The Grain-Size Lineup: A Test of a Novel Eyewitness Identification Procedure

Ruth Horry[a], Neil Brewer[b], and Nathan Weber[b]

[a]Flinders University, Adelaide, Australia; and Swansea University, Swansea, UK

[b]Flinders University, Adelaide, Australia

Accepted for publication in *Law and Human Behavior*. This article may not exactly replicate the final version published in the APA journal. It is not the copy of record.

Author note

Ruth Horry, School of Psychology, Flinders University, and Department of Psychology, Swansea University; Neil Brewer, School of Psychology, Flinders University; Nathan Weber, School of Psychology, Flinders University.

This research was supported by funding from the School of Psychology, Flinders University, and an ARC Linkage International Social Sciences Collaboration grant awarded to N. Brewer et al. We thank Alecia Alinejad for her assistance with data collection, and Carmen Lucas for her assistance with stimulus preparation.

Correspondence concerning this article should be addressed to Ruth Horry, Department of Psychology, College of Human and Health Sciences, Vivian Tower, Swansea University, Swansea, SA2 8PP. Email: r.horry@swansea.ac.uk.

Abstract

When making a memorial judgment, respondents can regulate their accuracy by adjusting the precision, or grain size, of their responses. In many circumstances, coarse-grained responses are less informative, but more likely to be accurate, than fine-grained responses. This paper describes a novel eyewitness identification procedure, the grain size lineup, in which participants eliminated any number of individuals from the lineup, creating a choice set of variable size. A decision was considered to be fine-grained if no more than one individual was left in the choice set or coarse-grained if more than one individual was left in the choice set. Participants (N = 384) watched two high-quality or low-quality videotaped mock crimes and then completed four standard simultaneous lineups or four grain size lineups (two target-present and two target-absent). There was some evidence of strategic regulation of grain size, as the most difficult lineup was associated with a greater proportion of coarse-grained responses than the other lineups. However, the grain-size lineup did not outperform the standard simultaneous lineup. Fine-grained suspect identifications were no more diagnostic than suspect identifications from standard lineups, while coarse-grained suspect identifications carried little probative value. Participants were generally reluctant to provide coarse-grained responses, which may have hampered the utility of the procedure. For a grain-size approach to be useful, participants may need to be trained or instructed to use the coarse-grained option effectively.

Keywords: Eyewitness identification, metacognition, grain size, confidence.

The Grain-Size Lineup: A Test of a Novel Eyewitness Identification Procedure

When reporting information, people generally aspire to be accurate and informative. Yet, under conditions of uncertainty, the goals of accuracy and informativeness compete, forcing individuals to seek a compromise between the two (Koriat & Goldsmith, 1996a). One option for balancing informativeness and accuracy is to regulate the precision, or grain size, of a response; fine-grained responses are more informative, but less likely to be accurate, than coarse-grained responses (Goldsmith, Koriat, & Weinberg-Eliezer, 2002; Yaniv & Foster, 1997). This paper describes a novel eyewitness identification procedure that allows participants to vary the grain size of their identification decisions. Participants eliminated members of the lineup one by one until they could eliminate no-one further, creating choice sets of variable size. They then rated their confidence that each person in the choice set was the offender (see Sauer, Brewer, & Weber, 2008). We investigated whether participants used the grain size option adaptively, and we compared the probative value of fine-grained and coarse-grained suspect identifications with suspect identifications from standard simultaneous lineups. Below, we outline the rationale for our approach, drawing on Koriat and Goldsmith's (1996a) strategic regulation of memory reporting framework and related research into the expression of "partial knowledge" in multiple-choice tests (Coombs, Milholland, & Womer, 1956; Dressel & Schmid, 1953).

The Strategic Regulation of Memory Reporting

Reporting information from memory is not as simple as searching through all the stored information and outputting all of the relevant material. Rather, memory reporting is guided by the goals of the individual within the context in which the information is reported (Koriat & Goldsmith, 1996b). The important work of Koriat, Goldsmith, and colleagues has identified two goals that drive memory reporting decisions: accuracy and informativeness (Ackerman & Goldsmith, 2008; Goldsmith, Koriat, & Pansky, 2005; Goldsmith et al., 2002;

Koriat & Goldsmith, 1996a; Pansky, Goldsmith, Koriat, & Pearlman-Avnion, 2009). When knowledge is certain, these goals do not conflict, as one can be both accurate and precise. However, under conditions of uncertainty, the goals of accuracy and informativeness may conflict. For example, if asked a difficult general knowledge question (such as "NATO includes how many member countries?"), it may prove impossible to provide an answer that is both precise and likely to be correct. Accordingly, individuals must *monitor* the likely accuracy of their memories, and *control* their output to optimise their reports within the goals and pragmatic constraints of the situation (Koriat & Goldsmith, 1996a).

Given complete freedom in their memory reports, people have two primary options for controlling their responses: they may withhold a response entirely (the *report option*; Koriat & Goldsmith, 1996a), or they made provide a less precise, coarse-grained response (the *grain size* option; Ackerman & Goldsmith, 2008; Goldsmith et al., 2002; Yaniv & Foster, 1997). Our focus is on the grain size option. In everyday interactions, respondents can freely vary the grain size of their responses. For example, when estimating the number of member countries of NATO, a respondent could report anything from a very fine-grained response (e.g., "33") to an extremely course-grained response (e.g., "5 to 90"). Of course, as grain size increases, the likelihood that the response is correct also increases, but the informativeness decreases. So how does an individual select the appropriate grain size for a memory report? The logical strategy for achieving high accuracy would be to provide extremely coarse estimates. However, such estimates are likely to be useless, and will also violate social norms of communication, potentially exposing the reporter to ridicule or scorn (Yaniv & Foster, 1995). Indeed, participants seem to be averse to providing extremely course estimates. In a striking example of this principle, Yaniv and Foster (1997) asked participants to provide interval estimates for general knowledge questions with numeric answers (e.g., "Date of the first trans-Atlantic flight?") such that they were 95% confident that the correct answer fell

within the interval. However, the intervals provided by the participants contained the correct answers only 43% of the time. In fact, on average, the participants would have needed to expand their intervals by a factor of 17 to achieve an accuracy rate of 95%. Ackerman and Goldsmith (2008) formalized this aversion to coarse-grained estimates in their dual-criterion model. They proposed that participants set both a confidence criterion and an informativeness criterion. If the participant cannot formulate a response that satisfies both criteria, then the participant will generally prioritise informativeness over accuracy, choosing to provide a more fine-grained alternative despite the lower probability that it is correct. Such a strategy enables the responder to avoid violating conversational norms. Indeed, Yaniv and Foster (1995) showed that judges tend to prefer inaccurate fine-grained responses (within a certain degree of error) to accurate coarse-grained estimates.

Importantly, the decision criteria that guide the selection of fine- versus course-grained information are under the control of the respondent and thus can be moved around strategically in response to changes in the relative importance of accuracy and informativeness (Ackerman & Goldsmith, 2008; Goldsmith et al., 2002). For example, an individual might select a different grain size when answering the same question in casual conversation with a friend than when testifying in a court of law. Experimentally, participants have been shown to lower their confidence criterion, selecting fine-grained responses more frequently, when high monetary incentives for informativeness are introduced (Goldsmith et al., 2002).

