Online Experiment

Jens Roeser and Harriet Smith[1]

[1] Nottingham Trent University

Author Note

Correspondence concerning this article should be addressed to Jens Roeser and Harriet Smith, 50 Shakespeare St, Nottingham NG1 4FQ. E-mail: jens.roeser@ntu.ac.uk

Online Experiment

Experiment

Method

Participants. 401 participants took part in the experiment (246 females, 150 males, 5 unknown), with an age range of 16–74 years (median = 33, $SD = 12.21$). This sample contains 211 native and 190 non-native speakers of English. We will focus on data from both native speakers of English and non-native speakers. An analysis focusing on the subset of native speakers can be found in Appendix A. Comparisons between native and non-native speakers of English can be found in Appendix B.

Results

Data were analysed in Bayesian linear mixed effects models (Gelman et al., 2014; McElreath, 2016). The R package brms (Bürkner, 2017, 2018) was used to model the data using the probabilistic programming language Stan (Carpenter et al., 2016; Hoffman & Gelman, 2014).

Posterior probability distributions of the statistically inferred parameter values were determined for all conditions in the Parade (serial, sequential) × FA warning (strong, standard) × Target (present, absent) design. Our statistical inference was based on posterior (i.e. statistically inferred) quantities which allow for direct statistical inference. We summarised these quantities as the most probable posterior parameter value μ and the interval around μ that contains 95% of the posterior probability mass; 95% Highest Posterior Density Interval (henceforth, HPDI) for non-symmetric posteriors (Hyndman, 1996; Liu, Gelman, & Zheng, 2015) and probability intervals (henceforth, PI) for symmetric posteriors (Lambert, 2018; Lee & Wagenmakers, 2014). Throughout we present modeled data that allow for direct statistical inference.[1]

[1] Models were fitted with weakly informative priors (see McElreath, 2016) and run with 20,000 iterations on

All models were fitted with maximal random effects structure (Barr, Levy, Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2015) with random intercepts for different lineups and by-lineup slope adjustments for Target and FA warning. Parade type was not included as random slope adjustments because of substantial presentation differences between serial and sequential parades that we address in the next section.

The advantages of using a Bayesian framework for hypothesis testing (Kruschke, 2014; Kruschke, Aguinis, & Joo, 2012) and parameter inference (Lambert, 2018; Lee & Wagenmakers, 2014) is well documented in the literature.

Response accuracy. The accuracy data were analyzed as binary responses (0 = incorrect, 1 = correct) in Bayesian linear mixed effects models with binomial link function. This analysis focuses on the inference of the conditional parameter values and the comparison of these values against the 10% chance-level threshold where 10% is the probability of marking a decision at random:

$$\frac{1}{\text{presented voices} + \text{target absent option}} \cdot 100 = \frac{1}{9+1} \cdot 100 = 10\%$$

Figure 1 illustrates the modelled probability density distributions for the response accuracy separated by serial and sequential parades and trials with strong and standard FA warning. Target absent and target present lineups indicate the bimodal distributions displayed in grey and yellow, respectively. Chance-level performance is indicated by the vertical dashed line.

The results indicates a high probability of incorrect decisions with a more diffuse probability density distribution in target present lineups indicating a large uncertainty about the true parameter value. The presence of the target voice in the lineups was associated with a higher accuracy of $\hat{\mu}$=22.74% in a 95% HPDI of [-4.55%, 73.55%] for serial parades with

3 chains with a warm-up of 10,000 iterations and no thinning. Model convergence was confirmed by the Rubin-Gelman statistic (Gelman & Rubin, 1992) and inspection of the Markov chain Monte Carlo chains.

