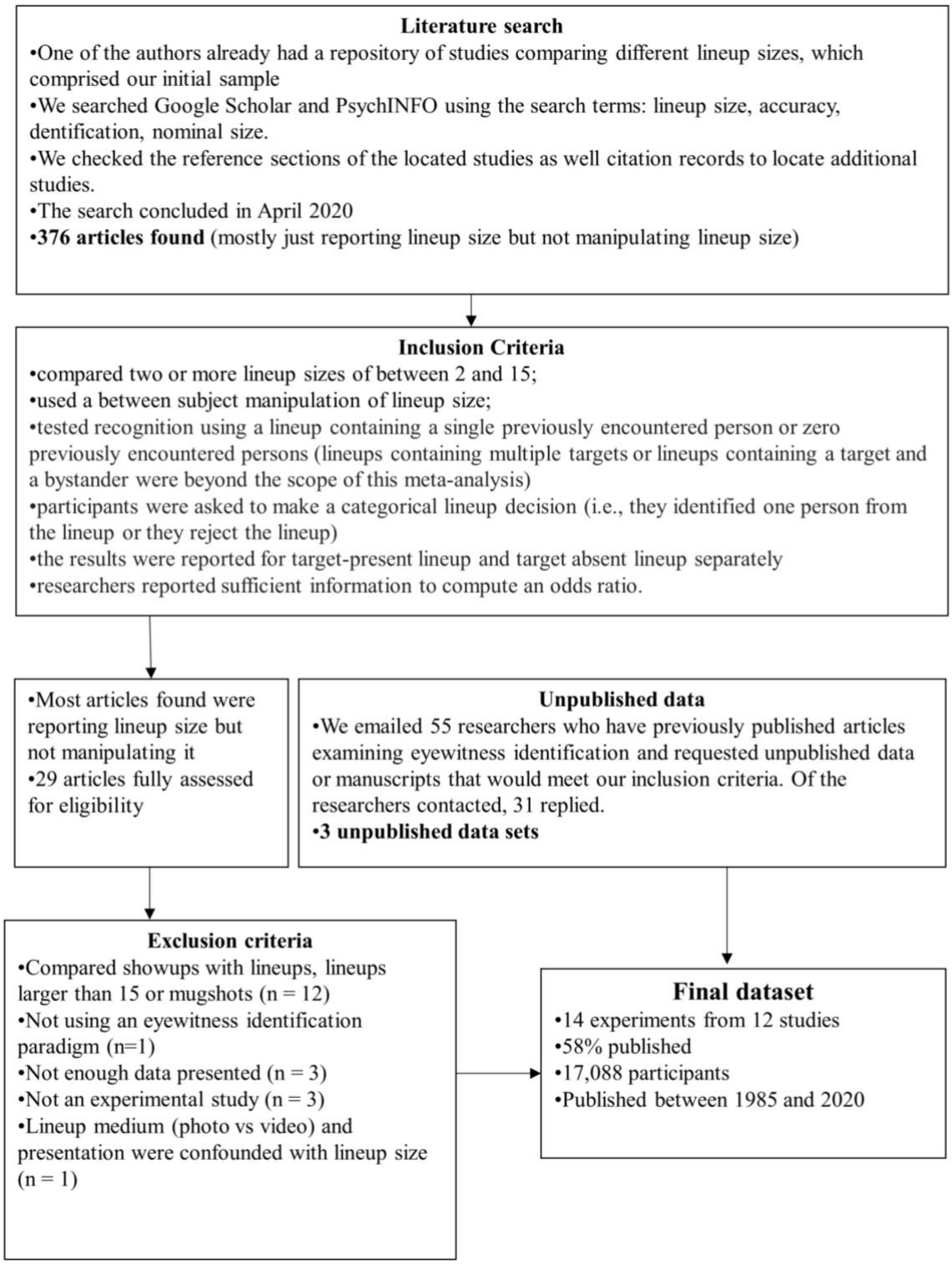