Much of the research on the strategic regulation of memory reporting has used semantic memory questions (e.g., Ackerman & Goldsmith, 2008; Goldsmith et al., 2002; Koriat & Goldsmith, 1996a), though recent research has extended the framework to eyewitness memory for event details (Evans & Fisher, 2011; Luna, Higham, & Martín-Luengo, 2011; Weber & Brewer, 2008). Weber and Brewer (2008), for example, showed their

participants a mock crime before asking them a series of questions regarding color-based and numerical details of the event (e.g., hair color of the offender, height of the offender). Using Goldsmith et al.'s (2002) two-phase procedure, the participants first provided a fine-grained response (exact color or specific number) and a coarse-grained response (overall tone or range of numbers) for each question, rating their confidence in each of these responses. In phase 2, the participants chose either the fine-grained or the coarse-grained response as their final response (or they could choose to withhold their response in Experiment 2). In support of Goldsmith et al.'s (2002) model, the participants were strategic in their selection of grain size; they volunteered the fine-grained response if confidence in that response was sufficiently high, otherwise they volunteered the course-grained response or withheld the response altogether.

Luna et al. (2011) expanded this general approach to event details that cannot be described by a numeric range or color. The participants viewed an event, and were then presented with a multiple-choice (5-alternative) test concerning the details of that event. In an adaptation of Goldsmith et al.'s (2002) two-phase procedure, participants first selected a single response (fine-grained) and a plural response consisting of three alternatives (coarse-grained), rating their confidence in each. They then chose whether they wished to volunteer the fine-grained or coarse-grained response in phase 2. In line with previous research, the participants selected the fine-grained response when their confidence in that response was high; otherwise, they selected the coarse-grained response. Luna et al.'s paper is important because it demonstrates that the grain size approach can be adapted for recognition decisions that involve choosing between discrete possibilities rather than decisions that fall on a continuum or range (see also Luna & Martín-Luengo, 2012; Higham, 2013).

In this research, we explored whether participants would strategically regulate grain size in an eyewitness identification task. However, rather than forcing participants to select a

specified number of alternatives, as in Luna et al. (2011), we wished to allow our participants complete freedom over the grain size of their responses. There are two obvious possible methods for allowing participants to vary grain size of eyewitness identification decisions: requiring participants to select alternatives for inclusion in a choice set and requiring participants to eliminate alternatives to create a final choice set. Both of these options have been explored within the domain of multiple-choice testing (e.g., Coombs et al., 1956; Dressel & Schmid, 1953).

Despite their logical equivalence, participants tend to produce coarser responses when asked to eliminate options from a set than when asked to include options in a set (Bereby-Meyer, Meyer, & Budescu, 2003; Weber, Woodard, & Williamson, 2013; Yaniv & Schul, 1997). It seems that participants set different decision criteria for excluding alternatives than for including alternatives. One consequence of this asymmetry is that the variability in grain size tends to be greater under exclusion instructions than under inclusion instructions, potentially rendering it a more useful tool for distinguishing between levels of knowledge. Indeed, pilot testing revealed that inclusion instructions were unlikely to be successful within the context of an eyewitness identification task. Of 19 participants tested (each of whom completed four lineups), only one ever included more than one lineup member in their choice set, indicating a strong resistance to producing a coarse-grained response. Thus, we required our participants to eliminate lineup members to create their final choice sets.

The Grain Size Lineup

In the grain size condition, participants saw six-person simultaneous lineups and were asked to eliminate faces, one by one, until they could no longer confidently rule anybody out, creating a choice set of variable size. We then asked the participants to rate their confidence (from 0 to 100% in 10% increments) that each face in the choice set was the offender (as in Sauer et al.'s, 2008, multiple-confidence procedure). Note that elimination methods have

been used in some prior research (e.g., Pozzulo & Lindsay, 1999), but these elimination lineup procedures have always required the participants to eliminate all but one person from the lineup. In other words, these participants could not control the grain size of their responses.

We assigned decisions from grain size lineups into four categories: 1) *fine-grained suspect identification* – the choice set included the suspect and no fillers; 2) *coarse-grained suspect identification* - the choice set included the suspect and at least one filler; 3) *filler identification* - the choice set included at least one filler but did not include the suspect; 4) *non-identification* – the choice set included no-one. Thus, the grain size lineup distinguished between two types of suspect identification: coarse-grained and fine-grained.[1] The participants saw two staged crime videos, each featuring two targets, and completed four lineups (one for each target). This multiple-lineup method increased our statistical power and added to the generalizability of our findings (Westfall, Kenny, & Judd, 2014), plus it allowed us to explore specific item differences in the use of the grain size option (cf. Brewer, Keast, & Sauer, 2010; Brewer & Wells, 2006).

We focused on two major research questions. First, would participants make use of the option to vary the grain size of their identifications, and would they use this option more frequently when memory strength was weaker? Second, how would the information gain provided by fine- and coarse-grained suspect identifications from grain size lineups compare to the information gain provided by suspect identifications from standard lineups? We outline our predictions for each of these questions in turn.

First, if the grain size procedure is to be useful, we needed to demonstrate that participants vary the grain size of their identification decisions. Furthermore, we needed to demonstrate that the proportion of coarse-grained responses increases under conditions of greater difficulty (Ackerman & Goldsmith, 2008; Goldsmith et al., 2002). To this end, we

experimentally manipulated the quality of encoding. We created two versions of our mock-crime videos, a high-quality and a low-quality version. To create the low-quality version, we halved the exposure duration to each of the targets and we added visual noise to the video. We predicted that the low-quality condition would lead to a greater proportion of coarse-grained decisions than the high-quality condition. As each participant completed four different lineups, we were also able to conduct an item analysis, capitalising on any a priori differences in stimulus difficulty. We predicted that the more difficult lineups would be associated with a greater proportion of coarse-grained suspect identifications than easier lineups.

Second, we examined the probative value of fine-grained and coarse-grained suspect identifications, and compared them to suspect identifications from standard lineups. A police lineup can be thought of as a hypothesis test; the police have a suspect, and they wish to test their hypothesis that the suspect is guilty (Wells & Luus, 1990). The eyewitness's decision provides information that either increases or decreases the probability that the suspect is guilty (Wells & Olson, 2002). The extent to which one should adjust their belief that the suspect is guilty is referred to as information gain. We predicted that fine-grained suspect identifications would provide more information than coarse-grained identifications. We also predicted that fine-grained suspect identifications might provide more information than suspect identifications from standard lineups, as these identifications were likely to include witnesses who would have provided a coarse-grained response if the option were presented.

Method

Participants and Design

A 2 (Lineup type: standard, grain size) × 2 (Encoding quality: high, low) × 2 (Target presence: target-present, target-absent) × 4 (Target number) mixed design was used, with

target presence and target number manipulated within participants. Each participant viewed two target-present and two target-absent lineups.

Participants were 385 undergraduates, postgraduates, and university staff, who received approximately $13 USD for participating in the study. One participant in the grain size lineup condition was excluded as he responded inaccurately on all three practice trials, indicating a failure to understand the task. Of the remaining 384 participants, 187 (49%) were male; the mean age was 21.31 years (*SD* = 5.19).

Participants were tested in groups of up to six, with each participant seated in a separate cubicle. Assignment of individuals to conditions was random; 192 participants were allocated to each of the standard lineup and grain size lineup conditions.

Materials

Mock crime videos. In the interests of generalizability, each of our participants viewed four targets across two staged crime events. Event 1 featured two male targets (Targets 1 and 2) of dissimilar appearance engaging in a drug deal. Event 2 featured two female targets (Targets 3 and 4), of dissimilar appearance, in a convenience store; one of the women steals several items before leaving the store.

We created high- and low-quality versions of the same two mock crime events. To create the low-quality versions, the resolution of the high-quality video was lowered and visual noise was added; furthermore, the exposure time to each target was approximately halved by halving the length of each of the shots that included any of the targets' faces. Across the four targets, full face exposure time varied from approximately 5 s to 8 s in the high-quality condition, and from 2.5 s to 4 s in the low-quality condition. The targets' faces were partially visible for a further 9 s to 20 s in the high-quality condition, and for a further 4.5 s to 10 s in the low-quality condition. Approximately half of the participants saw the high-quality videos, and the remainder saw the low-quality videos.

