Abstract

Unfamiliar voice identification is error-prone. Whilst the investigation of system variables may indicate ways of boosting earwitness performance, this is an under-researched area. Two experiments were conducted to investigate how methods of presenting voices during a parade affect accuracy and self-rated confidence. In each experiment participants listened to a target voice, and were later asked to identify that voice from a nine-person target present or target absent parade. In Experiment 1, accuracy did not vary across parades comprising 15 or 30 s sample durations. Overall, when the target was present, participants correctly identified the target voice with 39% accuracy. However, when the target was absent, participants correctly rejected the parade 6% of the time. There was no relationship between accuracy and confidence. In Experiment 2, performance with a serial procedure, in which participants responded after hearing all nine voices, was compared with a sequential procedure, in which participants made a decision after listening to each voice. Overall accuracy was higher with the sequential procedure. These results highlight the importance of system variable research in voice identification. Different methods of presenting voices have the potential to support higher levels of accuracy than the procedure currently recommended in England and Wales.

Keywords: voice parade, voice lineup, unfamiliar voice identification, system variables, earwitness

Voice parade procedures: Optimising witness performance

Witnesses who hear a perpetrator's voice, but do not see their face, may be required to try and identify the perpetrator from a voice parade. Voice identification evidence is error-prone, but can be decisive in court (Robson, 2017). In comparison to face identification, voice identification is relatively under-researched, and wide gaps in knowledge exist. The effect of system variables has been particularly neglected. Unlike estimator variables, which relate to the crime event, system variables can be controlled by police and legal professionals to try and minimise witness errors. In this paper we begin to address this oversight by investigating how voice procedures might be simplified and adapted. Specifically, we are interested in how methods of presenting voices during a parade affect both accuracy and self-rated confidence.

This paper focuses on adapting the voice parade procedure recommended in England and Wales (Home Office, 2003) but the research has international relevance. Voice parades are conducted in the US, Australia, Canada and across Europe (Broeders & Van Amelsvoort, 2001; Cantone, 2011; Laub, Wylie, & Bornstein, 2013; McGorrery & McMahon, 2017). Although voice parades may not take place as frequently as standard identity parades featuring faces, voice identification evidence can be pivotal in determining the outcome of a case (e.g. R v Khan & Bains, 2002, discussed in Nolan, 2003). In R v Khan and Bains (2002) the defendants were accused of murder by arson. A witness identified the younger defendant in a voice parade. Whilst the defence claimed there was insufficient evidence for the case to continue, the judge concluded that the voice parade provided a sufficient safeguard to allow it to continue, and the defendants were subsequently convicted. Voice identification evidence has been used over 150 times in British legal cases (Clifford, 1983). Over thirty years on, this figure is likely to be much higher. Nevertheless, there is scope for parades to be used even more widely. A recent Freedom of Information Request revealed that police forces in England and Wales vary greatly in their willingness to conduct voice parades, with some

police forces having made explicit policy decisions not to conduct them (Robson, 2017). However, if police forces are to conduct voice parades, they must be made less time-consuming to administer. At the same time, it is important that any changes support, rather than undermine, witness performance.

Problems associated with voice identification

Voice identification is error-prone, and significantly less accurate than face identification (McAllister, Dale, & Keay, 1993; Stevenage, Howland, & Tippelt, 2011). Memory for voices is particularly subject to interference (Stevenage et al., 2011), which suggests that identity-specific sound quality information is not clearly encoded, or is difficult to retrieve. This may be because the primary role of voices is to convey meaning through speech content (Fenn et al., 2011; Vitevich, 2003); attention is focused on what is being said, rather than the sound of the voice.

The results of lab-based studies employing a parade methodology underline the error-prone nature of voice identification. In a voice parade study, the participant is asked to select the target voice from amongst a number of 'foils' or distractor voices. If the participant picks out the target from a parade, this is referred to as a 'hit'. This represents a guilty suspect being selected by a witness. Although the results of previous studies vary widely, hit rates tend to be low (< 50%) or at chance-level (e.g. Kerstholt, Jansen, Van Amelsvoort, & Broeders, 2004, 2006; Ohman, Eriksson, & Granhag, 2011, 2013a, 2013b; Perfect, Hunt, & Harris, 2002). If the participant does not select the target there are three other possible outcomes. They could respond that the target is not present, therefore rejecting the parade. If the target is present this would be a 'miss', and if the target is in fact not present, this would be the correct response. Alternatively, they could select a foil, which would be a 'false alarm'. A false alarm in a real crime situation might involve an innocent suspect or a known

foil being selected. Although the latter possibility would be harmless to the suspect, it would highlight the unreliability of the witness' memory.

Previous studies have consistently found that false alarm rates are high in voice identification (Kerstholt et al., 2004, 2006; Memon & Yarmey, 1999; Philippon, Cherryman, Bull, & Vrij, 2007; Stevenage, Clark, & McNeill, 2012). For example, only one in 76 of Van Wallendael, Surace, Parsons and Brown's (1994) participants correctly responded that the target was not present. This suggests that based on memory, it is difficult to rule out voices belonging to other identities as a match. In line with the eyewitness literature on face lineups (e.g. Wells, 1984), it is also possible that high false alarm rates might be due to participants adopting a relative, rather than absolute, judgment strategy at test. Perhaps they tend to compare voices to each other, picking the voice that best matches their memory, so are unlikely to reject the parade. The alternative, an absolute judgment strategy, would be less likely to result in a positive identification because it involves comparing parade options directly to memory, rather than to each other (Dunning & Stern, 1994; Wells, 1984; Wells et al., 1998). In the eyewitness literature, relative and absolute judgments are particularly associated with simultaneous and sequential procedures respectively (Lindsay & Wells, 1985). According to the WITNESS model (Clark, 2003), identifications are based on a combination of absolute and relative judgements. One of the key assumptions of the model is that rejecting the line-up is based solely on absolute match information. Given that a participant's memory of a perpetrator is likely to be weak and error prone, particularly in respect to voice memory, this results in high false alarm rates when the target is absent. In sum, the evidence from lab-based studies therefore suggests that in a real case, not only is the witness unlikely to identify the perpetrator, but the risk of wrongful identification is high.

The confidence-accuracy relationship

Mock juror studies show that although lay people might have some knowledge of the error-prone nature of voice identification (Philippon, Cherryman, Bull, & Vrij, 2007a), jurors are likely to find voice identification evidence convincing (McAllister et al., 1993; Van Wallendael et al., 1994), and a confident witness can be persuasive (Brewer & Burke, 2002; Cutler, Penrod, & Stuve, 1988; Lindsay, Wells, & Rumpel, 1981). This is worrying, considering that previous studies suggest the relationship between confidence and unfamiliar voice identification accuracy is, at best, weak (Kerstholt et al., 2004; Ohman et al., 2011; Perfect et al., 2002; Saslove & Yarmey, 1980; Yarmey, 1995; Yarmey, Yarmey, & Yarmey, 1994). However, the majority of previous voice identification studies tend to use point biserial correlation to analyse the relationship, which risks underestimating its strength (Brewer & Wells, 2006). Recent eyewitness research has focused on confidence-accuracy calibration, an alternative method of analysing confidence-accuracy data. Such research highlights that the relationship might be more complicated, and in certain cases stronger, than previously assumed (Palmer, Brewer, Weber, & Nagesh, 2013; Wixted & Wells, 2017). Although early calibration studies suggest that the diagnosticity of confidence is more limited for unfamiliar voice identification compared to unfamiliar face identification (Olsson, Juslin, & Winman, 1998), it is not currently clear whether the relationship varies according to conditions. In line with Olsson (2000), we might expect the relationship to be weaker when the task is more difficult. Based on the eyewitness literature, we might also expect that the relationship would be stronger for choosers than non-choosers (Sauerland & Sporer, 2009; Sporer, Penrod, Read, & Cutler, 1995). This relationship has never been analysed separately in published earwitness studies, perhaps because high false alarm rates result in so few non-choosers.