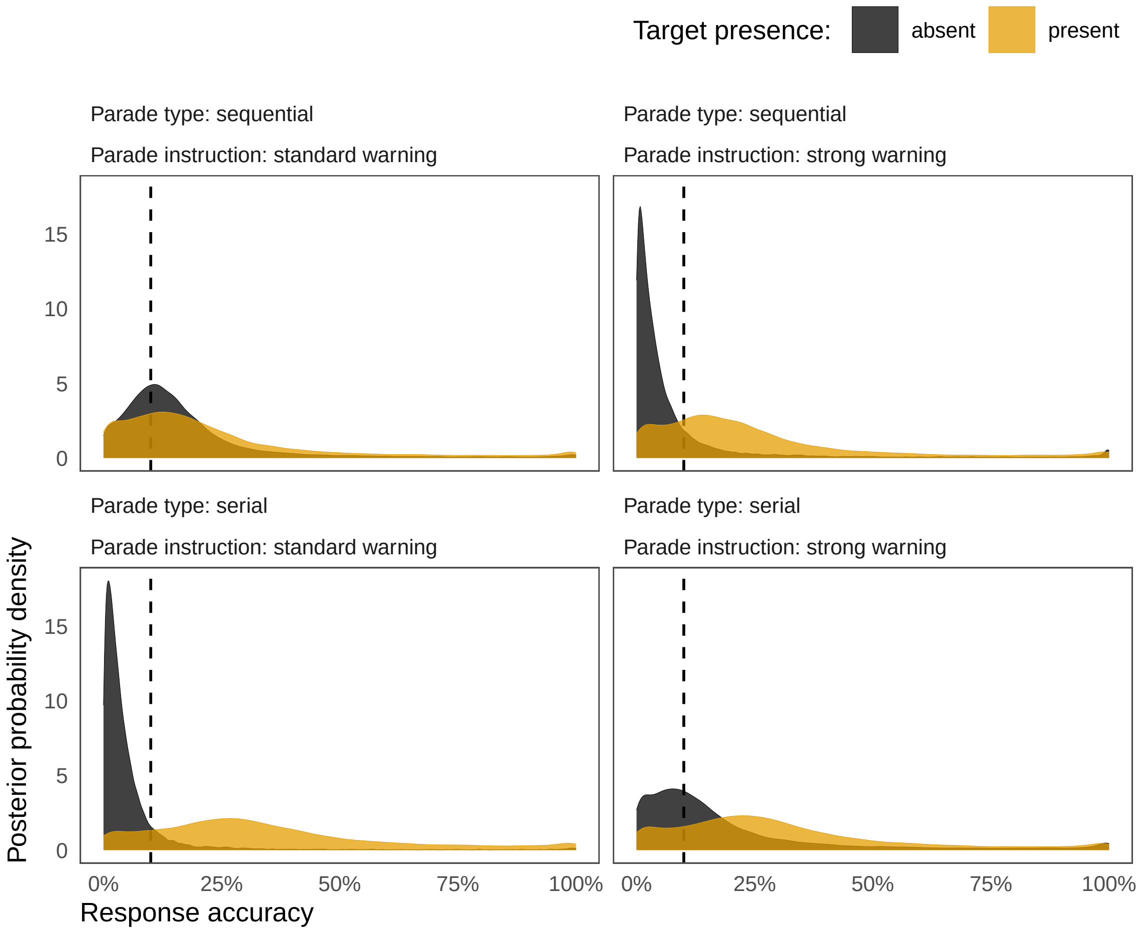

Figure 1. The posterior probability density of the response accuracy for target absent and target present lineups displayed in grey and yellow, respectively. The posterior probability is shown by Parade type × FA warning to illustrate the performance against chance-level (10%). The dashed line indicates chance-level performance.

standard FA warnings. This difference was less substantial when strong FA warnings were provided ($\hat{\mu}$=1.47%, 95% HPDI [-37.98%, 75.84%]). The reverse was found for sequential parades: Target presence increased the response accuracy by $\hat{\mu}$=10.54% (95% HPDI [-17.30%, 82.87%]) when strong FA warnings were provided with negligible evidence for a difference between target present and target absent trials for standard FA warnings ($\hat{\mu}$=0.20%, 95% HPDI [-23.41%, 45.64%]).

Figure 1 shows that the response accuracy for all conditions is non-different from chance. In particular for target absent lineups we observe that the posterior probability mass is either embracing or concentrated below chance level indicating a high probability of incorrect identifications. From these distributions we can infer the posterior probability of below chance-level response accuracies (the posterior density on the left of the vertical dashed line in Figure 1). In other words, we can determine the probability of incorrect accusation (i.e. false-positives for target absent lineups) and missed targets (for target present lineups). The resulting probabilities are illustrated in Figure 2.

