Evaluating Earwitness Identification Procedures: Adapting Pre-Parade Instructions and Parade Procedure

Harriet M. J. Smith[a], Jens Roeser[a], Nikolas Pautz[a], Josh P. Davis[b], Jeremy Robson[c], David Wright[d], Natalie Braber[d] and Paula C. Stacey[a]

[a] Department of Psychology, Nottingham Trent University

[b] School of Human Sciences, University of Greenwich

[c] Leicester De Montfort Law School, De Montfort University

[d] English, Communications and Philosophy, Nottingham Trent University

Author Note

Correspondence concerning this article should be addressed to Harriet M. J. Smith, Department of Psychology, Nottingham Trent University, 50 Shakespeare Street, Nottingham, NG1 4FQ. Telephone number: +44 (0) 115 8484535. Email: harriet.smith02@ntu.ac.uk

Abstract

Voice identification parades can be unreliable, as earwitness responses are error-prone. In this paper we tested performance across serial and sequential procedures, and varied pre-parade instructions, with the aim of reducing errors. The participants heard a target voice and later attempted to identify it from a parade. In Experiment 1 they were either warned that the target may or may not be present (standard warning) or encouraged to consider responding ‘not present’ because of the associated risk of a wrongful conviction (strong warning). Strong warnings prompted a conservative criterion shift, with participants less likely to make a positive identification regardless of whether the target was present. In contrast to previous findings, we found no statistically reliable difference in accuracy between serial and sequential parades. Experiment 2 ruled out a potential confound in Experiment 1. Taken together, our results suggest that adapting pre-parade instructions provides a simple way of reducing the risk of false identifications.

Keywords: voice identification, lineup instructions, earwitness, serial, sequential

Evaluating Earwitness Identification Procedures: Adapting Pre-Parade Instructions and Parade Procedure

Earwitness evidence is relevant when a witness is able to hear a perpetrator's voice while not being able to see their face. Such cases often relate to serious crimes, like rape or murder, where disguises may be worn (*R v Khan and Bains*, 2002, discussed in Nolan, 2003; *R v Flynn and St John*, 2008). Voice identification is error-prone; earwitnesses are likely to identify an incorrect voice (Kerstholt et al., 2004, 2006; H. M. J. Smith et al., 2020). As highlighted within the context of eyewitness identification, if an innocent suspect has been apprehended, a false identification increases the risk of an erroneous conviction (Innocence Project, 2020). Finding ways to reduce such errors is vital, but methods of adapting voice identification procedures to optimize earwitness performance are under-researched. In this paper we investigate the effect of pre-parade instructions and procedure type, with the aim of informing procedural changes that support earwitness performance.

Parade Types

During a voice parade[1] the suspect is presented amongst 'foil', or distractor voices. If the 'target' (i.e., perpetrator) is present the witness might identify the target voice (a 'hit') or a foil voice (a 'miss'). Alternatively, they might incorrectly reject the parade by responding 'not present'. If the target is absent the witness may incorrectly identify someone (a 'false alarm'), despite having the option to respond, 'not present'. There are various ways of presenting parade voices, but the Home Office (2003) recommends a serial procedure, requiring witnesses to listen to all 9 parade voices before making a decision. Working memory demands, which require storing the target voice along with all other voices in the parade, may contribute to task difficulty. An alternative method, the sequential procedure,

[1] We refer to parades rather than lineups because the term 'parade' is used by practitioners in England and Wales.

potentially reduces these demands as it involves responding 'yes' or 'no' after hearing each voice. H. M. J. Smith et al. (2020) found higher hits and lower false alarms with the sequential compared to the serial parade. Voice identification is subject to interference from intervening identity information (Stevenage et al., 2011). Posing a question after each voice may mitigate the effect of interference by demarcating the voices (H. M. J. Smith et al., 2020).

Patterns of performance across parade procedures have been thoroughly investigated in the context of face identification (e.g., Brewer et al., 2012; Carlson et al., 2008; Lindsay et al., 2009). Although false alarms might be lower when faces are presented sequentially rather than simultaneously (Clark et al., 2008; Steblay et al., 2011), the sequential procedure appears to lead to a stricter decision standard and overall lower rates of choosing (Ebbesen & Flowe, 2002; Mickes et al., 2012). This highlights the importance of considering both hit and false alarm rates when designing parades. However, the findings may not generalise across modalities because (1) cognitive processes involved in recognising faces and voices are not identical (Belin et al., 2011; Belin et al., 2004; Young et al., 2020), (2) faces can be presented simultaneously while voices cannot, and (3) listening to a voice likely involves focusing on the meaning of what is being said rather than identity-specific sound information (Fenn et al., 2011; Vitevitch, 2003).

The Effect of Instructions

Witnesses receive instructions prior to completing an identification parade. We know that the content of these instructions can influence an eyewitness's decision, and stronger warnings can reduce false alarms (Brewer & Wells, 2006; Clark, 2005; Malpass & Devine, 1981; Meissner et al., 2005; Steblay, 1997). No previous earwitness studies have systematically manipulated the content of instructions. This is an important omission considering the high false alarms associated with voice parades.

Unbiased instructions (i.e., warning witnesses that the perpetrator may not be present), are mandatory in England and Wales (Police and Criminal Evidence Act, Code D, 1986) and are included in guidelines in other common law jurisdictions (Fitzgerald et al., 2020). *Biased* instructions do not include this warning, creating an insinuation that the perpetrator is indeed present. Unbiased instructions seem to encourage eyewitnesses to raise their criterion for selecting the lineup member who looks most familiar (Brewer & Wells, 2006). Such instructions are associated with lower false alarms on target-absent face parades (Brewer & Wells, 2006; Clark, 2005), with the potential to halve the number of mistaken identifications (Malpass & Devine, 1981). The results of Steblay's (1997) meta-analysis suggested that lower false alarms were not at the cost of a reduction in correct identifications and overall lower rates of choosing. However, re-analysis of the data set by Clark (2005) revealed that unbiased instructions are associated with a criterion shift. That is, there are fewer false alarms *and* fewer correct identifications; guilty or not, the suspect is less likely to be identified.

To mitigate against possible miscarriages of justice, unbiased warnings are necessary. Therefore, it is important to consider what form they should take. Indeed, the effect of biased instructions may vary according to their exact wording. Lampinen et al. (2020) and Wilcock, Bull and Vrij (2005) found no difference between a standard and enhanced version, in which eyewitnesses are additionally reminded that an incorrect identification might lead to false imprisonment. However, the instructions may have been too demanding (Wilcock et al., 2005), or may not have produced a sufficient impression (Lampinen et al., 2020; Meissner et al., 2005). Meissner et al. (2005) compared the standard instruction to stronger, criterion-based instructions, in which participants were told to make a positive identification only if they were 100% sure. The stronger instruction improved discrimination, reducing false alarms but not hits.

Confidence

The wording of instructions may affect identification confidence. Leippe and colleagues (2009) found that in some circumstances biased instructions can contribute to inflated confidence ratings; a positive accuracy cue is associated with the parade member most closely matching the eyewitness' memory, regardless of 'guilt'. This translates to a weaker confidence-accuracy relationship following biased instructions. Such a pattern might be elusive for earwitnesses, as confidence-accuracy relationships tend to be weak or non-existent (e.g., Kerstholt et al., 2004; H. M. J. Smith et al., 2020; but see Sarwar et al., 2014). Previous studies have tended not to thoroughly address earwitness confidence, so it is unclear why this is the case. However, H. M. J. Smith et al. (2020) report that participants often record their confidence in the middle of the scale. The difficulty of voice identification might prompt noncommittal responding which effectively masks potential relationships from emerging (H. M. J. Smith et al., 2020).

The Current Study

The earwitness literature has not addressed the effect of pre-lineup instructions. Here we compare standard unbiased instructions (standard warning) to a strong warning, encouraging participants to consider responding 'not present'. We test the effect of warnings in serial parades (Home Office, 2003), and sequential parades, as these might be more appropriate for assessing voice identification performance (H. M. J. Smith et al., 2020). We expected that hits would be low, and the false alarms would be high, but that the strong warning would make participants less likely to false alarm in both types of parade. On balance, we do not expect this reduction in false alarms to be at the expense of hits (Meissner et al., 2005). We predict that accuracy on target-present and target-absent parades will be higher for the sequential compared to the serial procedure. Overall, we do not expect to observe a reliable relationship between confidence and accuracy.