Estimator variables and system variables

It is useful to distinguish between estimator and system variables when considering voice identification (Wells, 1978). The first category, estimator variables, refers to characteristics of the perpetrator, witness, or event. They have the potential to affect witness memory, and in terms of identification, they might affect hit rates, false alarm rates, or both. The majority of previous voice identification literature has focused on estimator variables. For example, it has been found that typical sounding voices are more difficult to identify than distinctive sounding voices (Mullennix et al., 2011; Sørensen, 2012; Yarmey, 1991), accuracy is higher when the initial target sample is longer (Cook & Wilding, 1997), and the listener is familiar with the language being spoken (Philippon, Cherryman, Bull, & Vrij, 2007b). Studies have focused on witness demographics, suggesting that older children (aged 11-13) are just as good as, if not slightly better than, adults at unfamiliar voice identification (Ohman et al., 2011, 2013a, 2013b). There is also some evidence for an own sex bias, with higher accuracy if the listener and speaker are the same sex (Roebuck & Wilding, 1993; Wilding & Cook, 2000).

The second category, system variables, relates to parade procedures (e.g. how the parade is conducted). They can be controlled by the legal justice system in order to minimise witness errors. It is concerning that very little previous voice identification research has focused on system variables (for exceptions, see Hollien, Bahr, & Harnsberger, 2014; Memon & Yarmey, 1999; Ohman et al., 2013b). Further research is urgently needed in order that existing procedures can be fine-tuned in a way that optimises earwitness performance. As demonstrated clearly in the eyewitness literature, system variable research points toward promising innovations with the potential to improve accuracy and the way that faces are presented leads to different patterns of performance (Brewer, Weber, Wootton, & Lindsay, 2012; Carlson, Gronlund, & Clark, 2008; Fitzgerald, Price, Oriet, & Charman, 2013; Pozzulo & Lindsay, 1999). Whilst it has previously been argued that presenting faces sequentially

reduces false alarm rates without compromising hit rates (e.g. Lindsay, Mansour, Beaudry, Leach, & Bertrand, 2009), this position has since been undermined by work showing that the simultaneous presentation of faces may in fact be preferable. Sequential presentation just makes people respond more conservatively, i.e. less likely to select a face regardless of whether the target face is present or absent in the parade (e.g. Mickes, Flowe, & Wixted, 2012). This debate underlines the importance of thorough research into the link between procedural changes and patterns of performance.

Although some of these innovations could be adapted to support voice identification, it would be unwise to assume that findings relevant to the visual modality can simply be applied to the auditory modality (Ormerod, 2001). Memory processes underlying faces and voices operate separately (Bruce & Young, 1986; Burton, Bruce, & Johnston, 1990), and faces provide stronger cues to identity than voices do (Ellis, Jones, & Mosdell, 1997; Hanley, Smith, & Hadfield, 1998). The cognitive processes involved with recognising people from faces and voices might be similar, but they are not identical (Belin, Bestelmeyer, Latinus, & Watson, 2011; Belin, Fecteau, & Bedard, 2004; Cook & Wilding, 1997; Stevenage et al., 2011). Procedures must be designed based on research focusing specifically on voice identification.

Current voice identification procedure in England and Wales

Before addressing the optimal voice parade design, it is useful to be familiar with current voice identification procedures in England and Wales. Advice for constructing parades were developed by DS John MacFarlane and Professor Francis Nolan (Home Office, 2003). The advice is not mandatory, and is still evolving in line with developments in research (McDougall, 2013). The guidelines provide valuable information about the preparation of material (e.g. foil selection). Other key recommendations include that the voice sample for each parade member should be one minute long, that the witness has to

listen to nine voices, and must listen to each of the voices at least once before making a selection (Home Office, 2003). Overall, the recommended procedure for administering the parade bears many similarities to the visual identity parade procedure outlined in Code D of the Police and Criminal Evidence Act (PACE). This implicit assumption that visual and auditory recognition operate similarly is potentially problematic. As explained above, since the original voice parade guidelines were developed, psychological research has emphasized the differences between face and voice processing. Building on the previous work of DS McFarlane and Professor Nolan, the procedures may therefore require updating.

System variables: Methods of presenting voices. In this paper we focus on two procedural aspects of the guidelines. We consider the potential for adapting the advice both in order to make the parade less time-consuming for the police, and to support witness performance. The first aspect relates to the length of the voice sample. From a practical point of view, constructing voice samples of one minute can be time-consuming for the police. The excerpts must be spliced together from different parts of the interview in which only the suspect is speaking. The time involved in doing this may increase the delay between the crime and the parade (Robson, 2017) and so risks the witness' memory degrading. Furthermore, listening to nine one-minute samples twice means that the parade itself will take around 20 minutes. Listening to each voice for a relatively long period of time (such as a minute) may enable people to build clearer auditory representations against which to compare to their memory of the perpetrator's voice. On the other hand, the high cognitive load involved in completing this task could in fact undermine performance, with people losing attention during the protracted procedure (Seale-Carlisle & Mickes, 2016). Hearing long vocal samples is unnecessary if identity information is extracted in a short period of time. Indeed, it has been shown that people can extract information about cues such as age, sex, emotion, and personality in less than a second (Bestelmeyer, Rouger, DeBruine, & Belin,

2010; McAleer, Todorov, & Belin, 2014). Such aspects can be detected in the acoustic spectrum of laryngeal voice production, but supralaryngeal features such as articulation are also relevant to voice recognition. The literature on memory for voices is not conclusive about whether duration or phonemic variety during initial exposure predicts later identification accuracy (Cook & Wilding, 1997; Roebuck & Wilding, 1993), but the importance of duration should not be ruled out because it is equally unclear what role it plays at retrieval (i.e. duration of parade samples).

The second aspect relates to the way voices are presented. Insufficient research has been undertaken to indicate whether the current procedure optimises witness performance. As shown in the eyewitness literature, false alarm rates are lower in a sequential procedure, where people are asked to make a decision after viewing each face, compared to when they look at all the faces at the same time and then make a decision (Lindsay et al., 2009). It has been argued that the sequential procedure encourages absolute judgements, whereby instead of comparing faces to each other and selecting the best match, people compare the face to their memory and only make a selection if the match is strong enough (Wells, 1984). As false alarm rates are typically high for voice identification it would be sensible to compare the serial procedure (where participants make a decision after listening to all the voices, as recommended by the Home Office) to the sequential procedure (where participants make a decision after listening to each voice).

Aims

Work to establish the best way of conducting a voice parade is long overdue. Here we focus on system variables with a view to informing the police and legal professionals about potential ways of minimising the risk of miscarriages of justice. Experiment 1 followed the Home Office procedure as closely as possible to provide a baseline against which further improvements could be made. We hypothesised that accuracy would be low, particularly for

target absent parades. While Home Office guidelines recommend the use of one-minute samples, such long samples make parades lengthy and problematic to both create and administer. Given that people can extract basic identity information from voices based on limited exposure (Bestelmeyer et al., 2010; McAleer et al., 2014), we chose shorter samples of 15 seconds (s). These samples are sufficiently long to provide acoustic variety and information about vocal idiosyncrasies that unfold over time such as speaking rate and intonation. It is however possible that longer sample durations lead to better performance as they leave a stronger memory trace (Cook & Wilding, 1997). Therefore we compared sample durations of 15 s to longer samples of 30 s. In Experiment 2 we compared the serial (Home Office) procedure to the sequential procedure. We expected that false alarm rates would be lower in the sequential procedure.

Experiment 1

This experiment follows the Home Office guidelines more closely than many previous voice identification studies, which often use 5, 6, or 7 voices instead of 9, and/or allow the participant to hear the voice only once (Cook & Wilding, 2001; Ohman et al., 2011, 2013a, 2013b; Perfect et al., 2002; Philippon et al., 2007b; Sørensen, 2012). One exception is McDougall, Nolan and Hudson (2015), who did use the exact Home Office procedure. However, their design encouraged intentional encoding. Before listening to the target voice, participants were told that they would later be asked to recognise it. Arguably, most crime situations involve incidental encoding; witnesses are likely to be focusing on the event rather than considering the possibility of having to attempt a subsequent identification. The two types of encoding affect memory performance differently (Haese & Czernochowski, 2015). In this experiment participants did not know that there would be an identification test until it happened. We compared samples of 15 s and 30 s rather than one minute for practical reasons.

Method

Design. This was a 2 x 2 between subjects factorial design. The factors were sample duration (15 s or 30 s) and target presence (present or absent). The dependent variables were identification accuracy and self-rated confidence.

Participants. There were 92 participants (69 female, 23 male), with an age range of 18-70 years (M = 32.34, SD = 12.25), and normal or corrected-to-normal hearing. Ethical approval for the experiment was granted by the [X] University's Business, Law and Social Science College Research Ethics Committee.