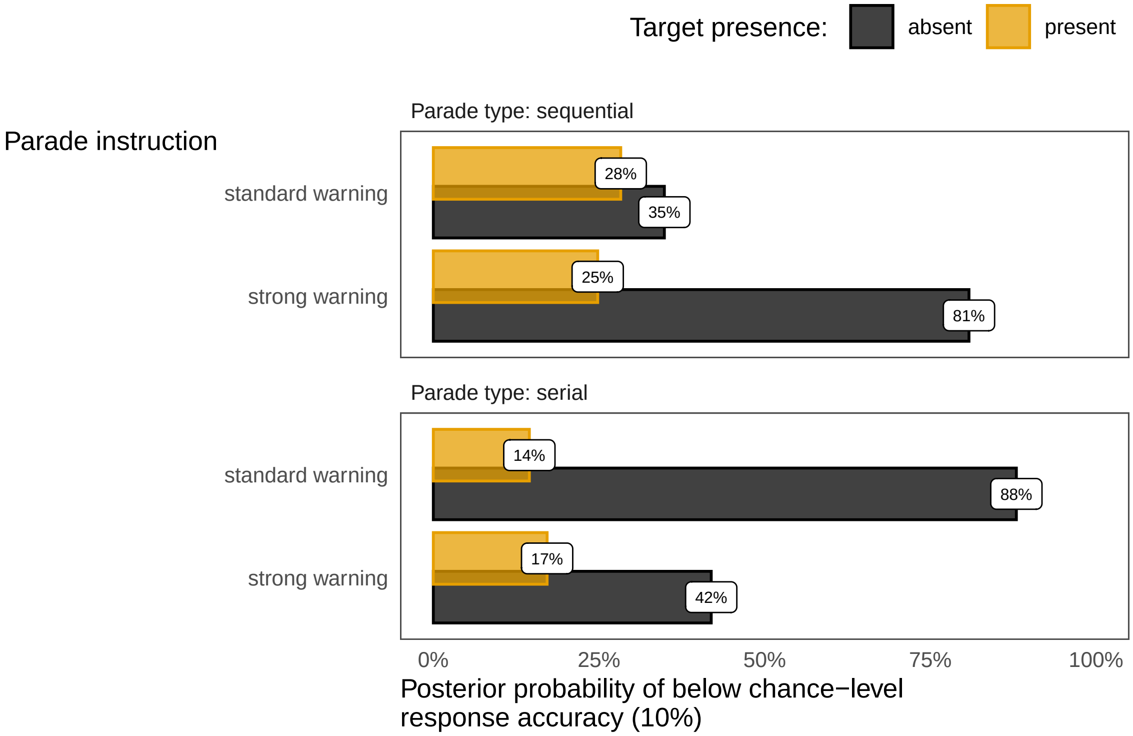

Figure 2. The posterior probability of below chance-level (10%) responses for target absent and target present lineups displayed in grey and yellow, respectively.

Figure 2 shows that the target absent trials are associated with a higher probability of below chance-level responses compared to target present trials. In other words, false alarms are more likely than misses (i.e. incorrectly rejecting a trial that contains the target voice). Striking is the high probability of false-positives when strong FA warnings were presented in sequential parades and for standard FA warnings in serial parades. In particular, in serial parades the probability of below chance-level responses was 46.04% lower but 45.96% higher in sequential parades when strong FA warnings were provided. In other words, FA warnings have opposite effects on the probability of observing false-positives for serial compared to sequential parades. The difference for target present lineups was negligible.

Confidence ratings. Confidence ratings (0 - 10) were analysed in cumulative mixed effects models for ordinal data (Liddell & Kruschke, 2018). Accuracy was included as a nonlinear predictor with adjustments for chance following Bürkner (2019) to prevent biases in the estimator. The posterior ratings show an overall high confidence. There was no evidence for any effects associated with response accuracy ($\hat{\mu}$=2.93, 95% PI [-1.36, 7.34]). The posterior confidence ratings for FA warnings and Target presence can be found in Table 1. There was negligible evidence was differences between confidence ratings.

Table 1

Posterior confidence ratings with most probable parameter value and 95% probability intervals in brackets for serial parades and sequential parades by factors Parade instruction (strong warning, standard warning) and Target presence (present, absent).