Experiment 1: Pre-Parade Instructions and Parade Procedure

Method

Design. We used a 2 x 2 x 2 between-subjects factorial design. The factors were parade type (serial, sequential), parade instructions (strong warning, standard warning), and target presence (present, absent). Voices identified as targets (1 = yes, 0 = no) and self-rated confidence (0-10) were the dependent variables.

Participants. We recruited 561 participants from [BLINDED] database. All participants had previously completed the Cambridge Face Memory Test + (CFMT+) and agreed to be contacted about future experiments. We removed data from 28 participants who reported uncorrected hearing problems. The final sample included 533 participants (337 female, 195 male, 1 prefer not to say) with an age range of 18–75 years (*M* = 45.84, *SD* = 12.55). Excluding 7 participants who had missing data, the mean CFMT+ scores (*M* = 86.37, *SD* = 9.62) were somewhat higher than typical (70.7; Bobak, Pampoulov, & Bate, 2016). The experiment was approved by the [BLINDED] University's Business, Law and Social Science College Research Ethics Committee.

Apparatus and materials. The voice stimuli were taken from the Dynamic Variability in Speech Database (DyViS) (Nolan et al., 2009). This database features 100 male speakers between the ages of 18 and 25, all with a Standard Southern British English (SSBE) accent. The speakers are recorded performing spoken tasks, such as a simulated police interview. All of the recordings used in this experiment were made in a sound-treated booth and were studio quality (44.1 kHz/16 bit) (Nolan et al., 2009). The recordings used in each voice parade were the same as those used by H. M. J. Smith et al. (2020): Thirty speakers were randomly selected from the database and assigned into three 10-speaker groups based on fundamental frequency (F0) (low, medium, and high). From each group, a target-absent and target-present parade were constructed; this meant that overall there were three target-present parades and three target-absent parades. In the target-present parade the target either appeared in an early position (position 3), or a late position (position 7). Target position varied within targets. The three target voice samples were taken from the recording of a telephone conversation during which the speakers discussed a crime.[2] These recordings were used for the encoding stage. The recording was edited so that it was 60s long and featured only the targets' side of the conversation. As all speakers were responding to the same scripted questions, the content was similar for all three targets. The voice parade speech samples were selected from the simulated mock police interview recordings. These recordings were used for the test stage. All interviewer speech content was removed and only excerpts featuring the interviewees were combined to produce 15s samples. The voice samples for each speaker were from different, randomly selected sections of the police interview, meaning that the content of speech varied across speakers. The content of the telephone-recording and interview samples did not overlap. All of the parades were fair and unbiased (Malpass & Lindsay, 1999), as reported in H. M. J. Smith et al. (2020): None of the parades were found to be biased towards the target, and there were several viable alternatives to each target among the foils; Tredoux's *E* varied from 3.80 to 7.22 across parades. Participants completed a wordsearch containing words for different types of fruit during the retention interval. The axes of the wordsearch were numerically labelled and participants were required to enter the coordinates of the X and Y axes for the first letter of each word. While completing the wordsearch, a recording of ambient noise featuring unintelligible speech sounds played in the background.

Procedure. Participants completed the experiment online hosted on Gorilla.sc (Anwyl-Irvine et al., 2019). They gave informed consent, set their volume, calibrated their headphone volume, completed a headphone screening test (Woods et al., 2017), and were

[2] All of the recordings were studio rather than telephone quality.

randomised to one of the eight conditions. The participants were also randomised to a speaker group (1, 2 or 3), and a target position (3 or 7; target present parades only), although these were not included as factors. Each participant completed a single trial.

Prior to the presentation of the target voice, participants were instructed that they would hear a voice recording and be asked questions relating to what they had heard; participants were not informed that they would be undertaking a voice parade. Participants were not able to go back to previously viewed pages at any point. On pages where timing was critical, it was not possible to progress until the task had been completed. Participants were asked to click 'Play' to listen to the 60s target voice sample when they were ready to begin. Once the voice sample had finished, participants automatically progressed to the next part of the experiment where they then completed the filler task (wordsearch) for 5 minutes. After 5 minutes had passed, the instructions for the parade appeared. Participants were reminded that at the beginning of the experiment they had heard a perpetrator discussing a crime. They were instructed that they were going to listen to a voice parade in order to try and identify the perpetrator they had heard speaking in the initial recording.

Participants either completed a serial or a sequential parade:

Serial parade. Each recording was presented on a separate page with the voice number visible while the recording was playing. Participants listened to each recording (15s) once. They were informed that they were going to hear 9 voices played one after the other. Following each recording they pressed the spacebar to indicate they were ready to proceed to the next voice. After listening to all 9 voices they read the following instruction in the standard-warning condition: “The perpetrator may or may not be present. If you think the perpetrator was present, please select the correct voice below. If you think the perpetrator was not present, please select 'The perpetrator was not present'. In the strong-warning condition, in addition to these instruction participants were reminded to, 'Please consider your response carefully. In a real case, selecting someone from the lineup when the perpetrator is not present could lead to a wrongful conviction". Below the instruction, participants were given one of ten options: voice 1–9 or, "The perpetrator was not present". When they had made a selection, they were asked to assess the confidence in their decision (0 = Not at all confident, 10 = Extremely confident).

Sequential parade. Participants were informed that they would hear a series of voice recordings with the objective of trying to identify the perpetrator. They were informed that following each of the voices they would be asked to decide if the voice belonged to the perpetrator. The participants were not informed about the number of voices that would be presented, only that after they had responded 'yes', no further voices would be played. Participants listened to a voice recording (15 s), and in the standard-warning condition were then asked:" Do you think Voice [number] belongs to the perpetrator?" In the strong-warning condition, in addition to this, they were reminded: "Please consider your response carefully. In a real case, selecting someone from the lineup when the perpetrator is not present could lead to a wrongful conviction". Participants selected "yes" or "no" and provided a confidence rating (0 = Not at all confident, 10 = Extremely confident) when they responded "yes" to a voice, or after they had responded "no" to all 9 voices. As in the serial procedure, participants pressed the spacebar after each voice had played to proceed to the next voice.

After completing the parade, participants in both the serial and sequential condition were invited to answer brief questions about their experience. Two of these questions served as a manipulation check, with responses provided on an 11-point slider rating scale and having a default starting position of 5. Participants were asked, "to what extent did you consider responding that the perpetrator was not present/responding 'no' to each voice? (0 = Did not consider it at all, 10 = Strongly considered it)", and "Before completing the parade, you were warned that the perpetrator may or may not be present. To what extent did this warning influence your decision(s)? (0 = The warning had no influence on my decision, 10 = The warning had a strong influence on my decision)". Our sample showed higher mean scores for strong warnings for both manipulation checks, supporting that our participants adapted their behaviour according to the presented instructions (see Appendix A for an overview).

Results

Data were analysed using the Bayesian modelling framework (Gelman et al., 2014; McElreath, 2016). We obtained the evidence for the alternative hypothesis from Bayes Factors (BF) using the Savage-Dickey method (Dickey et al., 1970; Wagenmakers et al., 2010).[3] We summarized the posterior as the most probable population value and the interval containing 95% of the posterior probability mass (i.e. Highest Posterior Density Interval [HPDI]).[4] The benefits of using a Bayesian approach for hypothesis testing (Kruschke et al., 2012; Kruschke, 2014) and parameter estimation (Lambert, 2018; Lee & Wagenmakers, 2014) are well documented in the literature. Here we report the results in brief. Supplementary information for Experiment 1 analyses is presented in Appendices B-D.[5]

Signal-detection analysis. We evaluated voice identification in the context of signal-detection theory to independently evaluate (1) the response criterion and (2) discriminability (*d'*; e.g., Wixted et al., 2016). The response criterion is an indicator of the overall willingness to make a positive identification. A criterion (*c*) that is statistically below 0 would indicate a liberal decision criterion, while a *c* that includes 0 would be indicative of a neutral decision

[3] While there is an ongoing debate on how to interpret the strength of a Bayes Factor, a common interpretation is that a BF larger than 3 indicates moderate support for the alternative hypothesis and a BF larger than 10 indicates strong support (Baguley, 2012; Jeffreys, 1961; Lee & Wagenmakers, 2014). For example, a BF of 10 indicates that the alternative hypothesis is 10× more likely than the null hypothesis.

[4] The R (R Core Team, 2020) package brms (Bürkner, 2017, 2018) was used to model the data. Models were run with 30,000 iterations on 3 chains with a warm-up of 15,000 iterations and no thinning. Model convergence was confirmed by the Rubin-Gelman statistic ($\hat{R}$= 1) (Gelman & Rubin, 1992) and inspection of the Markov-chain Monte-Carlo chains.