Lineups. Modal descriptions were created from the descriptions of three independent participants. Sixteen description-matched faces were sourced for each target from a large database. These images were evaluated by 24 mock witnesses who read the description of a target before seeing 17 faces presented simultaneously (the target plus the fillers). The mock witnesses crossed out any images that did not match the description. The four images that were eliminated most frequently for each target were removed, leaving 12 potential fillers per target.

To select the innocent suspects, 15 new participants rated the similarity of each of the targets to its 12 fillers on a seven-point scale (1 = *very dissimilar* to 7 = *very similar*). For each target, the filler with the highest similarity rating was designated as the innocent suspect. Mean similarity ratings for the innocent suspects ranged from 4.07 (*SD* = 1.39) to 5.27 (*SD* = 1.44). By selecting the best-matching filler as the innocent suspect, we provide a worst-case scenario and, therefore, the strongest test of the efficacy of the grain size procedure in discriminating guilty from innocent suspects.

Each participant saw a 6-person lineup that included 5 fillers randomly drawn from the pool of 11 fillers available for each suspect. We used this strategy to reduce the possibility that our results would be driven by a small number of fillers and to increase the heterogeneity of the retrieval conditions across witnesses. Each lineup was seen as a 2 × 3 array, with the order of the photographs randomized for each participant.

Procedure

Participants watched the two videos (in counterbalanced order) before completing a 15-minute filler task (a series of mazes). Each participant then saw four 6-person lineups (two target-present) in counterbalanced order. The participants were informed that they would see four lineups, each of which corresponded to one of the four people in the videos. They were also told that the target people may or may not appear in the lineups. They were not informed

as to the number of target-present and target-absent lineups they would see. Before each lineup, the participants were told which of the target people the lineup corresponded to (e.g., "the drug dealer from the park"), and they were reminded that the target may or may not be present.

In the standard lineup condition, the participants made their choices by clicking on a face or by clicking on a button at the bottom of the screen labelled *Not present*. Following their decision, they were asked to provide a confidence judgment from 0 to 100% (in 10% increments) that their decision was correct, using a slider bar.

In the grain size lineup condition, the participants saw the 2 × 3 array of photographs, underneath which were two buttons: *Eliminate all* and *Done*. The participants were told that their task was to eliminate any faces that they were confident did not match the specified target person. They were told that they could remove as many or as few faces as they wished, and that they did not need to leave anybody in the array. To exclude a face, the participants clicked on an image, at which point the image was grayed out yet still visible. To reverse a decision, the participants could click on a face again to put it back into the choice set. If the participants wished to exclude all of the faces, they could do so in two ways: by manually excluding all six images, or by clicking the *Eliminate all* button. To submit their final choice set, the participants clicked *Done*. At this point, any excluded faces were removed from the array, and the participants rated their confidence, from 0 to 100% (using a drop-down menu underneath each face), that each of the remaining faces was the relevant person from the video. When the participants were happy with their confidence ratings, they clicked "Done" to submit them. If all of the faces were excluded, the participants saw a new screen with a 0 to 100% rating scale, and were asked how confident they were that the target person had not been in the lineup.

Prior to completing the experimental trials, the participants completed three practice trials. The purpose of the practice trials was two-fold: to familiarise the participants with the grain size lineup; and to make it clear that leaving one image, multiple images, or no images in the choice set were all acceptable responses. In each practice trial, participants saw a 2 × 3 array of photographs of animals; their task was to exclude all of the images that were not of cats. The first practice trial included one cat, the second practice trial included three cats, and the third practice trial included no cats. Thus, the appropriate responses were to exclude all but one image in practice trial 1, all but three images in practice trial 2, and all of the images in practice trial 3. Participants were excluded if they got all three practice trials incorrect; one participant met this criterion.

Results

Overview

In most of the following analyses, we used mixed-effects logistic regression. The data were analysed using the lme4 package (Bates, Maechler, & Bolker, 2011) for R (R Core Team, 2013). The fixed effects (predictors) varied across analyses, but included: Lineup type (0 = standard lineup; 1 = rejection lineup); Target-presence (0 = target-absent; 1 = target-presence); Encoding condition (0 = poor-quality; 1 = high-quality); Target number (1, 2, 3, or 4); Grain-size (for grain-size lineups only; 0 = fine-grained; 1 = coarse-grained); and confidence (centered). Random intercepts for participant number and target number were included in all models (except where explicitly noted), which allowed the intercepts to vary separately for each participant and target (Wright & London, 2009). Where they improved the fit of the model, random slopes were also included. These are noted explicitly for each model. For brevity, only statistically significant effects ($p < .05$), and those central to our research questions, are reported in the text. The full details of each model, including all non-significant effects, can be found in the online Supplemental Materials.

When comparing fixed effects with two levels, we report the odds ratios (*ORs*). These were calculated by exponentiating the log-odds ratios (*lnOR*s) produced by the regression model. 95% Confidence Intervals (CIs) were calculated around the log-odds ratios, using the following formula: $lnOR \pm (1.96 * SE)$. The CIs were then exponentiated to be on the same scale as the ORs. If the CI around an OR excluded 1, then the association was deemed to be statistically significant.

Decision Outcomes from Grain Size and Standard Lineups

Before we addressed our main research questions, we compared the likelihood of each type of identification decision between standard and grain size lineups (see Table 1 for proportions). Each decision type was coded as a binary variable (e.g., filler identification versus any other decision) for the purposes of logistic regression analyses. For each outcome, the following fixed effects were included: Lineup Type, Target presence, Encoding condition, and all interactions between them. Target number and participant number were included as random intercepts. For the analyses of suspect identifications, the slope for Target presence was allowed to vary by target number.

In Analysis 1 (see Supplemental Materials for full details of each analysis), the outcome variable was suspect identifications. In this analysis, we included only fine-grained suspect identifications from grain size lineups. The likelihood of a suspect identification was higher if the lineup was target-present than if the lineup was target-absent, $OR = 7.24$, 95% CI [2.11, 24.90]. The likelihood of a suspect identification did not significantly differ between standard and grain size lineups, $OR = 1.11$, 95% CI [0.50, 2.42].

We then re-ran the above analysis, this time re-coding suspect identifications to include coarse-grained identifications from grain size lineups (Analysis 2). The interaction term between Target Presence and Lineup Type was statistically significant, $OR = 7.17$, 95% CI [2.69, 19.11]. To interpret this interaction, separate regression models were created for

target-present and -absent lineups, with lineup type as a fixed effect and target number and participant number as random intercepts. For target-present lineups (Analysis 3), the likelihood of a suspect identification was not statistically different for the standard and grain size lineups, *OR* = 1.36, 95% CI [0.97, 1.91]. However, for target-absent lineups (Analysis 4), the likelihood of an innocent suspect identification was significantly higher for grain size lineups than for the standard lineup, *OR* = 2.92, 95% CI [1.82, 4.67]. Thus, when coarse-grained identifications were considered as suspect identifications, the result was an increase in the likelihood of innocent suspect identifications in the grain size lineup as compared to the standard lineup. This increase in false identifications was not accompanied by any significant gain in correct identifications.

In Analysis 5, the outcome variable was filler identifications. Fillers were less likely to be identified from target-present lineups than from target-absent lineups, *OR* = 0.53, 95% CI [0.34, 0.81]. No other effects were statistically significant.

Finally, in Analysis 6, the outcome variable was non-identifications. The likelihood of a non-identification was significantly lower for target-present lineups than for target-absent lineups, *OR* = 0.55, 95% CI [0.36, 0.85]. In addition, the likelihood of a non-identification was significantly lower for grain size lineups than for standard lineups, *OR* = 0.46, 95% CI [0.29, 0.73].