Parade instruction	Target presence	Parade type: sequential	Parade type: serial
standard warning	absent	7.64 [7.13, 8.1]	7.62 [7.03, 8.19]
	present	7.81 [7.09, 8.29]	7.6 [7.01, 8.2]
strong warning	absent	7.92 [7.19, 8.33]	7.37 [6.88, 8.06]
	present	7.88 [7.14, 8.29]	7.29 [6.81, 7.98]

Parade comparison

It is non-trivial to compare the results from serial and sequential parades for one fundamental difference: In serial but not in sequential parades, participants listen to all voices before making a decision; in sequential parades, a positive response terminates the trial. This procedural difference suggests higher memory demands for participants in serial parades. While in serial parades, participants have to choose between 9 voices and the target absent option, they do not know the total number of options (voices) to choose from in sequential parades but instead evaluate each voice against the preceding voice(s). In other words, we would expect that the position of a voice in a lineup impacts on participants' responses in sequential but less so in serial parades. This comparison is illustrated in Figure 3. For each lineup position (Lineup ID) participants made a response that is either correct or incorrect. Colour was used to distinguish between serial and sequential parades. Solid lines indicate data with and dotted lines without false alarm warning. As for sequential parades only data until the first positive response were considered, we indicate the number of observations with the size of dots. Data are shown for target absent lineups in the upper panel and target present lineups in the lower panel.

There are two indicators of order-related differences between sequential and serial parades. We observed a lower accuracy for sequential parades, first, for voice 3 (and subsequent voices) in lineups without target voice, and second, for voice 4 and 5 in lineups that contain the target voice. This may be understood as indicator for increasing pressure to make a positive response in sequential parades and / or as resulting from an distribution of responses over all voices in serial parades.

Returning to the comparison of serial and sequential parades: From a statistical point of view, it is difficult to determine an accurate chancel-level threshold for participants that are tested in a sequential parade. The chance-level threshold used before derived from the number of correct options (1) over the sum of the number of possible candidate voices x and the target absent option (1): $\frac{1}{x+1} \cdot 100$.

For serial parades $x = 9$, so the chance-level is $\frac{1}{9+1} \cdot 100 = 10\%$. This is not the case for sequential parades. In sequential parades, participants made a binary decision for each voice but were also aware that only the first positive response is considered. For the first voice in a