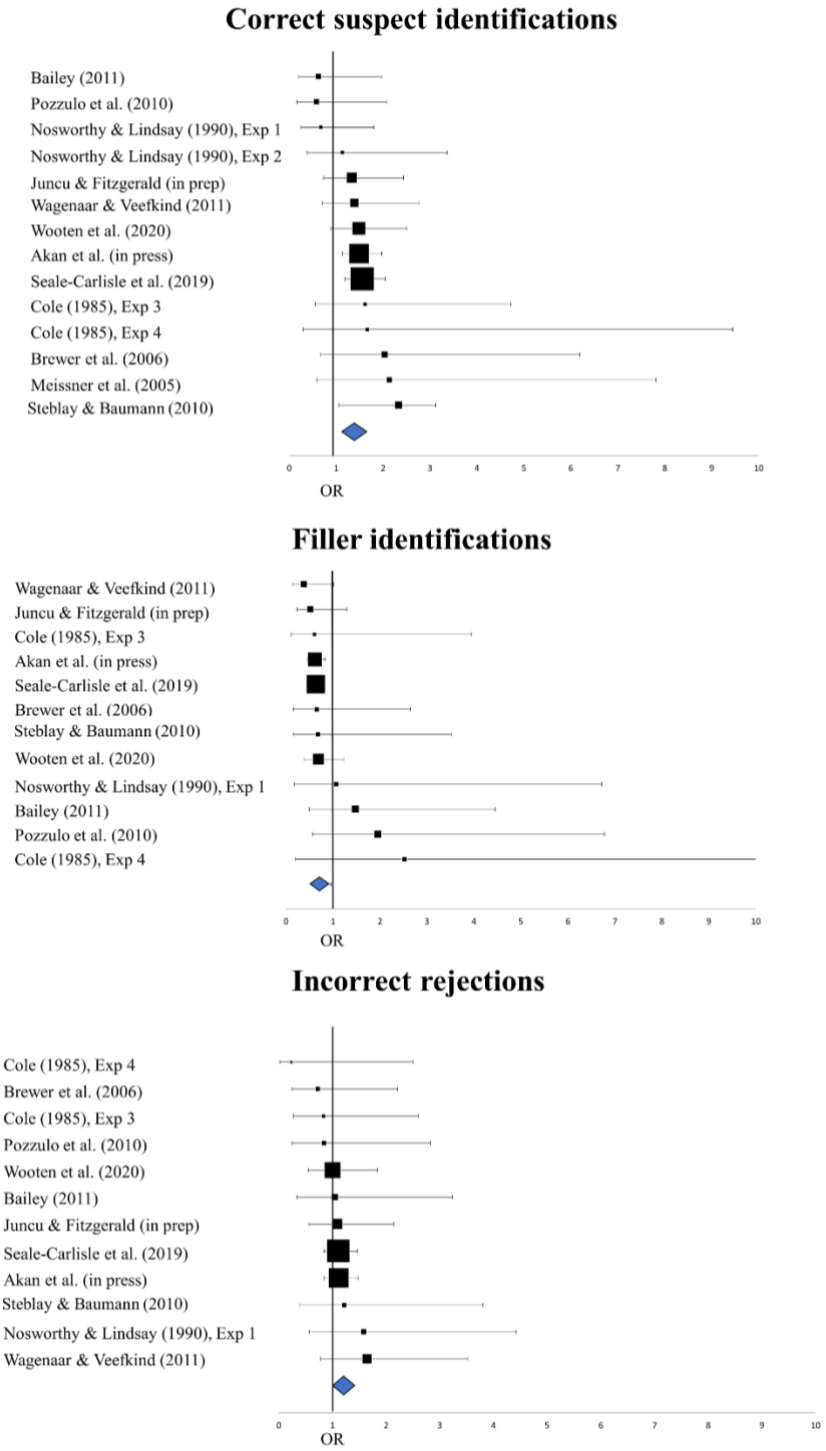