[5] Data and analysis scripts can be found on osf.io/x2dpc/?view_only=b74edc32c9494dcda5802a6efb4f0981.

criterion, and finally a *c* that is statistically above 0 would indicate a conservative decision criterion. The discriminability *d'* can be understood as a measure of sensitivity to the signal, i.e., the ability to distinguish between the target voice and fillers. Thus, a higher *d'* value indicates a better ability to identify the target voice. A *d'* value that is not meaningfully different from 0 would be indicative of, at best, a chance-level ability to detect the signal from the noise, while a *d'* value meaningfully different from 0 would indicate evidence that a listener is able to discriminate the signal from the noise. To measure *d*' between guilty suspects and innocent suspects rather than absolute discriminability between guilty suspects, innocents, and foil voices (Colloff et al., 2016), we removed filler IDs in target present parades.

As discussed in the introduction, we predict that a strong warning will increase sensitivity (*d'*) to the target voice. However, it is also plausible that strong warnings will reduce the propensity of listeners to false alarm by facilitating a conservative criterion shift.

A Bayesian framework was used to infer the model parameters (DeCarlo, 1998, 2010; Rouder et al., 2007; Rouder & Lu, 2005)[6]. As there was no designated innocent suspect in the target absent parades, we adjusted the false alarm rate by the number of voices in the parade, as commonly done in the eyewitness literature (e.g., see A. M. Smith et al., 2021)[7]. Table 1 provides a descriptive overview of all response types, including foil IDs. The analysis revealed that – as can be seen in Figure 1 – participants were less inclined to make a positive identification after receiving a strong false alarm (FA) warning in both serial and sequential parades. Collapsing the data across parade type, we found moderate evidence supporting a lower (i.e., more liberal) response criterion for standard FA warnings compared to strong FA warnings ($c = -0.09$, HPDI[8]: [-0.17 – -0.02], BF = 7.51). Further, collapsing the data across

[6] There was no evidence that unequal variance SDT models increased model performance, see Appendix D.

[7] We note that there is also an argument for estimating false alarm rates using the parade's resultant effective size (e.g., see A. M. Smith et al., 2021).

[8] HPDI is the Highest Posterior Density Interval

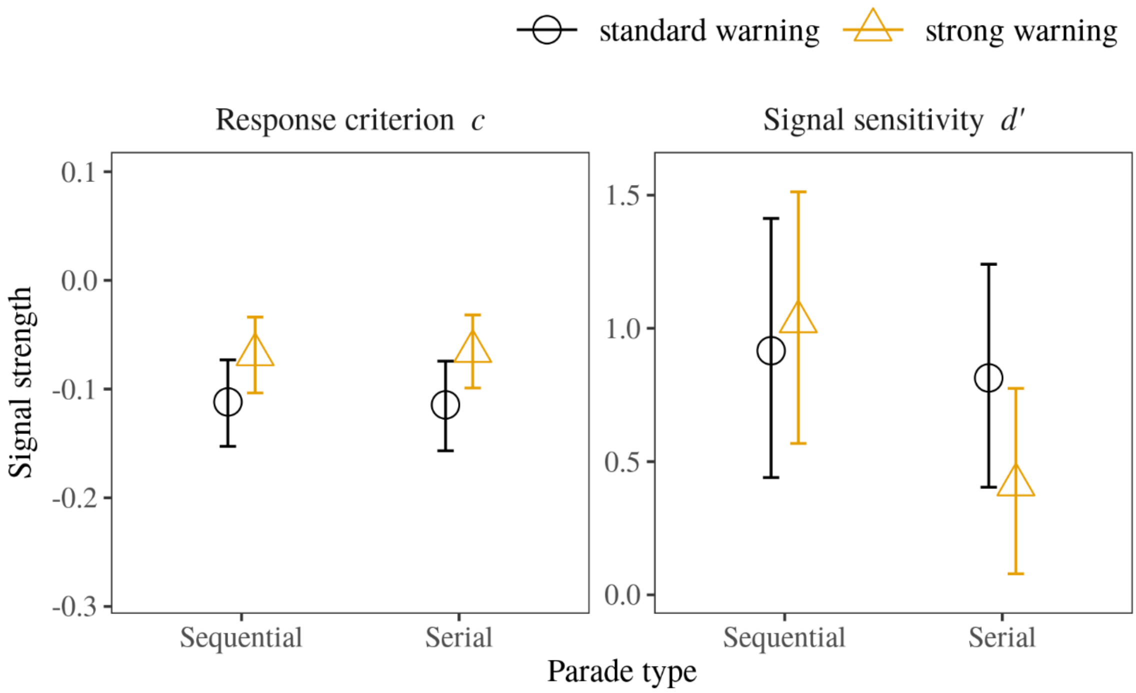
parade instructions, there was moderate evidence that participants displayed a better ability to distinguish between the target voice and ‘innocents’ for sequential parades: overall, signal sensitivity was higher for sequential parades compared to serial parades (*d’* = 0.71, HPDI: [-0.16 – 1.59], BF = 3.26). The results are summarised in Appendix B.

Table 1

Decision frequency with percentages in parentheses, Experiment 1

		Target-present			Target-absent	
Parade Type	Pre-parade Instructions	Target	Foil	Reject	Foil	Reject
Sequential	Standard Warning	25 (38%)	38 (58%)	3 (5%)	57 (85%)	10 (15%)
	Strong Warning	31 (47%)	32 (48%)	3 (5%)	50 (74%)	18 (26%)
Serial	Standard Warning	32 (47%)	30 (44%)	6 (9%)	54 (86%)	9 (14%)
	Strong Warning	28 (45%)	21 (34%)	13 (21%)	53 (73%)	20 (27%)
Total		116 (44%)	121 (46%)	25 (10%)	214 (79%)	57 (21%)

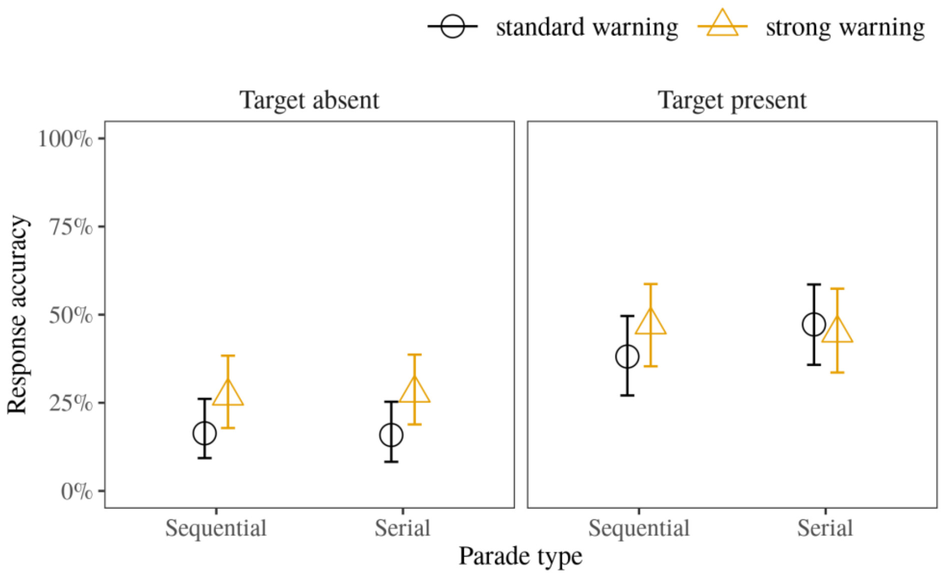
Figure 1. Parameter-value estimates with 95% HPDIs of the signal-detection theory model, the signal sensitivity d’ and the response criterion *c* (Experiment 1).



The results shown in the right panel of Figure 1 suggest that the parade type difference reported above might be driven by the strong FA warning conditions. By-parade type comparisons based on FA warning conditions showed moderate evidence for a higher signal sensitivity for strong warnings in sequential parades compared to serial parades (*d'* = 0.61, HPDI: [0.02 – 1.2], BF = 4.99); evidence for the same contrast was negligible for standard FA warnings (*d'* = 0.05, HPDI: [-0.54 – 0.75], BF = 0.69). This reflects what is shown in Figure 1: sensitivity appears similar for standard and strong warnings in a sequential parade, but appears lower for strong compared to standard warnings in a serial parade. However, there was negligible evidence for an interaction that would support this pattern (see Appendix B).