Apparatus and materials. The stimuli were taken from the Dynamic Variability in Speech Database (DyViS) (Nolan, McDougall, De Jong, & Hudson, 2009; available to download from the Economic and Social Data Service). This database features .wav files of 100 male speakers of Standard Southern British English performing a variety of spoken tasks, including a mock police interview. All speakers are aged between 18-25. As reported by Nolan et al. (2009), the speakers were recorded in a sound-treated booth using a Marantz PMD670 portable solid-state recorder and a sampling rate of 44.1 kHz. We randomly selected 30 speakers from the database, and using Audacity (version 2.2.1), measured the average fundamental frequency (F0) of interviewees' voices. For speech, F0 refers to the lowest frequency of the vibrations produced by the vocal folds, and F0 is closely related to the perception of pitch. The voices were put into three groups according to F0 (low, medium and high). We used F0 to group the voices because pitch plays a key role in voice similarity (Nolan, McDougall, & Hudson, 2011). From each group of 10 speakers we constructed a target present and target absent parade, so that overall there were three versions of the target present parade, and three versions of the target absent parade. In the target present parade, target position was counterbalanced. The target either appeared in an early position (position 3), or a late position (position 7). In group 1, the average F0 across the 10 speakers was 90 Hz

(*SD* = 7). The F0 of the target's voice was 88 Hz, while for the replacement foil it was 92 Hz. In group 2, the average F0 was 103 Hz (*SD* = 7). The F0 of the target's voice was 107 Hz, and for the replacement foil's voice it was 103 Hz. In group 3, the average F0 was 118 Hz (*SD* = 8). The F0 of the target's voice was 111 Hz, and the replacement foil's voice was 129 Hz. The speech samples for the three targets were taken from the recording of a telephone conversation, during which a crime was discussed. After editing, each recording was 30 s long, and only featured one side of the conversation. The content was similar across all three targets, because each of the speakers were responding to the same questions. As real voice parade samples are often made up of police interview excerpts (Home Office, 2003), the speech samples for the voice parade were taken from the mock police interview recordings. The parade voices were edited using Audition (version 2.2.1). Sections featuring the voice of the interviewer were deleted, and excerpts featuring the interviewees were spliced together to produce 15 s and 30 s samples. The 15 s and 30 s samples featured different spoken material. The voice samples were all taken from different sections of the interview, selected at random, so the content of speech differed across speakers. Furthermore, the content of the interview samples did not overlap with the content of the phone recording.

Pilot testing was undertaken to check for parade fairness for each of the parades. Five participants listened to the target voices and provided a description of the voice. They were asked how old they thought the person was, whether he had an accent, and if yes, what the accent was like. They were also asked to rate the voice pitch on a scale of 1 – 7 (1 = *low*, 7 = *high*), and the rate of speech on a scale of 1 – 7 (1 = *very slow*, 7 = *very fast*). Finally, they were asked whether the voice had distinctive features, and if so, what they were. From the resulting descriptions, a modal description was produced for each target voice. One group of participants (*N* = 34) acted as mock witnesses for the three target present parades, while another group (*N* = 31) acted as mock witnesses for the three target absent parades. The

participants read one of the modal descriptions of the target, then listened to the corresponding parade, attempting to identify the perpetrator from the description. This procedure was repeated for each of the three target present parades and each of the three target absent parades. We calculated Tredoux's *E* to provide a measure of the number of parade members fitting the description of the perpetrator (Tredoux, 1998). For parade 1, Tredoux's *E* was 7.22 (target present) and 5.19 (target absent). For parade 2 it was 5.90 (target present) and 3.80 (target absent). For parade 3 it was 5.50 (target present) and 4.35 (target absent). Overall these results show that amongst the foils there were several viable alternatives for each target. We also calculated whether there was a bias towards the target in each of the parades (Malpass & Lindsay, 1999). In each parade, the proportion of mock witnesses selecting the target did not differ from chance. That is, the critical ratios (parade 1: 0.12; parade 2: 1.71; parade 3: -2.82) did not exceed 1.96.

During the retention interval, participants completed a wordsearch containing words for different types of fruit (http://www.wordsearch-puzzles.co.uk). At the same time, they listened to a recording of ambient noise, which was made in a public lobby and featured unintelligible speech sounds.

The experiment ran on PsychoPy version 1.85.1 (Peirce, 2009), and all recordings were played binaurally through Sennheiser (HD205) headphones. The sound intensity was constant across participants. The volume was measured using a Svantek (977) sound level meter, and the headphones were places over a G.R.A.S. (RA0039) artificial ear simulator. The volume ranged between 65-75 dB.

Procedure. The participants were randomly allocated to a parade (1, 2 or 3) and a condition using an online research randomizer (Urbaniak & Plous, 2013).

Each participant completed a single trial. They were not told that they would have to complete a voice parade, but rather that they were being invited to take part in an experiment

about voice perception, and that after listening to a voice recording they would be asked questions about what they had heard. Participants listened to the 30 s sample of the target voice, then completed a wordsearch for five minutes whilst listening to the ambient noise. They were instructed to try and find as many words as they could. At the end of the five minutes, instructions for the parade were provided. Participants were told that they would hear nine voices, and that after they had listened to the parade twice, they would be asked to try and identify the perpetrator. They were informed that the perpetrator may or may not be present. The voice number was visible at the same time as the voice was playing. Participants had to listen to the whole of the recording and were not permitted to end the recording early. Both times the parade was heard, the voices were presented in the same order. Participants registered their response after hearing all the voices twice, by pressing a number (1-9) on the keyboard. They were told to press '0' if they thought the perpetrator was not present. No time limit was imposed on their decision. After making a selection, participants were asked, '[o]n a scale of 0-10, how confident are you that you have given the correct answer?' (0 = *not at all confident*, 10 = *extremely confident*). They were then debriefed.

Results

The data were analysed using the *R* package *rstanarm* (Gabry & Goodrich, 2016; R Core Team, 2016) to run Bayesian linear mixed effects models (Gelman et al., 2014; Kruschke, 2014; McElreath, 2016). The advantages of using a Bayesian framework for hypothesis testing are well documented in the literature (Kruschke, 2014; Kruschke, Aguinis, & Joo, 2012; Lambert, 2018; Nicenboim, & Vasishth, 2016; Sorensen, Hohenstein, & Vasishth, 2015). Accounting for the variance associated with stimuli is crucial in experimental psychology. Linear mixed models can be used to account for parade specific variance by treating it as random effect (intercept) with adjustments for conditional factors that might vary for individual parades by treating these as random slopes (details on the

model specification can be found below) (Baayen, Davidson, & Bates, 2008). Frequentist, as opposed to Bayesian, linear mixed models are notoriously subject to over-parameterisation, as indicated by convergence failure for maximal random effects structures (Barr, Levy, Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2015). This is usually related to an imbalance of model complexity and sample size and is not suitable for statistical inference. This, however, is not a problem in Bayesian settings because the models converge as the number of iterations approach infinity. Further, Bayesian models allow us to use posterior distributions to test our research hypotheses. All *R*-scripts and data can be accessed on GitHub (https://github.com/jensroes/Voice-line-up).

Descriptive analysis. The descriptive data for the response accuracy (in %) and the confidence ratings can be found in Table 1.

[INSERT TABLE 1 ABOUT HERE]

For the target present parades, the correct target identifications (hits), the incorrect rejections (misses), and false target identifications (false alarms) are shown in Table 2 with the proportions shown in parentheses.

[INSERT TABLE 2 ABOUT HERE]

Response accuracy. The accuracy data were analysed as binary responses (0 = inaccurate, 1 = accurate) using Bayesian linear mixed effects models with binomial link function. Models were fitted with maximal random effects structure (Barr et al., 2013; Bates et al., 2015); random intercepts were included for different parades (1, 2, 3) with slope-adjustments for target order (3 or 7), target presence (present, absent), sample duration (15 s, 30 s) and all interaction terms. Model predictors – target presence, sample duration and their interaction terms – were added incrementally to the intercept only model.

Statistically relevant effects were evaluated via model comparisons using out-of-sample predictions estimated using Pareto smoothed importance-sampling leave-one-out

cross-validation (PSIS-LOO) (Vehtari, Gelman, & Gabry, 2015, 2017). Predictive performance of the fitted models was estimated as the sum of the expected log pointwise predictive density (*elpd*) along with its standard error (*SE*). In other words, a higher *elpd* indicates that the model has better predictive performance than another model. The difference between the predictive quality of the models, and therefore statistically relevant effects, was expressed as *Δelpd* (shown with standard errors [*SE*] of the difference). This means that a negative difference *Δelpd* indicates that a model has a reduced predictive performance compared to another model.