Figure 3

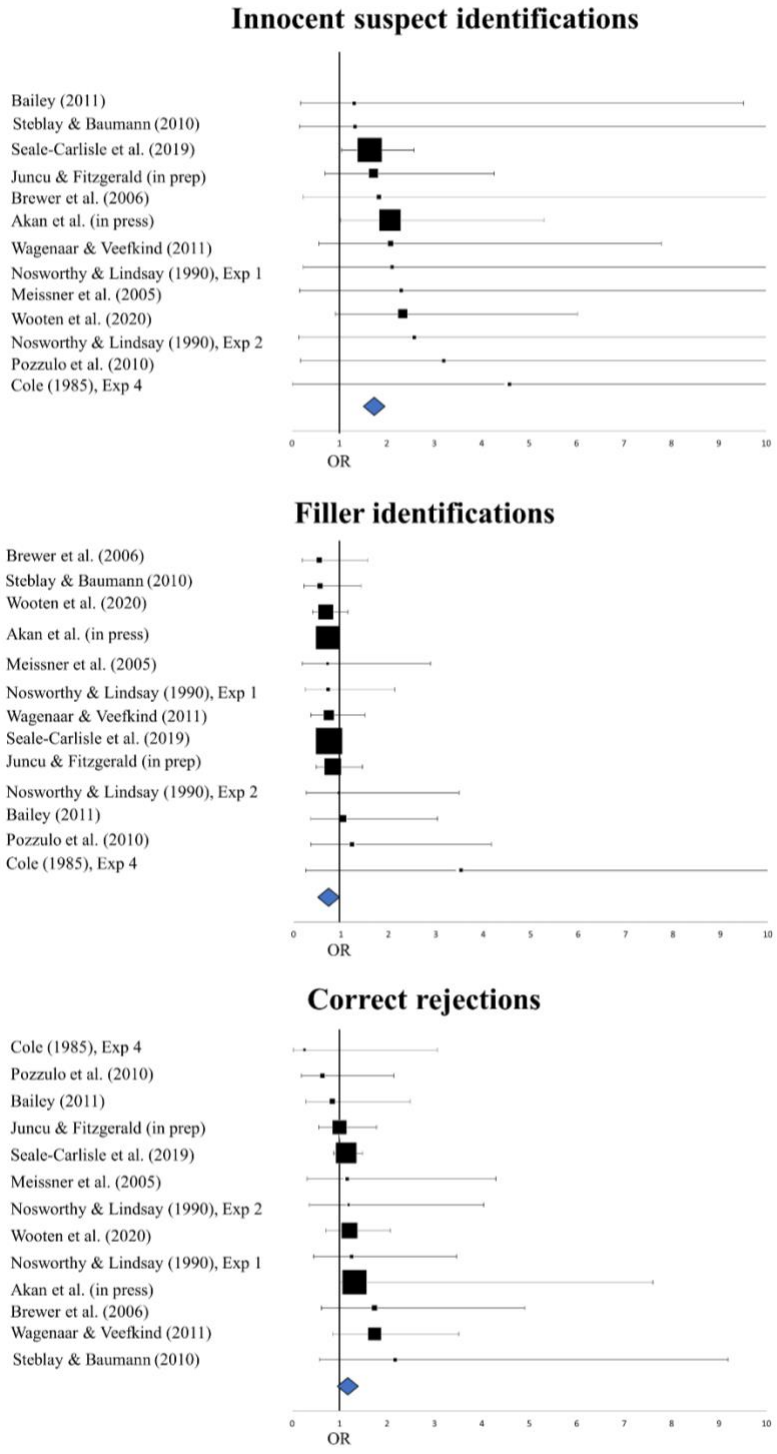
Effect of Lineup Size on Culprit-Present Lineup Outcomes



Note. The forest plots depict Odds Ratios (OR), with ORs above 1 denoting a decrease in the outcome as lineup size is increased. Box sizes are proportional to study weights. Horizontal lines are 95% CIs. Diamonds are summary effect sizes.