Accuracy analysis. We conducted accuracy analyses so that we could evaluate target presence, as target-present foil identifications were removed in the signal-detection analyses. Analysing accuracy also facilitated comparison with the results of H. M. J. Smith et al. (2020). We analysed response accuracy (0 = incorrect response, 1 = correct response) in a Bayesian logistic mixed model. Predictors were included for main effects and interactions of parade instructions (levels: standard warning, strong warning), parade type (levels: sequential, serial), and target presence (levels: present, absent). Evidence for all interactions was negligible (see Appendix C for full results). The cell means and 95% HPDIs are shown in Figure 2. There was negligible evidence for effects of parade type ($\hat{\beta}$ = -0.25, HPDI: [-1.43 – 0.94], BF = 0.65) or parade instructions ($\hat{\beta}$ = -0.94, HPDI: [-2.16 – 0.2], BF = 2.31). There was strong evidence that parades in which the target was present were more likely to result in an accurate decision than parades where the target was absent ($\hat{\beta}$ = -2.84, HPDI: [-4.02 – -1.64], BF > 100).

Figure 2. Posterior response accuracy with 95% HPDIs, Experiment 1.

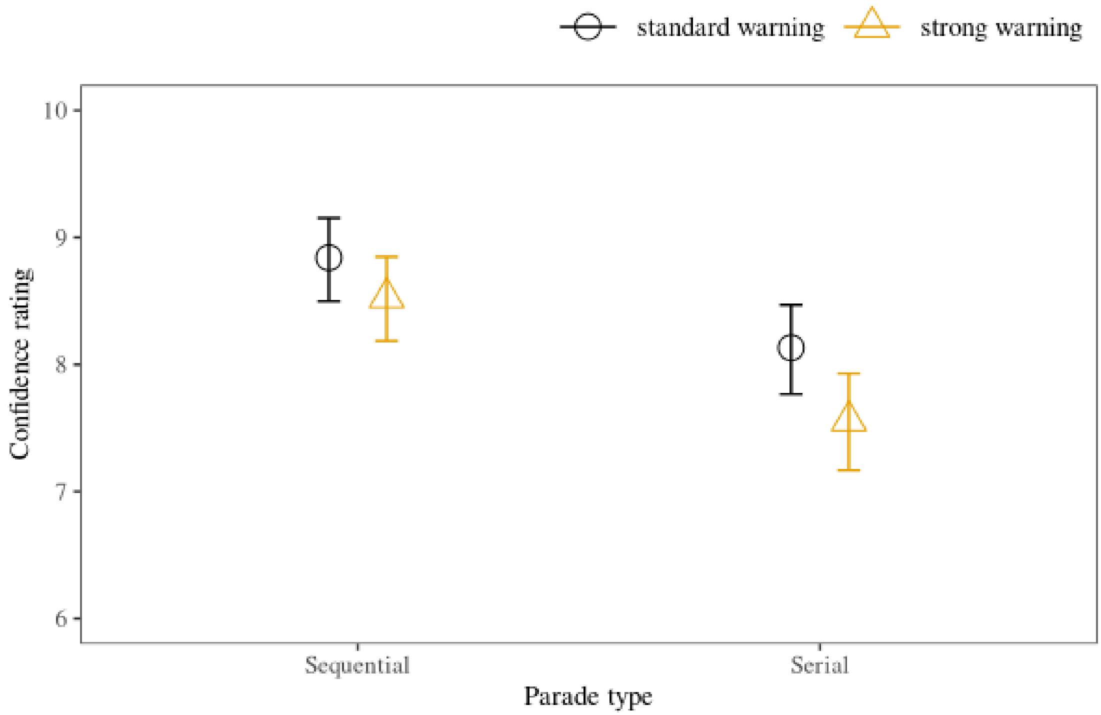


Confidence ratings. Confidence ratings, on a scale of 0-10 (0 = Not at all confident, 10 = Extremely confident), were analysed in cumulative models for ordinal data (Bürkner & Vuorre, 2019; Liddell & Kruschke, 2018).[9] We investigated the relationship between

[9] As in other identity perception research (H. M. J. Smith et al., 2020, 2021) we analyse the confidence ratings as ordinal rather than ratio data. The rationale for this, as can be found in the statistical modelling literature, is to avoid assuming equal intervals between ratings. It is not uncommon for ordinal data to be analysed using methods that assume metric responses. However, this practice can lead to errors in inference as the psychological distance between adjacent categories on psychometric scales is known to be non-identical and discrete (Liddel & Kruschke, 2018). Thus models that assume continuity and linearity are not suitable for ordinal data.

confidence ratings and accuracy separately for each condition of parade type and FA warning. Posterior cell means are shown in Figure 3 for each condition.

Figure 3. Posterior confidence with 95% HPDIs, Experiment 1.



We found strong evidence of a positive relationship between confidence and accuracy for the serial parade with a strong FA warning ($\hat{\beta}$ = 0.83, HPDI: [0.23 – 1.45], BF = 13.2), evidence was negligible for all other conditions (serial no warning: $\hat{\beta}$ = 0.63, HPDI: [0.02 – 1.23], BF = 2.4; sequential FA warning: $\hat{\beta}$ = 0.51, HPDI: [-.10 – 1.1], BF = 1.19; sequential no warning: ß = 0.46, HPDI: [-.15 – 1.12], BF = 0.95). In other words, participants were more confident about correct responses (than about incorrect responses) when receiving a strong FA warning but only in serial parades.

Discussion

Strong warning instructions were associated with a conservative criterion shift in both serial and sequential parades. This finding corroborates findings from the eyewitness literature showing that the content of instructions can influence witness decisions (Lampinen et al., 2020; Meissner et al., 2005). However, unlike Meissner et al.’s (2005) results, the strongly worded instruction did not improve discrimination.

We found no difference between sequential and serial voice parades in terms of accuracy. While we did find moderate evidence of higher sensitivity for sequential parades, it is likely that this result was driven by the strong warning conditions. These results appear to be inconsistent with the results of H. M. J. Smith et al. (2020), who did not conduct signal detection analyses, but reported higher accuracy on sequential voice parades compared to serial parades, with both including only standard unbiased warnings. Perhaps a strong warning prior to a sequential parade is worth exploring further as an identification procedure. Indeed, it is possible that the single-lap procedure adopted in Experiment 1 leads us to underestimate the sequential advantage reported by H. M. J. Smith et al. (2020). In H. M. J. Smith et al. (2020), participants listened to the serial parade twice before making a decision, but they listened only once in the present experiment. In Experiment 2 we address whether this might explain why we did not observe lower accuracy for serial compared to sequential parades.

The results are consistent with previous literature finding that accuracy was low (<50%), and false alarms were high (e.g., Kerstholt et al., 2004, 2006; H. M. J. Smith et al., 2020). Overall, as expected, we did not observe a reliable relationship between confidence and accuracy, and participants recorded surprisingly high levels of confidence despite low accuracy. However, they were reliably more confident about correct responses (than about incorrect responses) when receiving a strong FA warning in serial parades. It is feasible that in this condition the strong warning served to highlight the cognitive load and working memory demands associated with listening to 9 voices before making a decision.

Experiment 2: The Number of Laps

Home Office (2003) guidelines recommend that participants listen to each serial parade voice at least once before making a decision. In H. M. J. Smith et al. (2020) participants listened to all voices in the serial parade twice before making a decision, while in Experiment 1, participants heard serial parade voices once. The second 'lap' of voices was removed in Experiment 1 to avoid a potential confound between the two procedures, given that voices in the sequential parade are only heard once.

The lap effect has been addressed in the eyewitness literature. Yet, in the earwitness literature, there is no evidence to suggest an improved performance for 2 laps. Indeed, there is an associated risk that participants will adopt a more lenient response criterion on the second lap, being more likely to make a positive identification (Horry et al., 2015; Lindsay et al., 1991; Maclin & Phelan, 2007; Steblay et al., 2011). The majority of studies have tested the lap effect in the context of sequential parades, where participants respond after seeing each face, and so have the opportunity to identify different targets on each lap (but see Seale-Carlisle et al., 2019). This is different from implementing a second lap in a serial parade and only allowing participants to make a decision after considering all parade members twice. In Experiment 2 we test the effect of number of laps on earwitness parade responses.