The result of the model comparisons can be found in Table 3. The model with the main effect for target presence showed the highest predictive performance. Model comparisons showed that adding the main effect of target presence to the intercept-only model increased the predictive performance (*Δelpd* = -3.94, *SE* = 2.10). All other models rendered a lower predictive performance. Therefore, our inference is drawn from the statistical model with the main effect for target presence.

[INSERT TABLE 3 ABOUT HERE]

The posterior (statistically inferred) distribution of these models was used to derive the *maximum a posteriori*, the most probable value of the true (unknown) effect of interest $\hat{\mu}$ and the 95% Highest Posterior Densitiy Interval (henceforth, 95% HPDI), the shortest interval containing 95% of the posterior probability mass as indication of the certainty range in which the true (unknown) parameter value lies. HPDIs were used, as opposed to percentile intervals (also known as credible intervals), because HPDIs are more suitable for non-symmetric posteriors (Hyndman, 1996; Liu, Gelman, & Zheng, 2015), for example, bimodal or skewed posterior distributions as found in the results presented below.

The posterior distribution for the predictor target presence revealed that responses are more accurate by $\hat{\mu}$ = 25.98% for parades that included the target compared to parades that did not include the target (95% HPDI [14.49%, 38.65%]).

Chance performance for parades is 10%. The 95% HPDI for target absent parades contained 10% chance-level performance (95% HPDI [0.43%, 13.66%]) and is thus not different from chance. Accuracy for target present parades was above chance (95% HPDI [22.11%, 55.50%]).

The posterior distribution of the interaction model was used to assess chance-level performance for each sample duration type. Figure 1 illustrates the posterior probability distributions for each sample duration. Using the re-parameterised posterior distributions derived from the statistical model allows direct statistical inference. For each sample duration we observe a bimodal distribution representing target absent and target present trials, displayed in red and green, respectively. Horizontal bars indicate the 95% HPDI. For both sample durations we can see that the 95% HPDI contains chance-level performance (10%)[1]. From these posterior distributions we can infer that the posterior probability mass below chance-level (10%) is 44.72% for sample durations of 15 s and 34.67% for 30 s sample duration. So descriptively speaking, for the 30 s sample duration, 10.05% fewer responses were below chance-level.

[INSERT FIGURE 1 ABOUT HERE]

Figure 2 shows the posterior probability intervals of the inferred accuracy values with comparisons against chance-level performance for each condition. From these posterior distributions we can infer the posterior probability mass below chance-level (10%). For target absent parades we found a below chance-level performance of 88.70% for sample durations

[1] Note that we used the interaction model rather than the model with the highest predictive performance to display the posterior probability ranges of all conditions, including each sample duration, and to allow for variation between the levels of each factor.

of 15 s and 67.90% for 30 s sample duration. For target present parades we found a below chance-level performance of 0.73% for sample durations of 15 s and 1.43% for 30 s sample duration.

[INSERT FIGURE 2 ABOUT HERE]

Confidence ratings. Confidence ratings were analysed in Negative Binomial mixed effects models. Models were fitted with maximal random effects structure (Barr et al., 2013; Bates et al., 2015); random intercepts were included for parade items with slope-adjustments for target order, target presence (present, absent), sample duration (15 s, 30 s), accuracy (inaccurate, accurate) and all interaction terms.

Confidence ratings were modeled with predictors accuracy, target presence, sample duration and all interaction terms. Model predictors were added incrementally to the intercept only model. The results of the model comparisons can be found in Table 4. We found that adding main effects to the intercept-only model did not improve the predictive performance of the model. The intercept-only model was found to have the highest predictive performance suggesting that confidence ratings remained consistent across accuracy and target presence.

[INSERT TABLE 4 ABOUT HERE]

The posterior distribution for the confidence ratings showed a most probable value of $\hat{\mu} = 5.61$ (95% HPDI [4.92, 6.21]). This shows that participants' confidence about their response accuracy was generally neither low or high (confidence scale 0 to 10) and importantly, confidence ratings remained consistent across target present and target absent parades, response accuracy, and sample duration. We would have liked to explore whether the relationship between confidence and accuracy was stronger in choosers (Sauerland & Sporer, 2009; Sporer et al., 1995), but false alarm rates were high and there were so few non-choosers that it was not possible to run separate analyses.

Discussion

Overall accuracy was at chance-level, but consistent with performance observed in previous voice parade studies (Kerstholt et al., 2004, 2006; Ohman et al., 2011, 2013a, 2013b; Perfect et al., 2002). Target present responses were above chance-level (39% accuracy), with the most common type of error being a foil identification. In target absent parades, participants rarely correctly responded that the target was not present (6% accuracy). As has been found in previous studies (Kerstholt et al., 2004, 2006; Memon & Yarmey, 1999; Philippon et al., 2007; Stevenage et al., 2012; Van Wallendael et al., 1994), performance was significantly poorer on target absent parades compared to target present parades. This suggests that although there was a memory trace for the target, the memory was not sufficiently well encoded for participants to respond 'not present'. In other words, their representation of the target voice is broad, and tolerates sufficient within-person variability that it can incorporate voices that belong to a different identity. On the other hand, high false alarm rates may reflect that the serial presentation of voices encourages a relative judgment strategy (Dunning & Stern, 1994; Wells, 1984; Wells et al., 1998). There was no evidence for an effect of sample duration, even though 10% fewer 30 s responses were below chance level. This may be because the amount of identity information included in the 15 s and 30 s samples does not differ, or at least that it does not differ in the extent to which it activates the auditory representation for the target voice. Alternatively, if additional useful identity information *is* provided in the 30 s sample, the benefit of this may be offset by the high cognitive load and/or added interference associated with listening to longer samples of the foil voices. Although for practical reasons we compared 15 s and 30 s excerpts rather than 1 minute (as recommended by the Home Office), there is no reason to believe that an effect would appear at longer durations. This does not undermine our argument that longer excerpts are unnecessary.

Across conditions, participants were not particularly confident or unconfident, and did not register the difficulty and likely inaccuracy associated with target absent parades. Their confidence was not related to accuracy, and it did not vary across conditions. This finding sits well with other research predicting that the relationship between confidence and accuracy in unfamiliar voice identification is weak (Kerstholt et al., 2004; Ohman et al., 2011; Perfect et al., 2002; Saslove & Yarmey, 1980; Yarmey, 1995; Yarmey et al., 1994).

As the overall patterns of performance are consistent with previous research using different voice identification procedures, these results alone should not be viewed as reflecting negatively on the procedure currently recommended by the Home Office.

Experiment 2

Experiment 1 used a serial procedure, currently recommended by the Home Office. After they had heard each of the parade voices twice, the participants were asked which (if any) of the voices belonged to the target. False alarm rates were high, and overall accuracy was relatively low. As previous eyewitness research suggests that different identification procedures are likely to lead to different patterns of performance (Brewer et al., 2012; Carlson et al., 2008), in Experiment 2 we tested whether the same was true for voices. From a forensic point of view, the worst possible outcome from an identification procedure is that an innocent suspect is selected. We focused on the sequential procedure, in which participants make ‘yes’/‘no’ decisions after hearing each voice, because there is some evidence in the eyewitness literature that it may prompt an absolute judgment strategy and so reduce false alarm rates (Wells, 1984). Although previous studies have used sequential versions of a voice parade, and also observed high false alarm rates (Stevenage et al., 2012; Zetterholm, Sarwar, & Allwood, 2009), ours is the first to directly compare this procedure to the serial procedure. We expected that patterns of performance would vary between procedures, but that false alarm rates would be higher using the serial procedure.

Methods

Apart from the following exceptions, the materials and methods were identical to Experiment 1.

Design. The factors were parade type (serial or sequential) and target presence (present or absent).

Participants. There were 91 participants (67 female, 24 male), with an age range of 18-45 years (M = 20.80, SD = 4.40).

Procedure. As we did not find evidence for an effect of sample duration in Experiment 1, all of the parade samples were 15 s long. In the sequential condition, before completing the parade, participants were instructed that they were going to listen to a series of voices to try and identify the perpetrator, and that after listening to each voice once they would be asked whether the voice belonged to the perpetrator. They were not told how many voices they would hear in total, but that only their first 'yes' response would count. After listening to each voice for the full 15 s, participants responded by pressing 'Y' (for yes) if they thought the voice belonged to the perpetrator, and 'N' (for no) if they thought it did not. They were informed that the perpetrator may or may not be present. After listening to, and responding to, all of the nine voices in the parade, participants were asked, '[o]n a scale of 0-10, how confident are you that you have given the correct answer?' (0 = *not at all confident*, 10 = *extremely confident*).