Figure 4

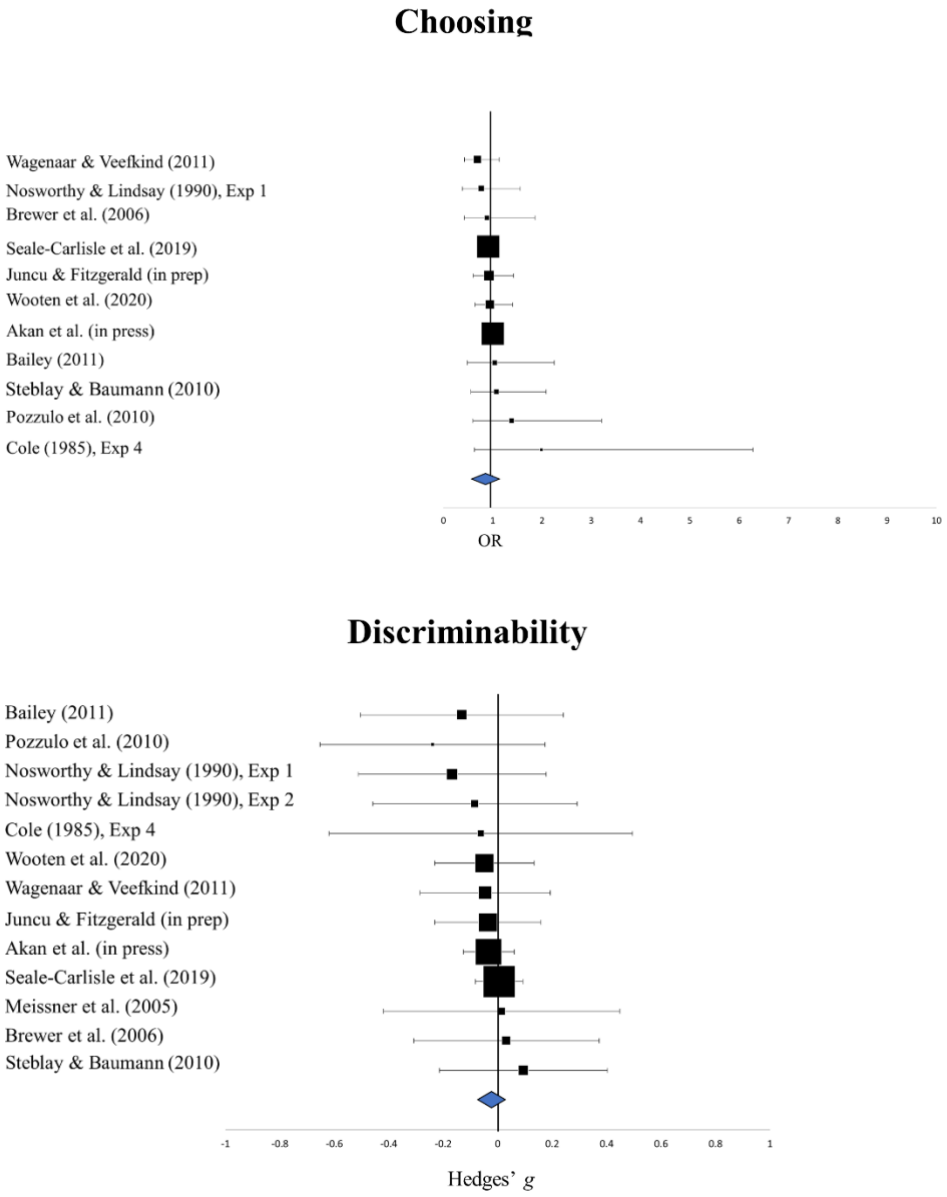
Effect of Lineup Size on Culprit-Absent Lineup Outcomes



Note. The forest plots depict Odds Ratios (OR), with ORs above 1 denoting a decrease in the outcome as lineup size is increased. Box sizes are proportional to study weights. Horizontal lines are 95% CIs. Diamonds are summary effect sizes.