Seale-Carlisle et al. (2019) reported no difference in serial visual parade outcomes as a function of the number of laps conducted. However, based on higher accuracy in sequential voice parades than serial parades in H. M. J. Smith et al. (2020) (2 laps), but observing no difference in Experiment 1 (1 lap), we tentatively expected a single serial lap to be associated with higher accuracy than 2 laps. We compare performance using a standard warning because this is consistent with the instructions provided in H. M. J. Smith et al. (2020). We did not include a strong FA warning condition because there is no reason to believe that providing a strong warning when completing 2 laps would optimize performance beyond the benefit observed for the 1-lap condition (Experiment 1).

Method

The method was identical to Experiment 1 except for the following amendments:

Design. We used a 2 x 2 between-subjects factorial design. The factors were number of laps (1 laps, 2 laps), and target presence (present, absent). Identification accuracy (1 = correct, 0 = incorrect) and self-rated confidence (0-10) were the dependent variables. As explained at the beginning of the Results section, the data for the 1-lap conditions (target present and target absent) were from Experiment 1.

Participants. We recruited 112 participants to the 2-lap condition. We removed data from: 1 participant with a missing CFMT+ score, and 4 participants who reported having uncorrected hearing problems. The final sample included 108 participants (69 female, 38 male, 1 preferred not to say) with an age range of 23-73 years (*M* = 48.72, *SD* = 12.14).

Procedure. All participants completed a serial parade with a standard warning and were randomly allocated to the target-present or target-absent condition. Participants were informed that they would listen to the parade twice before making a decision. The 9 voice samples were presented in the same order both times.

Results

Data were combined with the subset of trials from the serial parade, standard FA warning condition in Experiment 1 (*n* = 131): Participants from Experiment 1 listened to the serial parade once (1 lap) before making a decision; participants recruited for Experiment 2 listened to the serial parade twice (2 laps) before making a decision. There were no other differences across experiments.

We used the same tools for data analysis as in Experiment 1. Here we report the results in brief. Supplementary information for Experiment 2 analyses is presented in Appendices E – G.

Signal-detection theory model. We included number of laps (levels: 1 lap, 2 laps) as a fixed effect and estimated the response criterion and the signal sensitivity for each condition. Table 2 provides a descriptive overview of target and foil identifications, as well as parade rejections.

Table 2

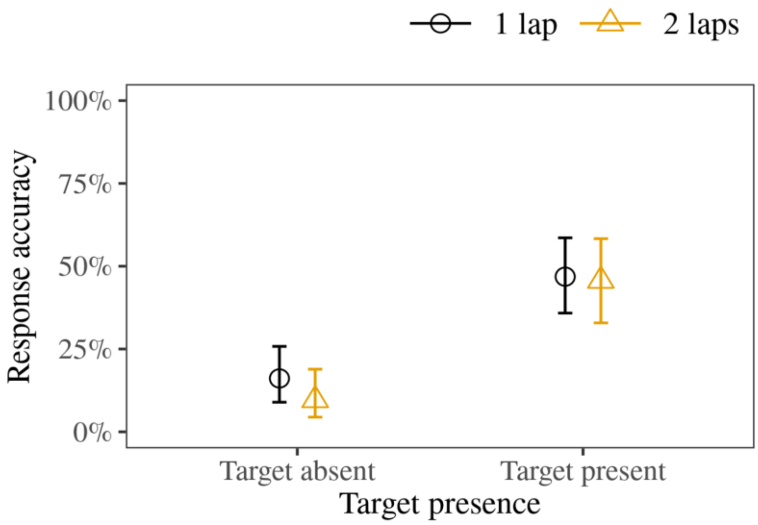
Decision frequency with percentages in parentheses, Experiment 2

Number of passes	Target-present			Target-absent	
	Target	Foil	Reject	Foil	Reject
1 pass	32 (47%)	30 (44%)	6 (9%)	54 (85%)	9 (15%)
2 passes	24 (45%)	23 (44%)	6 (11%)	51 (93%)	4 (7%)
Total	56 (46%)	53 (44%)	12 (10%)	105 (89%)	13 (11%)

There was negligible evidence of lap effects for both criterion and signal sensitivity. There was evidence to support the hypothesis that criterion was below zero and sensitivity (*d'*) was above zero for both 1- and 2-lap serial parades; see Appendix E for a full summary. These results suggest that listeners hearing either a 1- or 2-lap serial voice parade had a liberal response criterion and displayed some ability to distinguish the signal from the noise.

Accuracy analysis. We analysed response accuracy (0 = incorrect response, 1 = correct response) using Bayesian logistic mixed models. Predictors were the main effects of number of laps (levels: 1 lap, 2 laps), target presence (levels: present, absent), and their interaction. There was negligible evidence for an effect of number of laps (BF = 0.73) and by-target presence interaction (BF = 0.59). In other words, in this sample, increasing the number of voice presentations from 1 to 2 does not seem to affect the response accuracy. As in Experiment 1, we found higher accuracy for parades that included the target voice compared to parades without target voice ($\hat{\beta}$ = -2.74, HPDI: [-3.82 – -1.72], BF > 100); see Figure 4. See Appendix G for the full results.

Figure 4. Posterior response accuracy with 95% HPDIs, Experiment 2.



Confidence ratings. As in Experiment 1, confidence ratings (0 = Not at all confident, 10 = Extremely confident), were analysed in cumulative models for ordinal data. Collapsing by target-presence, the estimated cell means for both the 1-lap ($\hat{\mu}$ = 8.19, HPDI: [7.85 – 8.5]) and 2-lap ($\hat{\mu}$ = 8.21, HPDI: [7.85 – 8.55]) were similar.

We found evidence of a positive relationship between confidence and accuracy for the 2-lap parade ($\hat{\beta}$ = .84, HPDI: [0.12 – 1.5], BF = 4.75) and negligible evidence for the 1-lap parade ($\hat{\beta}$ = .64, HPDI: [-0.02 – 1.28], BF = 2.06). In other words, participants were more confident about correct responses (than about incorrect responses) after listening to the parade twice.

Discussion

We found no evidence that identification accuracy is influenced by hearing the serial parade once rather than twice before making a decision. Similarly, we found no evidence that the criterion or signal sensitivity differed between the lap conditions. Overall, it is unlikely that the number of laps explains the sequential advantage in terms of accuracy observed by H.

M. J. Smith et al. (2020), but not in Experiment 1. The results suggest there is no advantage in hearing serial parades twice before making a decision; this may have important implications for guidance and could ultimately save police forces time when implementing voice parades.

As in Experiment 1, overall accuracy was low and false alarms were high. There was a positive association with accuracy and confidence in the 2-lap condition, where listeners were more confident about correct responses (than about incorrect responses) when listening to the parade twice. However, participants do not appear to have reliable metacognitive awareness of the difficulty of voice identification; confidence ratings were high despite low accuracy. As such, indicators of confidence are unlikely to be informative of accuracy in a way that is useful to triers of fact.

General Discussion

The current paper reports that people are less likely to select a voice from a parade when given a ‘strong’ warning which asks them to consider their responses carefully to reduce the risk of a wrongful convictions. While this is true for both serial and sequential parades, discrimination is higher when a strong warning is given prior to a sequential parade than it is when a strong warning is given prior to a serial parade. There were no differences in accuracy between serial and sequential parades. The results of Experiment 2 suggest that the sequential advantage (following a standard warning) reported in H. M. J. Smith et al. (2020) may not be clear-cut.

Our results reveal overall low accuracy, high choosing rates, and particularly error-prone performance when the target voice is absent; this is in line with existing research (Kerstholt et al., 2004; Öhman et al., 2011, 2013a, 2013b; Philippon et al., 2007; H. M. J. Smith et al., 2020). One possible reason for high error rates may be that although voices differ from each other (between-person variability), the same voice can sound very different across utterances (within-person variability; Lavan, Burton, Scott, & McGettigan, 2019). Lower target-absent accuracy might indicate a bias towards (mis)attributing differences across the encoding and parade samples to within-person variability, making people unlikely to respond, ‘not present’. However, on a target-present parade, people can extrapolate stable features across encoding and parade samples and so accuracy is higher (Kerstholt et al., 2006).