Results

The data were analysed in exactly the same way as Experiment 1. All *R*-scripts and data can be accessed on GitHub (https://github.com/jensroes/Voice-line-up).

Descriptive analysis. The descriptive data for the response accuracy (in %) and the confidence ratings can be found in Table 5.

[INSERT TABLE 5 ABOUT HERE]

For the target present parades, the correct target identifications (hits), the incorrect rejections (misses), and false identifications (false alarm) are shown in Table 6 with proportions shown in parentheses.

[INSERT TABLE 6 ABOUT HERE]

Response accuracy. Random intercepts were included for different parades (1, 2, 3) with slope-adjustments for target order (3 or 7), target presence (present, absent), parade type (serial, sequential) and all interaction terms. Model predictors – target presence, parade type and their interaction terms – were added incrementally to the intercept only model and compared using PSIS-LOO.

The model comparisons can be found in Table 7. The model with main effects for target presence and parade type showed the highest predictive performance compared to the intercept-only model ($\Delta elpd$ = -48.16, SE = 6.07) also when added. Therefore, our inference is drawn from the statistical model with simple main effects for target presence and parade type.

[INSERT TABLE 7 ABOUT HERE]

The posterior distribution for the predictor target presence revealed that responses are more accurate by $\hat{\mu}$=10.06% for parades that included the target compared to parades that did not include the target (95% HPDI [6.47%, 13.38%]). The posterior response accuracy for sequential presentations showed a higher accuracy by $\hat{\mu}$=15.74% compared to serial presentations (95% HPDI [11.93%, 18.79%]).

Chance performance for parades is 10%. From these posterior distributions we can infer whether or not the HPDI contains chance-level performance and the posterior probability mass that is below chance-level. In target absent parades, the 95% HPDIs show that responses for sequentially presented parades contained chance-level performance (95% HPDI [5.36%, 36.69%]) with 11.57% of the posterior distribution below chance. Target

absent serial parades contain chance-level (95% HPDI [0.01%, 17.96%]) with 71.47% of the posterior probability mass below chance-level. In target present sequential parades, responses were consistently above chance-level (95% HPDI [15.34%, 55.23%]) with only 0.07% of the posterior probability below chance. Target present serial parades were at chance-level (95% HPDI [4.68%, 33.81%]) with 15.40% of the posterior probability below chance-level performance.

The posterior distribution of the interaction model was used to assess chance-level performance for each parade type. Figure 3 illustrates the probability distributions for each parade type. For each parade type we observe a bimodal distribution representing target absent and target present trials, displayed in red and green, respectively. Horizontal bars indicate the 95% HPDI. For both parade types we can see that the 95% HPDI contains chance-level performance (10%)[2]. From the posterior distribution we can infer that the posterior probability mass below chance-level (10%) is 43.43% for serial parade types but only 5.82% for sequential parades, revealing a better performance for the latter by 37.62% less responses below chance-level.

[INSERT FIGURE 3 ABOUT HERE]

Figure 4 shows the posterior probability intervals of the inferred accuracy values with comparisons against chance-level performance for each condition.

[INSERT FIGURE 4 ABOUT HERE]

Confidence ratings. Confidence ratings were analysed in Negative Binomial mixed effects models, just as they were in Experiment 1. Models were fitted with maximal random effects structure (Barr et al., 2013; Bates et al., 2015); random intercepts were included for different parade items with slope-adjustments for target order, target presence (present,

[2] Note that we used the interaction model rather than the model with the highest predictive performance to display the posterior probability ranges of all conditions, including each parade type, and to allow for varying differences across the factor levels.

absent), parade type (serial, sequential), accuracy (inaccurate, accurate) and all interaction terms.

Confidence ratings were modeled with the predictors accuracy, parade type, target presence, and all interaction terms. Model predictors were added incrementally to the intercept-only model. The model comparisons can be found in Table 8. Model comparisons revealed that no predictor improved the predictive performance of the model. Therefore, the intercept-only model was found to have the highest predictive performance suggesting that confidence ratings remained consistent across accuracy, target presence, and parade type.

[INSERT TABLE 8 ABOUT HERE]

The most probable posterior value for confidence ratings was $\hat{\mu}$=5.40 (95% HPDI [4.87, 5.97]) showing that the participants' confidence about their response accuracy was generally neither low nor high (confidence scale 0 to 10). Importantly, confidence ratings remained stable across target present and target absent parades, response accuracy, and parade type.

Discussion

As in Experiment 1, overall performance was low, and performance was higher in the target present condition. Only target present sequential parades were significantly above chance-level. Overall chance-level performance in the serial parade, and the main effect of target presence replicate the results of Experiment 1. However, in Experiment 2, performance in the target present condition was at chance-level. We expected that accuracy would be higher on target absent parades in the sequential condition, but not that there would be a main effect of parade type and that accuracy would be higher overall in the sequential condition. The sequential procedure does not appear to simply make participants more conservative as we might have expected based on the eyewitness literature (Mickes et al., 2012). If it had, lower accuracy in the target present condition would have been observed. These results are

therefore consistent with the conclusion that the sequential procedure improves discriminability. Perhaps interference is a particular problem in serial parades when the voices are heard in succession. Having to make a decision after each voice as in the sequential procedure might mitigate the effect of interference by demarcating each voice. In this way, it offers the participant an opportunity to disregard any previously heard voices, focus only on the one being heard, and therefore to respond more accurately. The confidence-accuracy results replicate Experiment 1. Confidence was neither high nor low, was not related to accuracy, and did not vary across conditions.

General Discussion

In two experiments we focused on the effect of system variables on voice identification accuracy, testing how different voice parade procedures affect accuracy and self-rated confidence. The results underline the error-prone nature of voice identification, particularly in parades when the target is not present. We support previous calls for unfamiliar voice identification evidence only to be admitted with caution (Ohman et al., 2013a; Ormerod, 2001). That said, we have shown that different methods of presenting voices have the potential to support accuracy to a greater extent than the procedure currently recommended by the Home Office. Whilst there was no evidence for an effect of sample duration (Experiment 1), there was an effect of parade type (Experiment 2), with participants in the sequential condition responding more accurately. An important consideration from an applied point of view is whether a witness' confidence is diagnostic of accuracy. We did not observe a relationship between confidence and accuracy in either Experiment 1 or 2.

Accuracy

Error-rates exceeded 60% in every condition across Experiment 1 and 2. This was a difficult task, perhaps made more difficult because the voices in each parade were similar in terms of pitch (Nolan et al., 2011), and because the initial speech sample was relatively short

(30 s) (Cook & Wilding, 2000). Nevertheless, the overall accuracy in both experiments sits within the range observed in previous studies (Kerstholt et al., 2004, 2006; Ohman et al., 2011, 2013a, 2013b; Perfect et al., 2002). In both experiments there was a main effect of target presence, with lower accuracy on target absent parades. This is unlikely to be due to participants adopting a relative decision strategy (Dunning & Stern, 1994; Wells, 1984; Wells et al., 1998). The effect was observed in both the serial (Experiment 1 and 2) and the sequential condition (Experiment 2) and, if anything, the latter should prompt an absolute rather than relative decision strategy (Wells, 1984). The differences in target present and target absent performance may be due to participants making use of different cues across the two conditions. Whilst some voice features are stable, others vary, and the same person can sound very different across utterances and occasions (Holmberg, Hillman, Perkell, & Gress, 1994; Hammersley & Read, 1996). Kerstholt et al. (2006) argue that in target present conditions people are able to isolate stable, more diagnostic information. When the target is absent, and it is not possible to do this, less diagnostic information becomes more influential, driving the bias towards positive identifications. However, foil identifications were also high in target present conditions (Experiment 1 and 2), so the bias towards positive identification does not only manifest when the target is absent. The overall difficulty of voice identification likely contributes to this bias (Bruer, Fitzgerald, Therrien, & Price, 2015; Stepan, Dehnke, & Fenn, 2017).