Figure 5

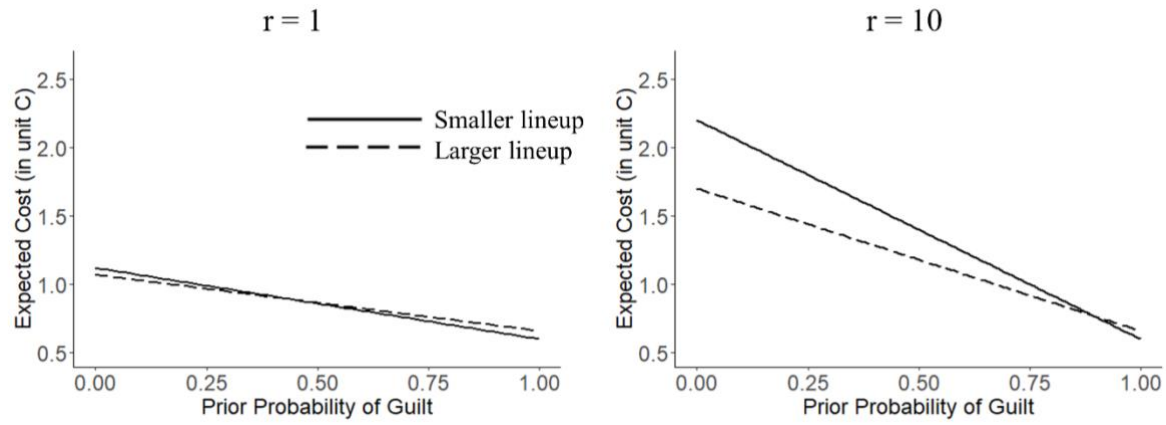
Effect of Lineup Size on the Choosing Rate and Suspect Guilt Discriminability



Note. The forest plots depict Odds Ratios (OR) for choosing and Hedges' *g* for discriminability. ORs above 1 denote a decrease in choosing as lineup size is increased, and *g* values greater than 0 denote a decrease in discriminability as lineup size increases. Box sizes are proportional to study weights. Horizontal lines are 95% CIs. Diamonds are summary effect sizes.

Figure 6

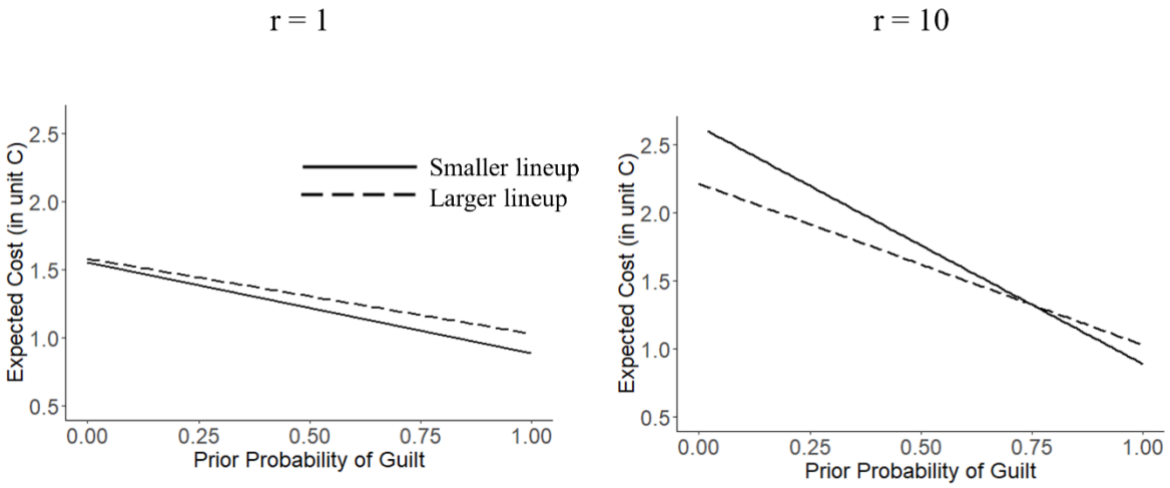
Equal Cost Analysis



Note. The expected costs of increasing lineup size, assuming equal costs for filler identifications and rejections (Equal Cost Analysis). The two slanted lines represent the expected costs of smaller lineups (solid line) and larger lineups (dashed line). *r* represents the ratio between the cost of incriminating an innocent suspect and the cost of failing to incriminate a guilty suspect.

Figure 7

Separate Cost Analysis



Note. The expected costs of increasing lineup size, assuming the cost of filler identifications and is double the cost of rejections (Separate Cost Analysis). The two slanted lines represent the expected costs of smaller lineups (solid line) and larger lineups (dashed line). *r* represents the ratio between the cost of incriminating an innocent suspect and the cost of failing to incriminate a guilty suspect.