Our results highlight the potential for mitigating high error-rates on target-absent voice parades by including strong warnings. According to the cue-belief model (Leippe, Eisenstadt, & Rauch, 2009), parade decisions are informed by a sense of familiarity, and a subjective likelihood judgment about memory accuracy. Testing conditions influence which kind of information is relied upon. As an extrinsic cue, the strong warning may communicate task difficulty and encourage a reliance on the subjective likelihood of being accurate. Although participants are less likely to commit false alarms, this is because a strong warning prompts a conservative criterion shift, reducing choosing rates on both target present and target absent parades. In contrast with previous eyewitness research, sensitivity was not higher in the strong warning condition (Meissner et al., 2005). This may be because voice identification is more challenging than face identification (Barsics, 2014), and so the fidelity of familiarity cues is particularly vulnerable to disruption by a strong warning communicating task difficulty. Participants may therefore have struggled to override the inclination to reject the parade. While reducing false alarms is of course a valid priority in the context of voice identification, accurate identification of suspects is also crucial. When considering if a strong warning should be applied, the two priorities must be weighed against each other. A strong warning is a simple but effective way of safeguarding innocent suspects, who would otherwise likely only be afforded chance-level protection owing to low accuracy on target-absent voice parades (Kerstholt et al., 2004; Öhman et al., 2011, 2013a, 2013b; Philippon et al., 2007; H. M. J. Smith et al., 2020).

The present results underline the clear need for replication and thorough testing before policy recommendations are made (Malpass et al., 2008). Experiment 1 did not replicate higher accuracy for sequential compared to serial voice parades (H. M. J. Smith et al., 2020). Comparing the two sets of results raised the question of whether the sequential advantage might be affected by the number of serial laps. The results of Experiment 2 suggest that our failure to replicate the accuracy results was not due to improved performance associated with listening to serial parades once (Experiment 1) rather than twice (H. M. J. Smith et al., 2020). Rather it is more likely due to the relatively noisy data, with error-prone performance subject to a host of factors encompassed by individual differences and stimulus effects. The sequential advantage may overall be more subtle than the results of H. M. J. Smith et al. (2020) suggest, particularly as we only observed a sequential advantage in terms of discrimination following a strong warning.

Preliminary evidence from Experiment 2 suggests that there is no benefit from presenting the serial parade twice, which is consistent with conclusions drawn from the eyewitness sequential lap effect literature (Maclin & Phelan, 2007; Steblay et al., 2011; Lindsay et al., 1991; Horry et al., 2015). Even if voice representations are strengthened during the second listen, this does not improve the ability to compare the parade voices to memory of the target voice. In fact, familiarity cues to the target might even be diluted because all of the voices have been heard previously.

In the sequential procedure we implemented a strict stopping rule, in which the parade was terminated following the first ‘yes’ response. In comparison to a procedure where multiple responses are permitted, this may have harmed overall performance by prompting a conservative criterion shift (see Horry, Fitzgerald, & Mansour, 2020). Although this procedure facilitates a comparison with H. M. J. Smith et al. (2020), who also used a first-yes-counts sequential voice parade, we cannot rule out the possibility that alternative versions of the sequential parade might elicit more accurate performance. However, as false alarms are so high in voice identification, it is reasonable to test performance using a procedure designed to encourage earwitnesses to make absolute rather than relative judgments, as absolute judgments are less likely to lead to positive identifications (Dunning & Stern, 1994; Wells, 1984; Wells et al., 1998). Whilst Horry et al. (2020) found that a first-yes-counts protocol reduced the hit rate and compromised discriminability for eyewitnesses in comparison to a sequential control condition with no first-yes-counts procedure, this is not what the present study or H. M. J. Smith et al. (2020) observed for earwitnesses when comparing the sequential first-yes-counts procedure to a serial procedure. A serial procedure bears some similarities to Horry et al.'s (2020) sequential control condition: Parade members are presented one after the other, and participants encounter all parade members regardless of their decision. However, we acknowledge that these two types of procedure are not equivalent. When comparing these two types of procedure in the context of face identification, Valentine, Darling and Memon (2007) observed higher correct identifications (but a similar false alarm rate) when participants made an identification decision at the end of the parade.

We note that in practice the police would not adopt a first-yes-counts procedure because a witness may never hear the suspect speak. Whilst the present findings are important because they extend the previous earwitness literature (H. M. J. Smith et al., 2020), in future research we will adopt a more applied focus, thoroughly testing voice identification performance using alternative versions of the sequential procedure.

Overall, we did not observe evidence for a reliable positive relationship between response accuracy and confidence in the context of voice identification. Although a positive relationship between accuracy and confidence tends to be observed in the context of unfamiliar face identification (Palmer et al., 2013; Wixted & Wells, 2017), for voices the relationship is often weak or even non-existent (Kerstholt et al., 2004; Ohman et al., 2011; Olsson et al., 1998). In the current study the relationship was unreliable and varied according to testing conditions. Accuracy was generally low but the associated confidence ratings were high, perhaps because people conflate the ease of familiar voice identification with the difficulty of unfamiliar voice identification (Stevenage, 2018). In previous lab-based voice identification studies however, we note that confidence ratings have reflected uncertainty (H. M. J. Smith et al., 2020), with responses tending to fall in the middle of the scale. These differences across studies could be due to the unique characteristics of the participant samples, with our sample being part of an online panel (via [BLINDED]), having an interest in recognition, and therefore being highly motivated to respond correctly to the parade.

Our participant sample were aged between 18 and 75 years (mean 46 years). For the purposes of this research, a sample with a lower mean age might have been preferable to ensure effects were less likely to be due to age-related hearing loss (Hoffman et al., 2017). However, this would reduce generalizability of findings and exclude participants who might be earwitnesses and be asked to listen to a voice parade. We do not believe that age limits our overall conclusions, but if anything may have added additional noise.

While a larger sample size would undoubtedly have increased the power of both experiments, our sample is substantially larger than those reported historically in earwitness literature with similar designs (e.g., Perfect et al., 2002; Kerstholt et al., 2006; Phillippon et al., 2013; H. M. J. Smith et al., 2020). We argue that any effect strong enough to have sufficient practical utility to be recommended as a procedural change would have been detected with our sample. Unlike eyewitness studies where researchers have the option of running several lineups and obtaining multiple data points per participant (Mansour, Beaudry,

& Lindsay, 2017), low voice identification accuracy (Kerstholt et al., 2004, 2006; H. M. J. Smith et al., 2020) and the higher risk of interference (Stevenage et al., 2011) makes this unwise in earwitness studies. The cost of larger samples is therefore prohibitive. As put forward by Lakens (2021), our sample size was justified by both resource constraints and heuristics (the general norm followed in the literature).

Conclusion

Our results underline the value of system variable research in voice identification to support the police and legal professionals in exploring the implications of potential procedural changes. We show that the serial procedure recommended by the Home Office can be adapted to provide additional protection for innocent suspects by using pre-parade instructions that encourage more conservative response behaviour. However, such behaviour risks guilty suspects avoiding identification. While this risk may be mitigated by using a sequential parade procedure, our results suggest that this is the extent of the sequential over serial procedure advantage. We also demonstrate that there appears to be no advantage to asking listeners to listen to voices in a serial parade twice before making an identification decision.

Acknowledgments: This research was supported by the UK Economic and Social Research Council as part of the project '[BLINDED]' (ES/S015965/1) awarded to [BLINDED]. We are grateful to Daisy Thomas (University of Greenwich) for assisting with data collection. We would like to thank the participants of the University of Greenwich Face and Voice Recognition Lab volunteer participant database for making this research possible.

Declaration of interest statement: No potential competing interest is reported by the authors.

Data availability statement: Data and analysis scripts can be found on osf.io/x2dpc/?view_only=b74edc32c9494dcda5802a6efb4f0981

References

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2018). Gorilla in our Midst: An online behavioral experiment builder. *Behavioural Research Methods, 52*(1), 388-407. https://doi.org/10.3758/s13428-019-01237-x

Barsics, C., 2014. Person Recognition Is Easier from Faces than from Voices. Psychologica Belgica, 54(3), pp.244–254. http://doi.org/10.5334/pb.ap

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. Basingstoke: Palgrave Macmillan.

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology, 102*(4), 711–725. https://doi.org/10.1111/j.2044-8295.2011.02041.x

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8*(3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. *Frontiers in Psychology, 7,* 1378. https://doi:10.3389/fpsyg.2016.01378

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

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science, 23*(10), 1208–1214. https://doi.org/10.1177/0956797612441217

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80* (1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10* (1), 395–411. https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science, 2* (1), 77–101. https://doi.org/10.31234/osf.io/x8swp

Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied, 14*(2), 118–128. https://doi.org/10.1037/1076-898X.14.2.118

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

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

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. Psychological Science, 27(9), 1227-1239. https://doi.org/10.1177%2F0956797616655789

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3* (2), 186–205. https://doi.org/10.1037/1082-989x.3.2.186

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology, 54* (3), 304–313. https://doi.org/10.1016/j.jmp.2010.01.001

Dickey, J. M., Lientz, B. P., & others. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a markov chain. *The Annals of Mathematical Statistics, 41* (1), 214–226. https://doi.org/10.1214/aoms/1177697203

Dienes, Z., Coulton, S., & Heather, N. (2018). Using Bayes factors to evaluate evidence for no effect: examples from the SIPS project. *Addiction, 113*(2), 240-246. https://doi.org/10.1111/add.14002

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology, 67*(5), 818-835.