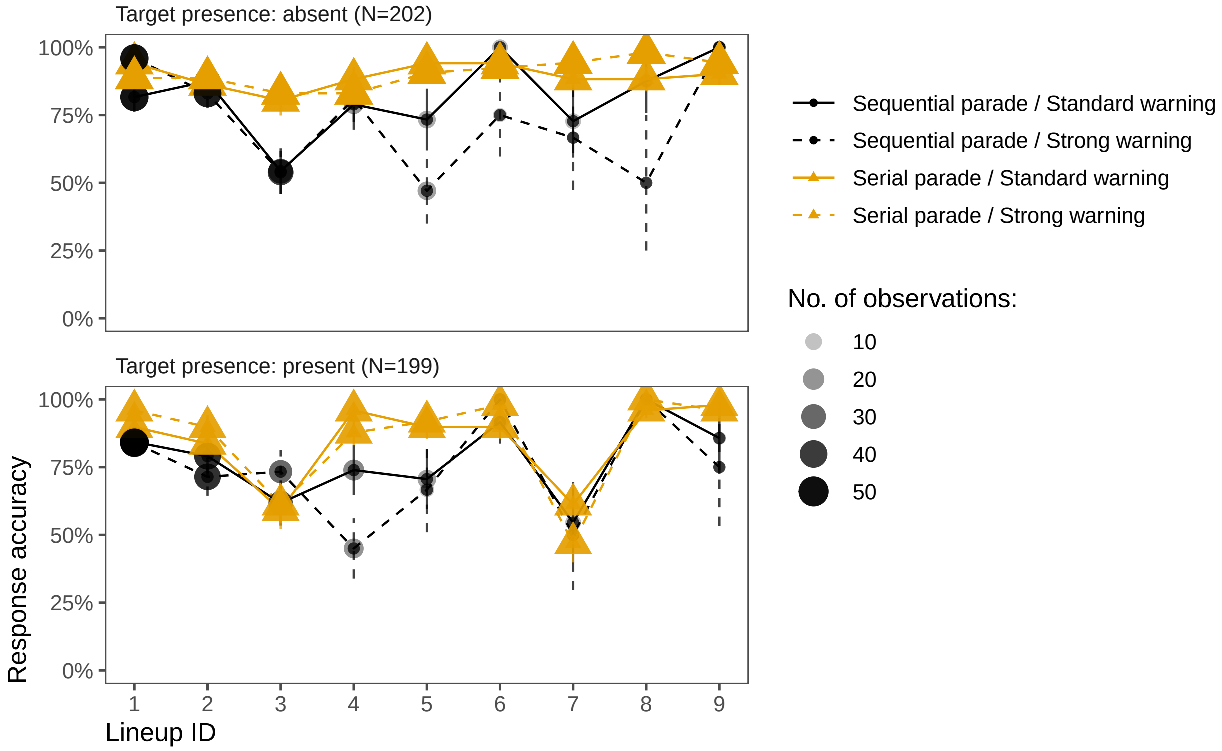

Figure 3. Observed response accuracy. Errorbars show 1×SEs. Response accuracy is shown by Lineup ID, the position of a voice in the lineup. Dot size indicate the number of observations. Response accuracy is shown by Target presence and Parade instructions.

sequential lineup, the probability of guessing correctly is $\frac{1}{2} \cdot 100 = 50\%$ as the number of response option is 2 (whether or not the voice is the target). If this decision for each voice in the lineup was independent of previous decision, this chance-level threshold would be the same for all 9 voices. However, to be able precede to a voice in the lineup, the previous voice(s) required a negative (non-target voice) response. Therefore the decisions are non-independent. The probability of making two non-independent correct decisions by chance is the product of the each decision individually: $\frac{1}{2} \cdot \frac{1}{2} \cdot 100 = 25\%$ (or $(\frac{1}{2})^2$). In other words, the probability of making x binary decision correctly by chance is $(\frac{1}{2})^x \cdot 100$ where x is the position of the voice in a sequential lineup. For example, the probability of guess three voices in a sequence correctly is $(\frac{1}{2})^3 \cdot 100 = 12.5\%$.

Note, that this entails that, for sequential parades, participants would have to correctly reject all 9 voices to correctly indicate the target voice as absent which has a chance-level of $(\frac{1}{2})^9 \cdot 100 \approx 0.195\%$. In the present design this is indistinguishable from strategic negative responses (e.g. pressing absent but not attending to voices at all; not committing to a voice). To avoid this, we can distinguish between whether a voice was identified as target voice (positive response) or not (negative response) rather than response accuracy. This approach is useful as positive responses flag up strategic decisions while response accuracy for negative responses can be the result of response strategies in lineups that do not contain the target voice. Further this consideration highlights the instances in which a voice was identified as target; this is arguably more directly related to the responses made by the participants without classifying responses as either correct or incorrect, similar to the standard treatment in signal detection theory models (DeCarlo, 2010).

Our analysis of response accuracy for voice lineups under different conditions in the previous sections is on a par with previous studies that compared sequential and serial parades by determining participants chance-level response accuracy for the entire lineup. However, the chance-level thresholds must be different for sequential parades, in which a positive respond often proceeds exposure to all voices, compared to serial parades, in which participants are exposed to all voices prior to their response. We suggested that for a comparison between sequential and serial parades we need to take into account the difference in how parades were presented to participants or, in other words, how participants were asked to respond to the lineups. Further, instead of response accuracy, a distinction between positive and negative responses allows us to assess the quality of serial and sequential voice parades. This is because the aim of the voice parade should be to reduce the number of incorrect accusations. In other words, the number of false positives should be ensured to be as low as possible, as each false positive could lead to an unjustified prosecution. A more direct way of looking at this is by determining the probability of identifying a voice as target voice (i.e. positive responses). In the following we present a statistical analysis that takes these features of the voice parade into account.

Positive response probability

As alternative to the analysis of the response accuracy in previous section, ordinal models can be used to infer the probability of selecting on of the different ordinal response categories (i.e. voices in the lineup; 1-9 or target absent). Because of the differences between serial and sequential models with regards to how voices were presented and how participants were allowed to respond, we used two different types of ordinal models.