The outlook may appear bleak for unfamiliar voice identification accuracy. However, other aspects of the results we present are a cause for optimism. From an applied point of view, observing no difference in sample durations of 15 s and 30 s is reassuring. Constructing voice samples for a parade can be extremely time-consuming. Samples often consist of excerpts from police interviews, and finding sufficient material poses a challenge. It would

appear that sample duration may not be as crucial as the Home Office guidelines (Home Office, 2003) suggest.

A further cause for optimism is that, as predicted from previous eyewitness literature findings (Brewer et al., 2012; Carlson et al., 2008; Fitzgerald et al., 2013; Pozzulo & Lindsay, 1999), system variable research offers great potential for improving earwitness performance. Until now, this area of research that has been somewhat overlooked by earwitness researchers (for exceptions see Hollien et al., 2014; Memon & Yarmey, 1999; Ohman et al., 2013a). Performance was significantly more accurate when tested using the sequential procedure. This may seem surprising considering that in the serial condition participants were able to listen to each of the voices twice, but only once in the sequential condition. The key features of sequential procedures that support performance are not clear. There are many differences between the serial and sequential procedure, and the effect of each one requires isolating and comparing before the key feature(s) can be identified. However, as memory for voices is particularly sensitive to interference (Stevenage et al., 2011), it seems feasible to hypothesise as a starting point that the sequential procedure helps by demarcating voices and therefore enabling people to make a decision about each voice on its own merits.

Confidence

As expected based on previous studies (Kerstholt et al., 2004; Ohman et al., 2011; Perfect et al., 2002; Saslove & Yarmey, 1980; Yarmey, 1995; Yarmey, Yarmey, & Yarmey, 1994), we found no evidence for a relationship between confidence and accuracy in either Experiment 1 or 2. Although we might have expected that the relationship between confidence and accuracy would be stronger when the task was easier (Olsson, 2000), there was no relationship even in the sequential target present condition, which descriptively speaking elicited the highest level of accuracy across the two experiments. Confidence did not vary across conditions, and participants consistently registered their confidence in the

middle of the scale (neither confident nor unconfident), regardless of whether their performance was at chance-level or above. Therefore, these results suggest that if participants are using strategies varying in effectiveness across target present and target absent conditions (Kerstholt et al., 2006), or serial and sequential conditions (Dunning & Stern, 1994; Wells, 1984; Wells et al., 1998) they do not have metacognitive awareness of doing so. Equally, it is feasible that low variance in responses and a tendency to register confidence in the middle of the scale might preclude the opportunity for a confidence-accuracy relationship to emerge. Unfortunately, descriptive confidence data has not been reported in detail in previous voice identification studies. However, it would certainly be interesting to know whether low variance should be expected, and if so, whether this might help to explain the lack of a relationship between confidence and accuracy.

The results underline the importance of being cautious about admitting voice identification evidence. As jurors are likely to find voice identification evidence convincing (McAllister et al., 1993; Van Wallendael et al., 1994), particularly when the witness is confident (Brewer & Burke, 2002; Cutler, Penrod, & Stuve, 1988; Lindsay, Wells, & Rumpel, 1981), they should be explicitly warned that witness confidence can be misleading.

Limitations and further research

The forensic implications of research into system variables in voice identification are significant, and further work should be undertaken as a matter of urgency. As such, future earwitness studies should strive for ecological validity, simulating the conditions of real voice parades as far as possible to maximise generalisability. The experiments conducted here used a forensically-oriented database (DyViS, Nolan et al., 2009), meaning that all stimuli were comparable to what an earwitness in an actual case would hear. Furthermore, care was taken to construct parades consisting of similar sounding voices (Home Office, 2003), and each parade was subjected to mock-witness testing (Tredoux, 1998). We do however acknowledge

that the retention interval was only five minutes long. Although similarly short retention intervals of 30 minutes or less are widely used in the earwitness and eyewitness literature (e.g. Brewer et al., 2012; Perfect et al., 2002; Philippon et al., 2007b; Seale-Carlisle & Mickes, 2016), more realistic intervals of days/weeks should be used in future research in order to thoroughly test procedures (Ohman et al., 2011). Nevertheless, even employing a retention interval of five minutes means that the participant cannot rely on sensory or short-term memory stores (Atkinson & Shiffrin, 1971; Lu, Williamson, & Kauffman, 1992), just as they would not be able to do during an actual voice parade.

It should be noted that the sample of participants used in each experiment differed in terms of age. In Experiment 1, the age range was wider, and the mean age was 10 years older than that of the sample in Experiment 2. Whilst the results of Experiment 2 broadly replicated Experiment 1 both in terms of overall accuracy and confidence, we cannot rule out the possibility that age might interact with the factors tested here. We recommend that future research on system variables addresses age effects because in an actual case, witnesses would be drawn from a broad sample of ages encompassing children and older adults. Age affects auditory acuity (Hoffman, Dobie, Losonczy, Themann, & Flamme, 2017), as well as parade decision-making processes, with children and older adults being particularly likely to commit false alarms on target absent face parades (Pozzulo & Lindsay, 1998; Searcy, Bartlett, & Memon, 1999).

In Experiment 1, serial target present performance was above chance (39%), but it was at chance-level in Experiment 2 (17%). It is worth considering whether the differences in ages across samples may have played a role in this inconsistency. The mean age of participants in both experiments was below 40, the age at which auditory acuity starts to degrade (Hoffman et al., 2017), so this is unlikely to have played a role. Indeed, the participants in Experiment 2 were drawn mostly from the undergraduate student population

and were younger. It is however possible that these participants may have been less motivated and therefore less likely to respond accurately. Nevertheless, this does not undermine evidence that the sequential parade is superior to the serial parade, as participants in this sample were randomly allocated to conditions.

Conclusion

This research addresses a gap in the under-researched area of voice identification, investigating ways of adapting voice parade procedures both to make them easier to conduct, and to support earwitness performance. We have shown that the procedure recommended by the Home Office may not be ideal. Not only is there scope to reduce the length of the overall parade, but accuracy is higher when the voices are presented in a different way. Our results also demonstrate that jurors should be sceptical about witness confidence when weighing up voice identification evidence. This is an extremely promising avenue of research, with global impact. Although we have focused here on the procedure used in England and Wales, voice parades are conducted all over the world. Any improvement in parade procedure has the potential to increase conviction rates and to reduce the real risk of miscarriages of justice in cases involving voice identification.

Acknowledgements: We are grateful to Evan Allen, who assisted with the data collection for Experiment 2.

References

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American, 225*(2), 82-91. doi: 10.1038/scientificamerican0871-82

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412. doi: 10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278. doi: 10.1016/j.jml.2012.11.001

Bates, D. M., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). Parsimonious mixed models. *arXiv Preprint arXiv:1506.04967*.

Belin, P., Bestelmeyer, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology, 102*(4), 711-725. doi: 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. doi: 10.1016/j.tics.2004.01.008

Bestelmeyer, P. E., Rouger, J., DeBruine, L. M., & Belin, P. (2010). Auditory adaptation in vocal affect perception. *Cognition, 117*(2), 217-223. doi: 10.1016/j.cognition.2010.08.008

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior, 26*(3), 353-364. doi: 10.1023/A:1015380522722

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. doi: 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. doi: 10.1177/0956797612441217

Broeders, A., & van Amelsvoort, A. (2001). A practical approach to forensic earwitness identification: constructing a voice line-up. *Problems of Forensic Sciences*, *47*, 237-245.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305-327. doi: 10.1111/j.2044-8295.1986.tb02199.x

Bruer, K. C., Fitzgerald, R. J., Therrien, N. M., & Price, H. L. (2015). Line-up member similarity influences the effectiveness of a salient rejection option for eyewitnesses. *Psychiatry, Psychology and Law*, *22*(1), 124-133. doi: 10.1080/13218719.2014.919688

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, *81*(3), 361-380. doi: 10.1111/j.2044-8295.1990.tb02367.x

Cantone, J. A. (2010). Do you hear what I hear: Empirical research on earwitness testimony. *Texas Wesleyan Law Review*, *17*, 123-142.

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. doi: 10.1037/1076-898X.14.2.118

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

Clifford, B. R. (1983). Memory for voices: The feasibility and quality of earwitness evidence. In S. M. A. Lloyd-Bostock and B. R. Clifford (Eds.), *Evaluating witness evidence*. Chichester, Great Britain: John Wiley & Sons

Cook, S., & Wilding, J. (1997). Earwitness testimony: Never mind the variety, hear the length. *Applied Cognitive Psychology*, *11*(2), 95-111. doi: 10.1002/(SICI)1099-0720(199704)11:2<95::AID-ACP429>3.0.CO;2-O

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior*, *12*(1), 41-55. doi: 10.1007/BF01064273

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.

Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra-and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology*, *88*(1), 143-156. doi: 10.1111/j.2044-8295.1997.tb02625.x

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. doi: 10.1080/17470218.2011.570353

Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis. *Psychology, Public Policy, and Law*, *19*(2), 151-164. doi: 10.1037/a0030618

Gabry, J., & Goodrich, B. (2016). *Rstanarm: Bayesian applied regression modeling via stan*. Retrieved from https://CRAN.R-project.org/package=rstanarm

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 & Hall.

Haese, A., & Czernochowski, D. (2015). Sometimes we have to intentionally focus on the details: Incidental encoding and perceptual change decrease recognition memory performance and the ERP correlate of recollection. *Brain and Cognition*, *96*, 1-11. doi: 10.1016/j.bandc.2015.02.003

Hammersley, R., & Read, J. D. (1996). Voice identification by humans and computers. In S. L. Sporer, R. S. Malpass & G. Koehnken (Eds.), *Psychological issues in eyewitness identification* (pp. 117-152). Hillsdale, NJ: Lawrence Erlbaum.

Hanley, J. R., Smith, S. T., & Hadfield, J. (1998). I recognise you but I can't place you: An investigation of familiar-only experiences during tests of voice and face recognition. *The Quarterly Journal of Experimental Psychology: Section A*, *51*(1), 179-195. doi: 10.1080/713755751

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 and Neck Surgery*,*143*(3), 274–285. doi:10.1001/jamaoto.2016.3527

Hollien, H., Bahr, R. H., & Harnsberger, J. D. (2014). Issues in forensic voice. *Journal of Voice*, *28*(2), 170-184. doi: 10.1016/j.jvoice.2013.06.011

Holmberg, E. B., Hillman, R. E., Perkell, J. S., & Gress, C. (1994). Relationships between intra-speaker variation in aerodynamic measures of voice production and variation in SPL across repeated recordings. *Journal of Speech, Language, and Hearing Research*, *37*(3), 484-495. doi: 10.1044/jshr.3703.484

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

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, *50*(2), 120–126.

Kerstholt, J. H., Jansen, N. J., 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. doi: 10.1002/acp.974

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

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). New York, NY: 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. doi: 10.1177/1094428112457829

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

Laub, C. E., Wylie, L. E., & Bornstein, B. H. (2013). Can the courts tell an ear from an eye: Legal approaches to voice identification evidence. *Law & Psychology Rev*iew, *37,* 119-158.

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. doi: 10.1348/135532508X382708

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

Lindsay, R. C., Wells, G. L., & Rumpel, C. M. (1981). Can people detect eyewitness-identification accuracy within and across situations? *Journal of Applied Psychology, 66*(1), 79-89. doi: 10.1037/0021-9010.66.1.79

Liu, Y., Gelman, A., & Zheng, T. (2015). Simulation-efficient shortest probability intervals. *Statistics and Computing*, *25*(4), 809-819. doi: 10.1007/s11222-015-9563-8

Lu, Z. L., Williamson, S. J., & Kaufman, L. (1992). Behavioral lifetime of human auditory sensory memory predicted by physiological measures. *Science*, *258*(5088), 1668-1670. doi: 10.1126/science.1455246

Malpass, R. S., & Lindsay, R. C. L. (1999). Measuring line-up fairness. *Applied Cognitive Psychology, 13,* S1‑S7. doi: 10.1002/(SICI)1099- 0720(199911)13:1+%3CS1::AID-ACP678%3E3.0.CO;2-9

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'Hello'? Personality impressions from brief novel voices. *PloS one*, *9*(3), e90779. doi: 10.1371/journal.pone.0090779

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. doi: 10.1080/00224545.1993.9712155

McDougall, K. (2013). Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades. *International Journal of Speech, Language & the Law*, *20*(2), 163-172.

McDougall, K., Nolan, F., & Hudson, T. (2015). Telephone transmission and earwitnesses: Performance on voice parades controlled for voice similarity. *Phonetica, 72*(4), 257-272. doi: 10.1159/000439385

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: Chapman & Hall.

McGorrery, P. G., & McMahon, M. (2017). A fair 'hearing': Earwitness identifications and voice identification parades. *The International Journal of Evidence & Proof, 21*(3), 262-286. doi: 10.1177/1365712717690753

Memon, A., & Yarmey, A. D. (1999). Earwitness recall and identification: Comparison of the cognitive interview and the structured interview. *Perceptual and Motor Skills, 88*(3), 797-807. doi: 10.2466/PMS.88.3.797-807

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

Mullennix, J. W., Ross, A., Smith, C., Kuykendall, K., Conard, J., & Barb, S. (2011). Typicality effects on memory for voice: Implications for earwitness testimony. *Applied Cognitive Psychology, 25*(1), 29-34. doi: 10.1002/acp.1635

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas – Part II. *arXiv Preprint arXiv:1602.00245*.

Nolan F. (2003). A recent voice parade. *International Journal of Speech, Language and Law, 10*(2), 277-291. doi: 10.1558/sll.2003.10.2.277Nolan, F., McDougall, K., & Hudson, T. (2011, August). Some acoustic correlates of perceived (dis) similarity between same-accent voices. In *International Congress of Phonetic Sciences (ICPhS)* (pp. 1506-1509).

Nolan, F., McDougall, K., de Jong, G., & Hudson, T. (2006, December). A forensic phonetic study of 'dynamic' sources of variability in speech: the DyViS project. In *Proceedings of the 11th Australasian International Conference on Speech science and Technology* (pp. 13-18).

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 and the Law*, *16*(1), 31-57. doi: 10.1558/ijsll.v16i1.31

Ohman, 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. doi: 10.1007/s11896-010-9076-5

Ohman, L., Eriksson, A., & Granhag, P. A. (2013a). 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. doi: 10.1002/jip.1381

Ohman, L., Eriksson, A., & Granhag, P. A. (2013b). Enhancing adults' and children's earwitness memory: Examining three types of interviews. *Psychiatry, Psychology and Law*, *20*(2), 216-229. doi: 10.1007/s11896-010-9076-5

Olsson, N. (2000). A comparison of correlation, calibration, and diagnosticity as measures of the confidence–accuracy relationship in witness identification. *Journal of Applied Psychology*, *85*(4), 504-511. doi: 10.1037/0021-9010.85.4.504

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

Ormerod, D. (2001). Sounds familiar? Voice identification evidence. *Criminal Law Review*, 595-622.

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. doi: 10.1037/a0031602

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2,* 1–8. doi:10.3389/neuro. 11.010.2008

Perfect, T. J., Hunt, L. J., & Harris, C. M. (2002). Verbal overshadowing in voice recognition. *Applied Cognitive Psychology*, *16*(8), 973-980. doi: 10.1002/acp.920

Philippon, A. C., Cherryman, J., Bull, R., & Vrij, A. (2007a). Lay people's and police officers' attitudes towards the usefulness of perpetrator voice identification *Applied Cognitive Psychology*, *21*(4), 539-550. doi: 10.1002/acp.1281

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

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. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; https://www.R-project.org/.

Robson, J. (2017). A fair hearing? The use of voice identification parades in criminal investigations in England and Wales. *Criminal Law Review, 1,* 36-50.

Roebuck, R., & Wilding, J. (1993). Effects of vowel variety and sample length on identification of a speaker in a line-up. *Applied Cognitive Psychology*, *7*(6), 475-481. doi: 10.1002/acp.2350070603

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology, 65*(1), 111-116. doi: 10.1037/0021-9010.65.1.111

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied, 15*(1), 46-62. doi: 10.1037/a0014560

Seale-Carlisle, T. M., & Mickes, L. (2016). US lineups outperform UK lineups. *Royal Society Open Science, 3,* 160300. doi: 10.1098/rsos.160300

Searcy, J. H., Bartlett, J. C., & Memon, A. (1999). Age differences in accuracy and choosing in eyewitness identification and face recognition. *Memory and Cognition, 27*(3), 538-552. doi: 10.3758/BF03211547

Sørensen, M. H. (2012). Voice line-ups: speakers' F0 values influence the reliability of voice recognitions. *International Journal of Speech, Language & the Law, 19*(2), 145-158.

Sorensen, T., Hohenstein, S., & Vasishth, S. (2015). Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists. *arXiv Preprint arXiv:1506.06201*.