Ebbesen, E. B., & Flowe, H. D. (2002). *Simultaneous v. Sequential lineups: What do we really know?* Retrieved May 18th 2020, https://tinyurl.com/yybdephv

Fenn, K. M., Shintel, H., Atkins, A. S., Skipper, J. I., Bond, V. C., & Nusbaum, H. C. (2011). When less is heard than meets the ear: Change deafness in a telephone conversation. *Quarterly Journal of Experimental Psychology, 64*(7), 1442–1456. https://doi.org/10.1080/17470218.2011.570353

Fitzgerald, R. J., Rubinova, R., & 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. XXX-XXX). New York, NY: Routledge

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7* (4), 457–472. https://doi.org/10.1214/ss/1177011136

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC Press.

Hoffman, H. J., Dobie, R. A., Losonczy, K. G., Themann, C. L., & Flamme, G. A. (2017). Declining prevalence of hearing loss in US adults aged 20 to 69 years. *JAMA Otolaryngology–Head & Neck Surgery*, *143*(3), 274-285. https://doi.org/10.1001/jamaoto.2016.3527

Home Office. (2003). *Home Office circular 057/2003: Advice on the use of voice identification parades*. http://webarchive.nationalarchives.gov.uk/20130308000037/http://www.homeoffice.gov.uk/about-us/corporate-publications-strategy/home-office-circulars/circulars-2003/057-2003/

Horry, R., Brewer, N., Weber, N., & Palmer, M. A. (2015). The effects of allowing a second sequential lineup lap on choosing and probative value. *Psychology, Public Policy, and Law, 21*(2), 121-133. https://doi.org/10.1037/law0000041

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

Innocence Project (2020). Innocence project. http://www.innocenceproject.org

Jeffreys, H. (1961). *The theory of probability* (Vol. 3). Oxford: Oxford University Press.

Jenkins, R., Tsermentseli, S., Monks, C. P., Robertson, D. J., Stevenage, S. V., Symons, A. E., & Davis, J. P. (2021). Are super-face-recognisers also super-voice-recognisers? Evidence from cross-modal identification tasks. *Applied Cognitive Psychology, 35*(3), 590-605. https://doi.org/10.1002/acp.3813

Kerstholt, J. H., Jansen, N. J. M., Van Amelsvoort, A. G., & Broeders, A. P. A. (2004). Earwitnesses: Effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology, 18*(3), 327–336. https://doi.org/10.1002/acp.974

Kerstholt, J. H., Jansen, N. J. M., Van Amelsvoort, A. G., & Broeders, A. P. A. (2006). Earwitnesses: Effects of accent, retention and telephone. *Applied Cognitive Psychology*, *20*(2), 187–197. https://doi.org/10.1002/acp.1175

Kruschke, J. K. (2014). *Doing bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). MA: Academic Press.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15* (4), 722–752. https://doi.org/10.1177/1094428112457829

Lakens, D. (2021). *Sample size justification.* PsyArXiv. https://doi.org/10.31234/osf.io/9d3yf

Lambert, B. (2018). *A student's guide to Bayesian statistics*. Thousand Oaks, CA: Sage.

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

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, *26*(1), 90–102. https://doi.org/10.3758/s13423-018-1497-7

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

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

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.31219/osf.io/9h3et

Lindsay, R. C., Lea, J. A., & Fulford, J. A. (1991). Sequential lineup presentation: Technique matters. *Journal of Applied Psychology*, *76*(5), 741-745. https://doi.org/10.1037/0021-9010.76.5.741

Lindsay, R. C., Mansour, J. K., Beaudry, J. L., Leach, A.-M., & Bertrand, M. I. (2009). Sequential lineup presentation: Patterns and policy. *Legal and Criminological Psychology*, *14*(1), 13–24. https://doi.org/10.1348/135532508X382708

MacLin, O. H., & Phelan, C. M. (2007). PC Eyewitness: Evaluating the New Jersey method. *Behavior Research Methods*, *39*(2), 242-247. https://doi.org/10.3758/BF03193154

Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology*, *66*(4), 482–489. https://doi.org/10.1037/0021-9010.66.4.482

Malpass, R. S., & Lindsay, R. C. (1999). Measuring lineup fairness. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *13*(S1), S1–S7. https://doi.org/10.1002/(sici)1099-0720(199911)13:1+3.0.co;2-9

Malpass, R. S., Tredoux, C. G., Compo, N. S., McQuiston-Surrett, D., MacLin, O. H., Zimmerman, L. A., & Topp, L. D. (2008). Study space analysis for policy development. *Applied Cognitive Psychology*, *22*(6), 789–801. https://doi.org/10.1002/acp.1483

Mansour, J. K., Beaudry, J. L., & Lindsay, R. C. L. (2017). Are multiple-trial experiments appropriate for eyewitness identification studies? Accuracy, choosing, and confidence across trials. *Behavior Research Methods*, *49*(6), 2235-2254. https://doi.org/10.3758/s13428-017-0855-0

McAllister, H. A., Dale, R. H., & Keay, C. E. (1993). Effects of lineup modality on witness credibility. *The Journal of Social Psychology*, *133*(3), 365-376. https://doi.org/10.1080/00224545.1993.9712155

McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press.

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., 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. https://doi.org/10.1037/a0030609

Nolan, F. (2003). A recent voice parade. *International Journal of Speech, Language and the Law - Forensic Linguistics*, *10*(2), 277–291. https://doi.org/10.1558/sll.2003.10.2.277

Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language & the Law*, *16*(1). https://doi.org/10.1558/ijsll.v16i1.31

Öhman, L., Eriksson, A., & Granhag, P. A. (2011). Overhearing the planning of a crime: do adults outperform children as earwitnesses? *Journal of Police and Criminal Psychology*, *26*(2), 118–127. https://doi.org/10.1007/s11896-010-9076-5

Öhman, L., Eriksson, A., & Granhag, P. A. (2013a). Enhancing adults' and children's earwitness memory: Examining three types of interviews. *Psychiatry, Psychology and Law, 20*(2), 216–229.

Öhman, L., Eriksson, A., & Granhag, P. A. (2013b). Angry voices from the past and present: effects on adults' and children's earwitness memory. *Journal of Investigative Psychology and Offender Profiling, 10*(1), 57–70. https://doi.org/10.1002/jip.1381

Olsson, N., Juslin, P., & Winman, A. (1998). Realism of confidence in earwitness versus eyewitness identification. *Journal of Experimental Psychology: Applied, 4*(2), 101-118. https://doi.org/10.1037/1076-898X.4.2.101

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, *19*(1), 55-71. https://doi.org/10.1037/a0031602

Philippon, A. C., Cherryman, J., Bull, R., & Vrij, A. (2007). Earwitness identification performance: The effect of language, target, deliberate strategies and indirect measures. *Applied Cognitive Psychology, 21*(4), 539–550. https://doi.org/10.1002/acp.1296

Police and Criminal Evidence Act, Code D. (1984/1986). Main methods used by the police to identify people in connection with the investigation of offences and the keeping of accurate and reliable criminal records. https://www.gov.uk/guidance/police-and-criminal-evidence-act-1984-pace-codes-of-practice

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

R. v Flynn & St John. (2008). EWCA Crim 970.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12* (4), 573–604. https://doi.org/10.3758/bf03196750

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, *72* (4), 621-642. https://doi.org/10.1007/s11336-005-1350-6

Sarwar, F., Allwood, C. M., & Zetterholm, E. (2014). Earwitnesses: the type of voice lineup affects the proportion of correct identifications and the realism in confidence judgments. *International Journal of Speech, Language & the Law*, *21*(1). https://doi.org/10.1558/ijsll.v21i1.139

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., Smalarz, L., Ditchfield, R., & Ayala, N. T. (2021). Evaluating the claim that high confidence implies high accuracy in eyewitness identification. Psychology, Public Policy, and Law, 27(4), 479–491. https://doi.org/10.1037/law0000324

Smith, H. M. J., Andrews, S., Baguley, T. S., Colloff, M. F., Davis, J. P., White, D., ... & Flowe, H. D. (2021). Performance of typical and superior face recognizers on a novel interactive face matching procedure. *British Journal of Psychology*, *112*(4), 964-991. https://doi.org/10.1111/bjop.12499

Smith, H. M. J., Bird, K., Roeser, J., Robson, J., Braber, N., Wright, D., & Stacey, P. C. (2020). Voice parade procedures: Optimising witness performance. *Memory*, *28*(1), 2–17. https://doi.org/10.1080/09658211.2019.1673427

Steblay, N. K., Dietrich, H. L., Ryan, S. L., Raczynski, J. L., & James, K. A. (2011). Sequential lineup laps and eyewitness accuracy. *Law and Human Behavior*, *35*(4), 262-274. https://doi.org/10.1007/s10979-010-9236-2

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

Steblay, N. M. (1997). Social influence in eyewitness recall: A meta analytic review of lineup instruction effect. *Law and Human Behavior*, *21*(3), 283–297.