To summarise, if only fine-grained suspect identifications were considered to be true suspect identifications, the grain-size lineup produced a similar proportion of correct suspect identifications and false suspect identifications as the standard lineup. However, when coarse-grained identifications were included, the likelihood of a misidentification of an innocent suspect increased substantially. Thus, the grain-size lineup elicited a larger number of potentially harmful errors than the standard lineup.

Adaptive Use of the Grain Size Option

Our first major research question was whether participants would use the grain size option adaptively. To answer this question, we conducted three analyses. First, we examined whether participants in the low-quality encoding condition were more likely to provide coarse-grained responses than participants in the high-quality encoding condition. Second, we examined whether lineups that were, a priori, more difficult (as determined by results from the standard lineup condition), were associated with a higher likelihood of coarse-grained responses than less difficult lineups. Third, we examined confidence ratings assigned to fine- and coarse-grained responses. For each of these analyses, we coded all responses to grain size lineups as fine-grained (the choice set included 0 or 1 faces) or coarse-grained (the choice set included more than one face). Of 768 total responses, 201 (26.17%) were coarse-grained.

First, we asked whether the likelihood of a coarse-grained response varied between the high- and low-quality encoding conditions. We conducted a mixed-effects logistic regression analysis, in which we predicted grain size (fine vs. coarse) from Encoding condition, Target presence, and their interaction (Analysis 7). Participant number and target number were included as random intercepts; no random slopes were included as they did not significantly improve the fit of the model. Contrary to predictions, the likelihood of a coarse-grained response did not significantly differ between the good and poor quality encoding conditions, $OR = 0.61$, 95% CI [0.31, 1.24], and no other fixed effects were statistically significant. Thus, participants did not appear to use the grain size option to compensate for the uncertainty created by a low-quality view.

Second, we asked whether lineups that were more difficult, for whatever reason, were associated with a higher likelihood of coarse-grained responses than lineups that were less difficult. To establish whether our lineups did, indeed, vary in difficulty, it was first necessary to explore item differences in performance on standard lineups. Specifically, we predicted correct suspect identifications from target-present lineups (Analysis 8) and false suspect

identifications from target-absent lineups (Analysis 9) from target number, with participant number as a random intercept. The proportions are shown in Table 2. Target 1 was used as the referent for all comparisons, as he was associated with the lowest proportion of correct identifications and the highest proportion of false identifications. The likelihood of a correct suspect identification was significantly higher for all other targets in comparison to Target 1: Target 2 versus Target 1, *OR* = 4.57, 95% CI [2.05, 10.21]; Target 3 versus Target 1, *OR* = 5.00, 95% CI [2.20, 11.40]; Target 4 versus Target 1, *OR* = 10.07, 95% CI [4.17, 24.34]. Furthermore, the likelihood of a false suspect identification was significantly lower for Target 3 than for Target 1, *OR* = 0.39, 95% CI [0.10, 0.97].

Having established that Target 1 was the most difficult target to identify, we examined whether participants were more likely to provide coarse-grained responses for Target 1 than for all other targets. To test this hypothesis, we ran two mixed-effects logistic regression in which we predicted grain size (fine vs. coarse) from Target number; Analysis 10 included only target-present lineups, while Analysis 11 included only target-absent lineups. In both analyses, participant number was included as a random intercept. See Table 3 for proportions. For target-present lineups, the likelihood of a coarse-grained response was significantly lower for Target 3 than for Target 1, *OR* = 0.30, 95% CI [0.10, 0.97], and for Target 4 than for Target 1, *OR* = 0.24, 95% CI [0.10, 0.61]. For target-absent lineups, the likelihood of a coarse-grained response was significantly lower for Target 2 than for Target 1, *OR* = 0.40, 95% CI [0.18, 0.89]. Thus, as predicted, participants provided a coarse grained response most frequently for the target who was most difficult to identify, providing some evidence of adaptive use of the grain size option.

Third, we examined participants' confidence ratings in fine- and coarse-grained responses. Goldsmith et al. (2002) showed that participants reported a coarse-grained response when confidence in the fine-grained response was low (and thus, the fine-grained

answer was likely to be incorrect). A direct test of this hypothesis requires that participants provide confidence judgments for those fine-grained answers that are volunteered and withheld (i.e., those that are discarded in favour of the coarse-grained answer). Our procedure provided us with confidence ratings for volunteered fine-grained responses but not for withheld fine-grained responses. However, when our participants gave a coarse-grained response, they provided a separate confidence rating for each face that was not excluded (as opposed to a single confidence rating in their coarse-grained answer, as in most grain size studies, e.g., Goldsmith et al., 2002; Luna et al., 2011; Weber & Brewer, 2008). We therefore created a proxy for confidence in withheld fine-grained responses, which required the following assumptions: 1) When a participant provided a coarse-grained response, the face awarded the highest confidence rating was the face that would have been provided as the fine-grained response; and 2) the confidence rating assigned to that face would have been the same whether the response was fine-grained or coarse-grained. To clarify, if a participant identified two faces and provided confidence ratings of 60% for Face A and 20% for Face B, we assumed that Face A would have been provided as the fine-grained response, and it would have been assigned a confidence rating of 60%. Based on Goldsmith et al.'s model, we predicted that mean confidence in fine-grained responses would be higher than the mean confidence associated with the highest-rated face in coarse-grained responses.

To test this prediction, we created a new confidence variable, which was coded as the maximum confidence rating assigned to any one face in the choice set. We created a mixed-effects regression model, predicting confidence (which was centered around 0 prior to analyses) from grain size (Analysis 12). Participant number and target number were included as random intercepts. We included only those participants who made at least one fine-grained identification decision and at least one coarse-grained identification decision (n = 81). We excluded trials in which all faces were eliminated for two reasons: first, these trials cannot be

classified as fine or coarse-grained; second, it is unclear how a confidence judgment in a non-identification would relate to the confidence that would have been assigned to the best match in the lineup. Consequently, we could not make any predictions for these trials. This left 277 trials in the analysis, of which 130 (46.93%) were coarse-grained. Grain size was a significant predictor of confidence, β = -6.29, (SE = 2.36), indicating that confidence was approximately 6 percentage points lower for coarse-grained than for fine-grained decisions. Thus, it seems that participants provided a fine-grained response if confidence in that response was sufficiently high; otherwise they provided a coarse-grained response, which is consistent with strategic regulation of grain size.

Information gain

Our second main research question centred around the probative value of fine- and coarse-grained suspect identifications. To answer this question, we used a Bayesian information gain approach, as advocated by Wells and Olson (2002). From experimental data, we can estimate the probability that a suspect will be identified given that he is guilty (pSuspect ID|Suspect = Guilty) and the probability that a suspect will be identified given that he is not guilty (pSuspect ID|Suspect ≠ Guilty). Dividing the former by the latter gives the *diagnosticity*, or the degree to which one should adjust their belief in the suspect's guilt given the outcome: (pSuspect ID|Suspect = Guilty)/(pSuspect ID|Suspect ≠ Guilty). Diagnosticity can vary from 0 to infinity, with a value of 1 indicating that the suspect was no more likely to be identified when guilty than when innocent. If we wish to estimate a suspect's guilt given that he was identified (pSuspect = Guilty|Suspect ID), we must consider the *prior probability* that the suspect was guilty. In experiments, the prior probability (i.e., the target-absent base-rate) is usually set to .50, but in the real world the target-absent base-rate is unknown, and likely varies greatly across jurisdictions. Therefore, Wells and Olson advocate an approach whereby information gain is plotted across the entire range of prior probabilities.

We calculated diagnosticity, and plotted information gain curves, for high-confidence (90-100%), medium-confidence (60-80%), and low-confidence (0-50%) suspect identifications (fine-grained, coarse-grained, and standard) separately. We used three confidence bins to increase the reliability of the diagnosticity estimates; the particular cut-off points were chosen to divide the data as evenly as possible across the three categories. The information gain curves are shown in Figure 1, and the diagnosticity estimates are shown in Table 4, with 95% CIs. Note that the CIs were calculated using log-transformed diagnosticity ratios, which were then back-transformed for ease of interpretation. Also shown in Table 4 are the results of inferential tests comparing the diagnosticity estimates at each level of confidence, based on the procedure recommended by Tredoux (1998).