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118*(3), 315-327. doi: 10.1037/0033-2909.118.3.315

Stepan, M. E., Dehnke, T. M., & Fenn, K. M. (2017). Sleep and eyewitness memory: Fewer false identifications after sleep when the target is absent from the lineup. *PloS one, 12*(9), e0182907. doi: 10.1371/journal.pone.0182907

Stevenage, S. V., Clarke, G., & McNeill, A. (2012). The “other-accent” effect in voice recognition. *Journal of Cognitive Psychology, 24*(6), 647-653. doi: 10.1080/20445911.2012.675321

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

Tredoux, C. G. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior*, *22*(2), 217-237. doi: 10.1023/A:1025746220886

Urbaniak, G. C., & Plous, S. (2013). Research Randomizer (Version 4.0) [Computer software]. Retrieved on June 22, 2013, from http://www.randomizer.org/

Van Wallendael, L. R., Surace, A., Parsons, D. H., & Brown, M. (1994). 'Earwitness' voice recognition: Factors affecting accuracy and impact on jurors. *Applied Cognitive Psychology*, *8*(7), 661-677. doi: 10.1002/acp.2350080705

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. doi: 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), 333-342. doi: 10.1037/0096-1523.29.2.333

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

Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology*, *14*(2), 89-103. 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

Wilding, J., & Cook, S. (2000). Sex differences and individual consistency in voice identification. *Perceptual and Motor Skills*, *91*(2), 535-538. doi: 10.2466/pms.2000.91.2.535

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. doi: 10.1177/1529100616686966

Yarmey, A. D. (1991). Voice identification over the telephone. *Journal of Applied Social Psychology*, *21*(22), 1868-1876. doi: 10.1111/j.1559-1816.1991.tb00510.x

Yarmey, A. D. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law*, *1*(4), 792-816. doi: 10.1037/1076-8971.1.4.792

Yarmey, A. D., Yarmey, A. L., & Yarmey, M. J. (1994). Face and voice identifications in showups and lineups. *Applied Cognitive Psychology*, *8*(5), 453-464. doi: 10.1002/acp.2350080504

Zetterholm, E., Sarwar, F., & Allwood, C. M. (2009). Earwitnesses: The effect of voice differences in identification accuracy and the realism in confidence judgments. *FONETIK 2009*, 180-185.

Figure 1. Posterior probability distribution of inferred response accuracy. The posterior probability is shown by sample duration to illustrate the performance against chance-level (10%). The dashed line indicates chance-level performance. The horizontal bars indicate 95% HPDIs, the range containing the posterior probability for target absent and target present by sample duration. The accuracy for target absent parades and target present parades are displayed in red and green, respectively.

Figure 2. Posterior probability intervals of accuracy values. The dots indicate $\hat{\mu}$, the most probable parameter value, and error bars show the 95% HPDI for each condition interred from the interaction model. Dashed lines indicate chance-level. Chance-level performance was found for conditions in which the HPDI crosses the dashed line.

Figure 3. Posterior probability distribution of inferred response accuracy. The posterior probability is shown by parade type to illustrate the performance against chance-level (10%). The dashed line indicates chance-level performance. The horizontal bars indicate 95% HPDIs, the range of containing the posterior probability for target absent and target present by parade type. The accuracy for target absent parades and target present parades are displayed in red and green, respectively.

Figure 4. Posterior probability intervals of accuracy values. The dots indicate $\hat{\mu}$, the most probable parameter value, and error bars show the 95% HPDI for each condition interred from the interaction model. Dashed lines indicate chance-level. Chance-level performance was found for conditions in which the HPDI crosses the dashed line.

Table 1

Mean accuracy in % and median confidence ratings with standard deviations (SD) and number of observations (N).

		Accuracy			Confidence		
Target presence	Sample duration	Mean	*SD*	*N*	Median	*SD*	*N*
Absent	15 s	4.35	20.85	23	5.50	1.53	18
Absent	30 s	8.00	27.69	25	7.00	2.81	24
Present	15 s	39.13	49.90	23	6.00	2.00	15
Present	30 s	38.10	49.76	21	6.00	2.11	17

Table 2

Target present responses: frequency of hits, misses and false alarms.

Sample duration	Hit	Miss	False alarm
15 s	9 (.39)	2 (.09)	12 (.52)
30 s	8 (.38)	1 (.05)	12 (.57)
Total (N = 44)	17 (.39)	3 (.07)	24 (.55)

Note. Proportions in parentheses.

Table 3

Model comparisons for accuracy data (Experiment 1).

Model	*Δelpd*	*ΔSE*	*elpd*	*SE*
Main effect: target presence	0	0	-43.81	5.86
Main effect: target presence, sample duration	-1.13	0.45	-44.94	6.09
Main effects and interaction	-2.61	1.5	-46.41	6.72
Intercept-only model	-3.94	2.1	-47.74	5.68
Main effect: sample duration	-4.48	2.21	-48.29	5.89

Note. Models are ordered from the model with the highest predictive performance elpd (with standard error [SE]) in the top row. Δelpd shows the difference in the predictive performance (with standard error [ΔSE]) of the best fitting model with main effect of target presence compared to all remaining models.

Table 4

Model comparisons for confidence ratings (Experiment 1).

Model	*Δelpd*	*ΔSE*	*elpd*	*SE*
Intercept-only model	0	0	-169.74	4.54
Main effect: accuracy	-0.48	1.01	-170.22	4.56
Main effect: target presence	-0.7	0.25	-170.43	4.52
Main effect: sample duration	-0.82	0.44	-170.55	4.65
Main effect: target presence, sample duration, accuracy	-1.94	1.24	-171.68	4.6
All main effects and two-way interactions	-5.61	1.76	-175.35	4.88
All main effects and three-way interactions	-7.27	2.51	-177	5.16

Note. Models are ordered from the model with the highest predictive performance elpd (with standard error [SE]) in the top row. Δelpd shows the difference in the predictive performance (with standard error [ΔSE]) of the best fitting model with main effect of target presence compared to all remaining models.

Table 5

Mean accuracy in % and median confidence ratings with standard deviations (SD) and number of observations (N).

		Accuracy			Confidence		
Target presence	Parade type	Mean	*SD*	*N*	Median	*SD*	*N*
Absent	Sequential	17.39	38.76	23	5.00	2.18	23
Absent	Serial	9.52	30.08	21	6.00	1.91	21
Present	Sequential	39.13	49.90	23	6.00	2.44	23
Present	Serial	16.67	38.07	24	6.00	1.99	24

Table 6

Target present responses: frequency of hits, misses and false alarms.

Parade type	Hit	Miss	False alarm
Sequential	9 (.39)	1 (.04)	13 (.57)
Serial	4 (.17)	3 (.12)	17 (.71)
Total (*N* = 47)	13 (.28)	4 (.09)	30 (.64)

Note. Proportions in parentheses.

Table 7

Model comparisons for response accuracy (Experiment 2).

Model	*Δelpd*	*ΔSE*	*elpd*	*SE*
Main effect: target presence, parade type	0	0	-48.21	6.08
Main effect: parade type	-0.01	1.61	-48.22	5.72
Main effect: target presence	-0.23	1.79	-48.44	5.69
Intercept-only	-0.5	2.43	-48.7	5.5
Main effects and interaction	-1.1	0.56	-49.31	6.29

Note. Models are ordered from the model with the highest predictive performance elpd (with standard error [SE]) in the top row. Δelpd shows the difference in the predictive performance (with standard error [ΔSE]) of the best fitting model with main effect of target presence compared to all remaining models.

Table 8

Model comparisons for confidence ratings (Experiment 2).

Model	*Δelpd*	*ΔSE*	*elpd*	*SE*
Intercept-only model	0	0	-204.86	5.22
Main effect: accuracy	-0.05	1.52	-204.91	5.36
Main effect: target presence	-0.9	0.26	-205.77	5.32
Main effect: parade type	-0.93	0.36	-205.79	5.25
Main effect: target presence, sample duration, accuracy	-2.33	1.05	-207.19	5.23
All main effects and two-way interactions	-4.23	2.37	-209.09	5.54
All main effects and three-way interactions	-6.7	2.83	-211.56	5.86

Note. Models are ordered from the model with the highest predictive performance elpd (with standard error [SE]) in the top row. Δelpd shows the difference in the predictive performance (with standard error [ΔSE]) of the best fitting model with main effect of target presence compared to all remaining models.