Stevenage, S. V. (2018). Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings. *Neuropsychologia*, *116*, 162-178. https://doi.org/10.1016/j.neuropsychologia.2017.07.005

Stevenage, S. V., Howland, A., & Tippelt, A. (2011). Interference in eyewitness and earwitness recognition. *Applied Cognitive Psychology*, *25*(1), 112–118. https://doi.org/10.1002/acp.1649

Valentine, T., Darling, S., & Memon, A. (2007). Do strict rules and moving images increase the reliability of sequential identification procedures? *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *21*(7), 933-949. https://doi.org/10.1002/acp.1306

Vehtari, A., Gelman, A., & Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv Preprint arXiv:1507.02646*.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27* (5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vitevitch, M. S. (2003). Change deafness: The inability to detect changes between two voices. *Journal of Experimental Psychology: Human Perception and Performance, 29*(2), 333. https://doi.org/10.1037/0096-1523.29.2.333

Vuorre, M. (2017) Sometimes I R: Bayesian Estimation of Signal Detection Models. Retrieved from https://mvuorre.github.io/posts/2017-10-09-bayesian-estimation-of-signal-detection-theory-models/

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

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

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

Wilcock, R. A., Bull, R., & Vrij, A. (2005). Aiding the performance of older eyewitnesses: enhanced non-biased line-up instructions and line-up presentation. Psychiatry, Psychology and Law, 12(1), 129–140. https://doi.org/10.1375/pplt.2005.12.1.129

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

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

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064-2072. https://doi.org/10.3758/s13414-017-1361-2

Young, A. W., Frühholz, S., & Schweinberger, S. R. (2020). Face and Voice Perception: Understanding Commonalities and Differences. *Trends in Cognitive Sciences*, *24*(5), 398–410. https://doi.org/10.1016/j.tics.2020.02.001

Appendix A: Manipulation checks

Table A1

Mean scores of post-experimental manipulation checks (standard deviation in parentheses).

	Influence of Pre-parade Instructions [1]		Consider 'Not Present' [2]	
Parade Type	Strong Warning	Standard Warning	Strong Warning	Standard Warning
Sequential	4.42 (3.15)	3.26 (2.95)	4.75 (3.35)	3.76 (3.14)
Serial (1 pass)	4.93 (3.03)	3.24 (2.7)	5.28 (3.43)	3.7 (2.7)
Serial (2 pass)	—	3.79 (2.81)	—	3.86 (3.33)

Note. [1] 'Before completing the parade, you were warned that the perpetrator may or may not be present. To what extent did this warning influence your decision(s)?' 0: The warning had no influence on my decision; 10: The warning had a strong influence on my decision.

[2] 'To what extent did you consider responding that the perpetrator was not present/responding 'no' to each voice?' 0: Did Not Consider it at all; 10: Strongly Considered it.

Appendix B: Signal detection analysis (Experiment 1)

Table B1

Estimates of the signal-detection theory model. Criterion c represents willingness to respond target present and d' indicates the signal sensitivity (Experiment 1).

	Response criterion		Signal sensitivity	
Predictor	*c* with HPDI	BF	*d'* with HPDI	BF
Main effects and Interaction				
FA warning	-0.09 [-0.17 – -0.02]	7.51	0.32 [-0.59 – 1.15]	1.1
Parade type	0 [-0.07 – 0.08]	0.39	0.71 [-0.16 – 1.59]	3.26
Parade type × FA warning	0.01 [-0.07 – 0.08]	0.39	-0.52 [-1.37 – 0.37]	1.69

Note. HPDI = Highest Posterior Density Interval; BF = Bayes Factor in support of the alternative over the null hypothesis.

Appendix C: Accuracy analysis (Experiment 1)

Table C1

Accuracy results with main effects and interactions of FA warning, Parade type, Target presence (Experiment 1).

Predictor	$\hat{\beta}$ with HPDI	BF
Main effects		
Target presence	-2.84 [-4.02 – -1.64]	> 100
Parade Type	-0.25 [-1.43 – 0.94]	0.65
FA warning	-0.94 [-2.16 – 0.2]	2.31
Interactions		
FA warning × Target presence	-0.66 [-1.76 – 0.6]	0.98
Parade type × Target presence	0.17 [-1.01 – 1.38]	0.64
FA warning × Parade type	-0.22 [-1.48 – 0.89]	0.68
Target presence × FA warning × Parade type	0.4 [-0.85 – 1.51]	0.71

Note. ß = most probable parameter value; HPDI = Highest Posterior Density Interval; BF = Bayes Factor in support of the alternative over the null hypothesis; main effects and interaction are sum coded.

Appendix D: Model comparison (Experiment 1)

We compared the signal-detection model to an unequal variance model, a model with two different variance components for the distributions of target-present and target absent trials, which are frequently used in the literature (see Wixted et al., 2016). Apart from the two variance components, all parameter values were the same. The predictive performance was compared using leave-one-out cross-validation. The out-of-sample predictive performance was determined via Pareto smoothed importance-sampling (Vehtari et al., 2015, 2017) and estimated as the expected log predictive density ($el\hat{p}d$) and the difference between the two models ($\Delta el\hat{p}d$). We found negligible evidence that would support the use of an unequal variance model: $\Delta el\hat{p}d$= -0.35 (SE = 1.30), fit of equal variance model: $el\hat{p}d$= -206.43 (SE = 10.58).

Appendix E: Signal detection analysis (Experiment 2)

Table E1

Parameter estimates of the signal-detection theory. Criterion c represents willingness to respond target present and d' indicates the signal sensitivity (Experiment 2).

	Response criterion		Signal sensitivity	
Predictor	*c* with HPDI	BF	*d'* with HPDI	BF
Main effect				
No. of laps	-0.04 [-0.11 – 0.03]	0.62	-0.17 [-0.76 – 0.44]	0.71
Cell means				
1 lap	-0.11 [-0.16 – -0.08]	-	0.76 [0.4 – 1.23]	-
2 laps	-0.15 [-0.22 – -0.11]	-	0.68 [0.21 – 1.1]	-

Note. HPDI = Highest Posterior Density Interval; BF = Bayes Factor in support of the alternative over the null hypothesis.

Appendix F: Model comparison (Experiment 2)

We compared the signal-detection model to an unequal variance model, a model with two different variance components for the distributions of target-present and target absent trials, which are frequently used in the literature (see Wixted et al., 2016). Apart from the two variance components, all parameter values were the same. Predictive performance was compared using leave-one-out cross-validation. The out-of-sample predictive performance was determined via Pareto smoothed importance-sampling (Vehtari et al., 2015, 2017) and estimated as the expected log predictive density ($el\hat{p}d$) the difference between the two models ($\Delta el\hat{p}d$). We found negligible evidence that would support the use of an unequal variance model: $\Delta el\hat{p}d$= 0.04 (SE = 0.59), fit of equal variance model: $el\hat{p}d$= -75.81 (SE = 8.43) $el\hat{p}d$

Appendix G: Accuracy analysis (Experiment 2)

Table G1

Accuracy results with main effects and interactions of FA warning, Parade type, Target presence (Experiment 2).

Predictor	$\hat{\beta}$ with HPDI	BF
Main effects and Interactions		
Target presence	2.74 [-3.82 – -1.72]	> 100
No. of laps	0.38 [-0.64 – 1.48]	0.73
No. of laps × Target presence	0.22 [-0.83 – 1.28]	0.59

Note. $\hat{\beta}$ = most probable parameter value; HPDI = Highest Posterior Density Interval; BF = Bayes Factor in support of the alternative over the null hypothesis.