When confidence was high, all three types of suspect identification were informative (see Figure 1A), and diagnosticity did not significantly vary across the identification types, $\chi^2(2) = 1.10$, $p > .10$. High-confidence responses are important as they are most likely to be relied upon by investigators and to contribute to a prosecution. A criticism of diagnosticity is that it tends to favour the situation that minimizes false identifications, even if that reduction is accompanied by a sizeable drop in the correct proportion of correct identifications (Wixted & Mickes, 2012). Thus, we compared the proportion of correct high-confidence correct identifications from standard lineups with the proportion of correct fine-grained identifications from grain size lineups in a mixed-effects logistic regression (Analysis 13). Participant number and target number were included as random intercepts. The likelihood of a high-confidence suspect identification was not significantly different for standard and grain size lineups, $OR = 0.81$, 95% CI [0.52, 1.25]. The proportion of high-confidence correct identifications from standard lineups was 17.19% (95% CI [13.75%, 21.28%]), and the proportion of high-confidence fine-grained identifications from grain size lineups was 14.58% (95% CI [11.40%, 18.46%]).

For medium-confidence suspect identifications, however, diagnosticity did differ significantly across the identification types, $\chi^2(2) = 11.06$, $p < .01$. As is clear from Figure 1B, standard identifications and fine-grained identifications were informative, but coarse-grained identifications were not. Pairwise comparisons (with a Bonferroni corrected alpha of .017) revealed that diagnosticity was significantly lower for the coarse-grained responses than for the standard lineups, $\chi^2(1) = 9.42$, $p < .01$, and for the fine-grained responses, $\chi^2(1) = 9.42$, $p < .01$. Diagnosticity did not significantly differ for the standard lineups and the fine-grained responses, $\chi^2(1) = 0.21$, $p > .10$.

Finally, when confidence was low, information gain was negligible (see Figure 1C). However, diagnosticity did significantly differ for the three types of decision, $\chi^2(2) = 12.31$, $p < .01$. Pairwise comparisons (Bonferroni corrected) showed that diagnosticity was significantly lower for the coarse-grained responses than for responses from standard lineups, $\chi^2(1) = 8.62$, $p < .01$, and for fine-grained responses, $\chi^2(1) = 8.27$, $p < .01$. In fact, low confidence coarse-grained responses were indicative of innocence, as demonstrated by a diagnosticity ratio of less than one, and an information gain curve that skews to the right (Figure 1C). The standard lineups and fine-grained responses did not significantly differ from each other, $\chi^2(1) = 0.30$, $p > .10$.

Discussion

In this paper, we tested a novel eyewitness identification procedure, inspired by research into the regulation of grain size in memory reporting (Ackerman & Goldsmith, 2008; Goldsmith et al., 2002; Pansky et al., 2009). Participants who completed a grain size lineup eliminated lineup members one by one until they could eliminate no-one further, in a method adapted from the psychometric testing literature (Coombs et al., 1956). Our key findings were: 1) Participants appeared to use the grain size option adaptively, providing a higher proportion of coarse-grained responses for the most difficult lineup, and providing

coarse-grained responses when confidence in the fine-grained response was relatively low; 2) Fine-grained suspect identifications were no more probative than suspect identifications from standard lineups; and 3) Coarse-grained suspect identifications were probative only if the confidence rating awarded to the suspect was high. We discuss each of these key findings in turn, after which we discuss the broader theoretical and applied implications of this research.

Our first hypothesis was that participants would use the grain size option adaptively, providing a greater proportion of coarse-grained responses under conditions of increased uncertainty. We found some support for this hypothesis, as participants were more likely to provide coarse-grained decisions for the most difficult lineup. Furthermore, confidence in fine-grained responses was higher than confidence in the highest-rated face in coarse-grained responses, which is consistent with Goldsmith et al.'s (2002) contention that grain size selection is guided primarily by confidence in the fine-grained response. However, the participants were no more likely to provide coarse-grained decisions when the encoding conditions had been poor than when they had been good. This finding parallels results from Weber and Perfect's (2012) work examining *don't know* responses to showups. They found that the frequency of *don't know* responses was unaffected by a manipulation of retention interval (immediate versus 3-week delay), despite a substantial effect of the delay on forced-report accuracy. The most likely explanation for these findings appears to be that confidence is less sensitive to increases in difficulty than accuracy itself (e.g., Weber & Brewer, 2004). If the choice of grain size is based on confidence, and if confidence is relatively insensitive to factors that decrease accuracy, then a likely consequence is over-reporting of fine-grained responses. An interesting avenue for future research is the degree to which participants can be helped to accurately monitor the likely accuracy of their decisions (e.g., Lane, Roussel, Villa, & Morita, 2007), and the consequence of this improved monitoring for the regulation of grain size.

Our final hypothesis was that fine-grained responses would be more probative than coarse-grained responses, and possibly more probative than standard suspect identifications. When the confidence rating given to the suspect was very high (90-100%) or very low (0-50%), the three types of suspect identifications did not significantly differ in their informativeness. Given a high confidence rating, all three types of suspect identification were informative; but given a low confidence rating, no type of suspect identification was informative. However, when confidence was medium (60-80%), coarse-grained suspect identifications were significantly less informative (in fact, they were uninformative) than fine-grained and standard suspect identifications.

Contrary to expectations, there were no circumstances in which fine-grained suspect identifications were more informative than standard suspect identifications. Participants who selected only one face even when they had the opportunity to identify more, were no more accurate than those who could choose, at most, a single face. Our speculation is that the effectiveness of the grain size lineup was hampered by a general reluctance to provide a coarse-grained response. Indeed, only a minority of responses were coarse-grained, and there were many participants who never provided a coarse-grained response (approximately 45% of the sample). We note that an even stronger reluctance was observed in a pilot study in which participants selected faces for inclusion in a choice set. Consistent with prior research (e.g., Bereby-Meyer et al., 2003; Weber et al., 2013; Yaniv & Schul, 1997), requiring participants to eliminate items, rather than including items, did appear to ameliorate this problem somewhat. If the grain-size lineup is to be useful, it will be important to encourage greater use of the coarse-grain option when participants are uncertain.

It is possible that willingness to violate the informativeness criterion forms a cognitive trait within individuals, such that some participants have a stronger aversion than others to providing coarse-grained estimates (see Kantner & Lindsay, 2012, for evidence of reliable

individual differences in response biases). This possibility was noted by Coombs et al. (1956), who suggested that there may be individual differences in the willingness to go beyond one's sure knowledge. A second, non-exclusive, possibility is that the eyewitness identification task instils in participants a strong expectation that their responses be as informative as possible, and that providing a coarse-grained response is seen as a violation of social norms (Ackerman & Goldsmith, 2008; Yaniv & Foster, 1997). Despite providing instructions and practice trials that emphasized the acceptability of coarse-grained responses, some participants may have nonetheless considered coarse-grained responses to be insufficiently informative. Interestingly, recent research focusing on eyewitnesses' memory reporting of fine- and coarse-grained details also strongly suggests that coarse-grained details appear to be generally considered as insufficiently informative to report (McCallum, Brewer, & Weber, 2015). Both individual differences and situational variations in willingness to provide coarse-grained responses provide potentially fruitful avenues for future research, and would contribute greatly to our theoretical understanding of how participants use grain size to regulate accuracy.

Coarse-grained suspect identifications were substantially less probative than fine-grained identifications. Importantly, this effect runs counter to the informativeness-accuracy trade-off reported elsewhere (e.g., Weber & Brewer, 2008; Goldsmith et al., 2002; Yaniv & Foster, 1997), whereby volunteered coarse-grained responses are more accurate than volunteered fine-grained responses. However, we included a condition that has not been investigated in any prior investigation of grain size regulation: a condition in which the correct alternative is not present. In the target-absent condition, only 14.86% of fine-grained identifications included the innocent suspect; however, 48.06% of coarse-grained identifications from target-absent lineups included the suspects. Thus, the chances of committing a harmful error, given a target-absent lineup, were considerably greater when a

coarse-grained response was provided than when a fine-grained response was provided. Of course, this pattern will vary depending upon how the innocent suspect is selected. We chose "worst-case" innocent suspects, who were rated as being most similar to the target. This pattern of a marked increase in errors for coarse-grained responses may be less pronounced for innocent suspects who less closely resemble the culprit.

Our results highlight that there are specific situations in which coarse-grained responses are both less informative *and* less accurate than fine-grained responses. A potential avenue for future research would be to explore grain size regulation in response to unanswerable or misleading questions, in which participants are asked to recall information that was not part of the witnessed event (e.g., Scoboria, Mazzoni, & Kirsch, 2008; Waterman, Blades, & Spencer, 2001). Our data suggest that in such circumstances, coarse-grained responses may be more likely to include some critical piece of misinformation than fine-grained responses, with potentially damaging effects on accuracy.

We note that there is some overlap between our conceptualization of the grain-size method and the relative-absolute judgment framework of eyewitness identification decisions (Wells, 1984). Relative judgments involve comparisons between lineup members such that a witness identifies the person who looks most like their memory of the culprit, relative to the other members of the lineup. Absolute judgments involve a direct comparison between an individual face in the lineup and the memory of the offender. If the grain-size procedure worked optimally, fine-grained responses would capture these strong, absolute judgments (e.g., "It's definitely number 5!"), while coarse-grained responses might capture more relative decisions (e.g., "It could be number 4 or 5…."). However, the degree to which relative and absolute judgments map on to fine- and coarse-grained decisions is not likely to be perfect, as the decision to report a fine- or coarse-grained response is a complex one that is influenced by many factors (Ackerman & Goldsmith, 2008; Goldsmith et al., 2002).

A final theoretical point is that fillers can provide important information about the state of a witness's memory and the likelihood that any decisions made are correct. The focus in eyewitness identification research is usually on suspect identifications, with filler identifications dismissed as non-harmful errors (except when used as a proxy for innocent suspect identifications in designs with no designated innocent suspect). Indeed, recently advocated analytic techniques such as receiver operating characteristics (ROC) analyses collapse filler identifications together with non-identifications (see Mickes, Flowe, & Wixted, 2012). Our results complement those of several other studies that have shown that fillers are important. For example, Wells and Olson (2002) showed that filler identifications provide exculpatory value; given that a filler is identified, the suspect is usually more likely to be innocent than guilty (see also Wells, Yang, & Smalarz, 2015). In another recent example, Charman and Cahill (2012) showed that memory for fillers in a subsequent recognition test was negatively related to target-present identification accuracy. Finally, in Sauer et al.'s (2008) multiple-confidence procedure, participants rated their confidence that each person in the lineup was the culprit; the difference in confidence between the highest-rated face and the other faces in the lineup was a powerful predictor of accuracy (see also Brewer, Weber, Wootton, & Lindsay, 2012). We add to these findings here, showing that a witness who is unable to eliminate all of the fillers is less likely to have identified the guilty culprit than one who is able to eliminate all of the fillers.

Relatedly, witnesses to real crimes do sometimes choose more than one individual from a lineup. Wells, Steblay, and Dysart (2015) found that 30 of 494 actual eyewitnesses made multiple identifications from a single lineup. Very few laboratory studies allow multiple identifications, and so we know very little about this type of response. Though multiple identifications seem to form a small minority of lineup decisions, we could potentially gain

valuable insights into the decision-making process through studying them in a more systematic manner.

This study is a first step in adapting the grain size approach for use in eyewitness identification tasks. We have shown that some, though not all, participants, are able to regulate the grain size of their identification decisions in response to the difficulty of the task. However, the grain-size lineup fared no better than the standard lineup in any key regard; fine-grained suspect identifications were no more probative than standard suspect identifications, and coarse-grained identifications carried virtually no probative information. The utility of the procedure appeared to be hampered by a general reluctance to provide coarse-grained responses. Future refinements to the procedure may include training procedures, incentives, or instructions to encourage participants to become more conservative in their use of fine-grained decisions, which may allow us to identify a subset of decisions with higher probative value than decisions of a similar confidence level from standard lineups, as well as a range of more graded responses of differing evidence strength.

Though procedures such as this may seem somewhat radical, several prominent researchers have called for more innovative research into eyewitness identification procedures that breaks away from the simultaneous-sequential dichotomy that has dominated the field for three decades (e.g., Wells, Memon, & Penrod, 2006). Indeed, promising new techniques are currently evolving whereby witnesses make no identification decisions at all - they simply judge the extent to which each member of the lineup resembles their memory of the culprit (Brewer et al., 2012; Sauer et al., 2008). Only with theoretically grounded yet innovative research will the field move forward toward procedures that maximize the information available to investigators and the judicial system and that minimize the likelihood of miscarriages of justice based upon wrongful identifications.

References

Ackerman, R., & Goldsmith, M. (2008). Control over grain size in memory reporting – with and without satisficing knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1224-1245. doi: 10.1037/a0012938

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects models using s4 classes.* Available at www.r-project.org

Bereby-Meyer, Y., Meyer, J., & Budescu, D. V. (2003). Decision making under internal uncertainty: The case of multiple-choice tests with different scoring rules. *Acta Psychologica, 112,* 207-220. doi: 10.1016/S0001-6918(02)00085-9

Brewer, N., Keast, A., & Sauer, J. D. (2010). Children's eyewitness identification performance: Effects of a Not Sure response option and accuracy motivation. *Legal and Criminological Psychology, 15,* 261-277. doi: 10.1348/135532509X47822

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science, 23,* 1208-1214. doi: 10.1177/0956797612441217

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

Charman, S. D., & Cahill, B. S. (2012). Witnesses' memories for lineup fillers postdicts their identification accuracy. *Journal of Applied Research in Memory and Cognition, 1,* 11-17. doi: 10.1016/j.jarmac.2011.08.001

Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement, 16,* 13-37. doi: 10.1177/001316445601600102

Dressel, P. L., & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement, 13,* 574-595. doi: 10.1177/001316445301300404

Evans, J. R. & Fisher, R. P. (2011). Eyewitness memory: Balancing the accuracy, precision and quantity of information through metacognitive monitoring and control. *Applied Cognitive Psychology, 25,* 501-508. doi: 10.1002/acp.1722

Goldsmith, M., Koriat, A., & Pansky, A. (2005). Strategic regulation of grain size in memory reporting over time. *Journal of Memory and Language, 52,* 505-525. doi: 10.1016/j.jml.2005.01.010

Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). Strategic regulation of grain size in memory reporting. *Journal of Experimental Psychology: General,* 131, 73-95. doi:10.1037/0096-3445.131.1.73

Higham, P. A. (2013). Regulating accuracy on university tests with the plurality option. *Learning and Instruction, 24,* 26-36. doi: 10.1016/j.learninstruc.2012.08.001

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition, 40,* 1163-1177. doi: 10.375/s13421-012-0226-0

Koriat, A., & Goldsmith, M. (1996a). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490-517. doi: 10.1037/0033-295X.103.3.490

Koriat, A., & Goldsmith, M. (1996b). Memory metaphors and the real-life/laboratory controversy: Correspondence versus storehouse conceptions of memory. *Behavioral and Brain Sciences, 19,* 167-188. doi: 10.1017/S0140525X00042114

Lane, S. M., Roussel, C. C., Villa, D., & Morita, S. K. (2007). Features and feedback: Enhancing metamnemonic knowledge at retrieval reduces source-monitoring errors.

Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 1131-1142. doi: 10.1037/0278-7393.33.6.1131

Luna, K., Higham, P. A., & Martín-Luengo, B. (2011). Regulation of memory accuracy with multiple answers: The plurality option. *Journal of Experimental Psychology: Applied.* doi: 10.1037/a002327

Luna, K., & Martín-Luengo, B. (2012). Improving the accuracy of eyewitnesses in the presence of misinformation with the plurality option. *Applied Cognitive Psychology, 26*, 687-693. doi: 10.1002/acp.2845

McCallum, N., Brewer, N., & Weber, N. (2015). Memorial monitoring and control: How confidence and social and financial consequences affect eyewitnesses' reporting of fine-grain information. *Manuscript under review.*

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*, 361-376. doi: 10.1037/a0030609

Pansky, A., Goldsmith, M., Koriat, A., & Pearlman-Avnion, S. (2009). Memory accuracy in old age: Cognitive, metacognitive, and neurocognitive determinants. *European Journal of Cognitive Psychology, 21*, 303 – 329. doi: 10.1080/09541440802281183

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

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General,* 137, 528-547. doi: 10.1037/a0012712

Scoboria, A., Mazzoni, G., & Kirsch, I. (2008). "Don't know" responding to answerable and unanswerable questions during misleading and hypnotic interviews. *Journal of Experimental Psychology: Applied, 14,* 255-265. doi: 10.1037/1076-898X.14.3.255

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

Waterman, A. H., Blades, M., & Spencer, C. (2001). Interviewing children and adults: The effect of question format on the tendency to speculate. *Applied Cognitive Psychology, 15,* 521-531. doi: 10.1002/acp.741

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied, 10*, 156-172. doi: 10.1037/1076-898X.10.3.156

Weber, N. & Brewer, N. (2008). Eyewitness recall: Regulation of grain size and the role of confidence. *Journal of Experimental Psychology: Applied,* 14, 50-60. doi: 10.1037/1076-898X.14.1.50

Weber, N., & Perfect, T. J. (2012). Improving eyewitness identification accuracy by screening out those who say they don't know. *Law and Human Behavior, 36*, 28-36. doi: 10.1037/h0093976

Weber, N., Woodard, L., & Williamson, P. (2013). Decision strategies and the confidence–accuracy relationship in face recognition. *Journal of Behavioral Decision Making, 26*, 152-163. doi: 10.1002/bdm.1750

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

Wells, G. L. & Luus, C. A. E. (1990). Police lineups as experiments – social methodology as a framework for properly conducted lineups. *Personality & Social Psychology Bulletin, 16,* 106-117. doi: 10.1177/0146167290161008

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

Wells, G. L. & Olson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied,* 8, 155-167. doi: 10.1037/1076-898X.8.3.155

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2015). Double-blind photo lineups using actual eyewitnesses: An experimental test of a sequential versus simultaneous lineup procedure. *Law and Human Behavior, 39,* 1-14. doi: 10.1037/lhb0000096

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,* 99-122. doi: 10.1037/lhb0000125

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143,* 2020-2045. doi: 10.1037/xge0000014

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace Receiver Operating Characteristic analysis. *Perspectives on Psychological Science, 7,* 275-278. doi: 10.1177/1745691612442906

Wright, D. B. & London, K. (2009). Multilevel modelling: Beyond the basic applications. *British Journal of Mathematical & Statistical Psychology,* 62, 439-456. doi: 10.1348/000711008X327632

Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making, 10,* 21-32. doi: 10.1002/(SICI)1099-0771(199703)10:1<21:AID-BDM243>3.0.CO;2-G

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General, 124,* 424-432. doi: 10.1037/0096-3445.124.4.424

Yaniv, I. & Schul, Y. (1997). Elimination and inclusion procedures in judgment. *Journal of Behavioral Decision Making, 10,* 211-220. doi: 10.1002/(SICI)1099-0771(199709)10:3<211::AID-BDM250>3.0.CO;2-J

Footnote

[1]Note that filler identifications could also be split into fine- and coarse-grained decisions. However, as all filler identifications are known to be incorrect, we treated them as a single category of response here.

Table 1.

Proportion of Decision Outcomes from Grain Size Lineups [with 95% Confidence Intervals], by Target-Prese

	Target-present				T	
	Suspect ID				Suspect ID	
Lineup type	Fine	Coarse	Filler ID	Non-ID	Fine	Coars
				High-quality view		
Grain size	43.9	12.2	31.1	12.8	5.1	10.7
	[37.1, 50.9]	[8.4, 17.6]	[25.1, 37.9]	[8.8, 18.2]	[2.8, 9.1]	[7.1, 15
Standard	49.5	-	29.7	20.8	8.9	-
	[42.7, 56.3]	-	[23.8, 36.3]	[15.8, 26.9]	[5.7, 13.6]	-
				Low-quality view		
Grain size	27.7	14.4	37.2	20.7	8.5	15.4
	[21.7, 34.5]	[10.1, 20.1]	[30.6, 44.3]	[15.6, 27.1]	[5.3, 13.4]	[11.0, 21
Standard	35.7	-	35.7	28.6	7.7	-
	[29.1, 42.9]		[29.1, 42.9]	[22.5, 35.5]	[4.6, 12.5]	
				Overall		
Grain size	35.9	13.3	34.1	16.7	6.8	13.0
	[31.3, 40.9]	[10.2, 17.0]	[29.6, 39.0]	[13.3, 20.7]	[4.7, 9.7]	[10.0, 16
Standard	43.0	-	32.6	24.5	8.3	-
	[38.1, 48.0]		[28.1, 37.4]	[20.4, 29.0]	[6.0, 11.5]	

Table 2

Proportions of Suspect identifications [with 95% Confidence Intervals] from Standard Lineups as a Function of Target Number

Target presence	Target number			
	Target 1	Target 2	Target 3	Target 4
Target-present	23.33	39.00	45.74	62.00
	[15.80, 33.05]	[30.02, 48.80]	[36.04, 55.78]	[52.21, 70.90]
Target-absent	14.71	5.43	5.10	7.61
	[9.12, 22.85]	[2.34, 12.1]	[2.20, 11.39]	[3.73, 14.88]

Table 3

Proportions of Coarse-grained Responses [with 95% Confidence Intervals] from Grain Size Lineups as a Function of Target Number

Target-presence	Target number			
	Target 1	Target 2	Target 3	Target 4
Target-presence	35.42	24.21	23.16	18.37
	[26.58, 45.38]	[16.71, 33.72]	[15.82, 32.58]	[11.95, 27.18]
Target-absent	35.42	21.65	24.74	26.60
	[26.58, 45.38]	[14.62, 30.84]	[17.23, 34.18]	[18.72, 36.32]

Table 4.

Diagnosticity Estimates [with 95% Confidence Intervals] for Suspect Identifications Made with High, Medium, and Low Confidence.

Identification type	Diagnosticity	95% Confidence Intervals		Difference test
		Lower	Upper	
		High confidence		
Standard	22.00	6.98	69.37	$\chi^2(2) = 1.10$, $p > .10$
Fine-grained	11.20	4.54	27.65	
Coarse-grained	8.00	1.01	63.66	
		Medium confidence		
Standard	5.73	3.07	10.69	$\chi^2(2) = 11.06$, $p < .01$
Fine-grained	4.69	2.62	8.40	
Coarse-grained	1.40	0.73	2.67	
		Low confidence		
Standard	2.00	1.16	3.46	$\chi^2(2) = 12.31$, $p < .01$
Fine-grained	2.63	1.18	5.85	
Coarse-grained	0.65	0.39	1.09	
		Overall		
Standard	5.16	3.63	7.24	$\chi^2(2) = 51.01$, $p < .001$
Fine-grained	5.31	3.58	7.87	
Coarse-grained	1.02	0.71	1.47	

Supplemental Materials

Below are the full regression models for each of the analyses reported in the manuscript.

For the random effects, the standard deviation is provided, which is an estimate of the variability of the random effect. For fixed effects, the log-odds ratio (*lnOR*) and its standard error are provided, along with a Wald's *z* test and associated *p* value.

Analysis 1: Comparing suspect identifications from standard lineups with fine-grained suspect identifications from grain size lineups

Random effects (SD)				
Participant (intercept)	0.60			
Target (intercept)	0.38			
Target presence (slope)	1.07			
Fixed effects	**lnOR**	**SE**	**Wald's *z***	***p***
Intercept	-2.69	0.36	-7.54	<.001
Lineup Type	0.10	0.40	0.26	.79
Target-presence	1.98	0.63	3.13	.002
Encoding condition	0.14	0.39	0.35	.73
Lineup Type × Target presence	-0.54	0.46	-1.17	.24
Lineup Type × Encoding condition	-0.69	0.58	-1.18	.24
Target presence × Encoding condition	0.50	0.44	1.13	.26
Lineup type × Target presence × Encoding condition	0.89	0.66	1.36	.18

Analysis 2: Comparing suspect identifications from standard lineups with fine-plus-coarse-grained suspect identifications from grain size lineups

Random effects (SD)				
Participant (intercept)	0.43			
Target (intercept)	0.37			
Target presence (slope)	0.75			
Fixed effects	**lnOR**	**SE**	**Wald's *z***	***p***
Intercept	-2.62	0.34	-7.68	<.001
Lineup Type	1.39	0.33	4.15	<.001
Target-presence	1.97	0.50	3.97	<.001
Encoding condition	0.15	0.38	0.39	.69
Lineup Type × Target presence	-1.09	0.40	-2.76	.006
Lineup Type × Encoding condition	-0.69	0.46	-1.48	.14
Target presence × Encoding condition	0.45	0.43	1.06	.29
Lineup type × Target presence × Encoding condition	0.71	0.55	1.29	.20

Analysis 3: Comparing the likelihood of a correct suspect identification from standard lineups with the likelihood of a suspect identification (either fine or coarse) from grain size lineups

Random effects (SD)				
Participant (intercept)	0.69			
Target (intercept)	0.52			
Fixed effects	**lnOR**	**SE**	**Wald's *z***	***p***
Intercept	-0.35	0.29	-1.22	.22
Lineup Type	0.31	0.17	1.79	.07

Analysis 4: Comparing the likelihood of a false suspect identification from standard lineups with the likelihood of a suspect identification (either fine or coarse) from grain size lineups

Random effects (SD)				
Participant (intercept)	0.57			
Target (intercept)	0.37			
Fixed effects	**lnOR**	**SE**	**Wald's *z***	***p***
Intercept	-2.61	0.31	-8.29	<.001
Lineup Type	1.07	0.24	4.38	<.001

Analysis 5: Comparing filler identifications from standard lineups and grain size lineups

Random effects (SD)				
Participant (intercept)	0.38			
Target (intercept)	0.25			
Fixed effects	**lnOR**	**SE**	**Wald's *z***	***p***
Intercept	0.02	0.20	0.11	.91
Lineup Type	0.02	0.22	0.10	.92
Target-presence	-0.64	0.22	-2.90	.004
Encoding condition	-0.07	0.22	-0.33	.74
Lineup Type × Target presence	0.05	0.31	0.15	.88
Lineup Type × Encoding condition	0.22	0.31	0.71	.48
Target presence × Encoding condition	-0.21	0.31	-0.68	.49
Lineup type × Target presence × Encoding condition	-0.22	0.43	-0.51	.61

Analysis 6: Comparing non-identifications from standard lineups and grain size lineups

Random effects (SD)				
Participant (intercept)	0.00			
Target (intercept)	0.00			
Fixed effects	**lnOR**	**SE**	**Wald's *z***	***p***
Intercept	-0.33	0.16	-2.06	.04
Lineup Type	-0.77	0.23	-3.40	<.001
Target-presence	-0.59	0.22	-2.63	.009
Encoding condition	0.01	0.21	-0.07	.95
Lineup Type × Target presence	0.35	0.33	1.04	.30
Lineup Type × Encoding condition	0.22	0.31	0.71	.48
Target presence × Encoding condition	-0.44	0.32	-1.38	.17
Lineup type × Target presence × Encoding condition	-0.38	0.48	-0.79	.43

Analysis 7: Likelihood of a coarse-grained response

Random effects (SD)				
Participant (intercept)	1.50			
Target (intercept)	0.28			
Fixed effects	**lnOR**	**SE**	**Wald's *z***	***p***
Intercept	-1.19	0.30	-4.00	.04
Target-presence	-0.41	0.28	-1.48	.14
Encoding condition	-0.49	0.36	-1.37	.17
Target presence × Encoding condition	0.56	0.39	-1.44	.15

Analysis 8: Likelihood of a correct identification from a standard lineup as a function of target number

Random effects (SD)				
Participant (intercept)	0.80			
Fixed effects	**lnOR**	**SE**	**Wald's *z***	***p***
Intercept	-2.12	0.37	-5.72	<.001
Target 1 vs Target 2	1.52	0.41	3.72	<.001
Target 1 vs Target 3	1.61	0.42	3.84	<.001
Target 1 vs Target 4	2.31	0.45	5.08	<.001

Analysis 9: Likelihood of a false identification from a standard lineup as a function of target number

Random effects (SD)				
Participant (intercept)	1.52			
Fixed effects	**lnOR**	**SE**	**Wald's *z***	***p***
Intercept	-2.59	0.81	-3.18	.001
Target 1 vs Target 2	-1.01	0.60	-1.68	.09
Target 1 vs Target 3	-1.19	0.59	-2.01	.04
Target 1 vs Target 4	-0.76	0.55	-1.38	.17

Analysis 10: Likelihood of a coarse-grained response from a target-present lineup as a function of target number

Random effects (SD)				
Participant (intercept)	1.45			
Fixed effects	**lnOR**	**SE**	**Wald's *z***	***p***
Intercept	-0.79	0.30	-2.61	.009
Target 1 vs Target 2	-0.78	0.42	-1.87	.06
Target 1 vs Target 3	-0.94	0.43	-2.18	.03
Target 1 vs Target 4	-1.41	0.47	-3.00	.003

Analysis 11: Likelihood of a coarse-grained response from a target-absent lineup as a function of target number

Random effects (SD)				
Participant (intercept)	1.27			
Fixed effects	**lnOR**	**SE**	**Wald's *z***	***p***
Intercept	-0.84	0.30	-2.94	.005
Target 1 vs Target 2	-0.92	0.41	-2.26	.03
Target 1 vs Target 3	-0.59	0.39	-1.53	.13
Target 1 vs Target 4	-0.50	0.39	-1.30	.19

Analysis 12: Predicting confidence from grain size

Random effects (SD)			
Participant (intercept)	10.26		
Target (intercept)	1.82		
Residual	18.84		
Fixed effects	***β***	**SE**	***t***
Intercept	2.77	2.16	1.28
Grain size	-6.29	2.36	-2.66

Analysis 13: Comparing proportion of high-confidence correct identifications: standard lineups vs fine-grained responses

Random effects (SD)				
Participant (intercept)	0.84			
Target (intercept)	0.89			
Fixed effects	**lnOR**	**SE**	**Wald's *z***	***p***
Intercept	-2.03	0.49	-4.16	<.001
Lineup type	-0.21	0.22	-0.94	.35