Models were used to account for two essential differences between parades. First, in serial parades, but not in sequential parades, participants listened to all voices before responding. In other words, participants knew the total number of response options. Cumulative ordinal model were used to model data from serial parades. Second, in sequential parades a participant's response to a particular voice is conditional on negative responses to all preceding voices. For example, responding to the third voice required a negative response to voice one and two; a positive response terminates the decision process. This property is accounted for by sequential process models (Bürkner & Vuorre, 2019; Tutz, 1997).

Sequential process models, in contrast to cumulative ordinal models, do not have an explicit upper bound for the response scale. This is important because any positive response terminates the decision process unless no positive response as made at all; the parade terminates after voice 9. This is important as participants were aware of the total number of options in serial parades but not in sequential parades. Target present responses varied from 1-9 but only in serial parades, participants made explicit target absent responses. We labelled target absent responses as 10 for convenience; in sequential parades, in which no explicit target absent response was given, target absent responses were implied by a negative response to voice 9.

Both types of ordinal models were fitted in generalized Bayesian models (brms

R-package; Bürkner & Vuorre, 2019; Bürkner, 2017). Sequential models were fitted the distribution family sratio with cloglog link-function and ordinal models were fitted with the family cumulative with probit link-function.

Model fit. Models for serial and sequential parades were fitted separately with the selected lineup voice ID (1-9, 10 for target absent responses) as response variable and as predictor variables the expected response (Target present: 3, 7; Target absent: 10) and FA warning (strong warning, standard warning).

As responses should, to some extent, depend on the whether the target voice was voice 3 or 7 (or absent), we fitted additional models with category-specific effects for the predictor expected response. This takes into account that this predictor might have a different effect on the response categories. We tested whether category-specific effects increased the predictive performance of the model. This was assessed in out-of-sample predictions estimated using Pareto smoothed importance-sampling leave-one-out cross-validation (PSIS-LOO) (Vehtari, Gelman, & Gabry, 2015, 2017). Predictive performance was estimated as the sum of the expected log pointwise predictive density ($\widehat{elpd}$). The difference between the predictive quality of the models expressed as $\Delta\widehat{elpd}$. The results of the model comparisons can be found in Table 2. The $\Delta\widehat{elpd}$ values show the difference compared to the best fitting model (top row for each serial and sequential parades). FA warnings showed negligible evidence for increased model performance in serial and sequential parades. Category-specific effects increased the model fit for sequential parades but not for serial parades.

NOTE: I'd omit the model comparisons. This is because it doesn't add much. We do not have enough data to make strong claims based on model comparisons. I think more interesting should be the inferred data in the next subsection.

Model outcome. The results are shown in Figure 4. Shown is the statistically inferred probability to respond with each lineup voice ID. These probabilities are shown separately for parades in which the target was voice 3, 7 or absent. The response probability

Table 2

Model comparisons of ordinal models for response probability. Predictive performance was indicated as expected log pointwise predictive density ($\widehat{elpd}$). The top row of each serial and sequential parades shows the models with the highest predictive performance with differences ($\Delta\widehat{elpd}$) relative to the model with the highest predictive performance.

Formula	$\Delta\widehat{elpd}$	$\widehat{elpd}$
Serial		
Expected response	0 (0)	-455 (6.8)
cs(Expected response)	-6.1 (6.1)	-461 (9.7)
cs(Expected response) + Parade instruction	-6.8 (6.3)	-461.7 (9.9)
Sequential		
cs(Expected response)	0 (0)	-403.5 (11.3)
Expected response	-0.1 (5.4)	-403.7 (10.3)
cs(Expected response) + Parade instruction	-0.2 (1.4)	-403.8 (11.4)

Note. Standard errors are shown in parentheses. Category-specific effects are indicated cs.

for sequential parades is shown in Figure 4a; Figure 4b shows the response probability for serial parades.

These results highlight differences between serial and sequential parades that are associated with the presentation design (i.e. voice order). The response probability for voices early in the lineups is higher for sequential parades compared to serial parades leading to a higher false positives. In particular, voice 3 was frequently chosen in the target absent condition which can only be explained as order effect. Target voice 7 has a low probability to be identified correctly; in fact the probability to select voice 7 is non-different in target

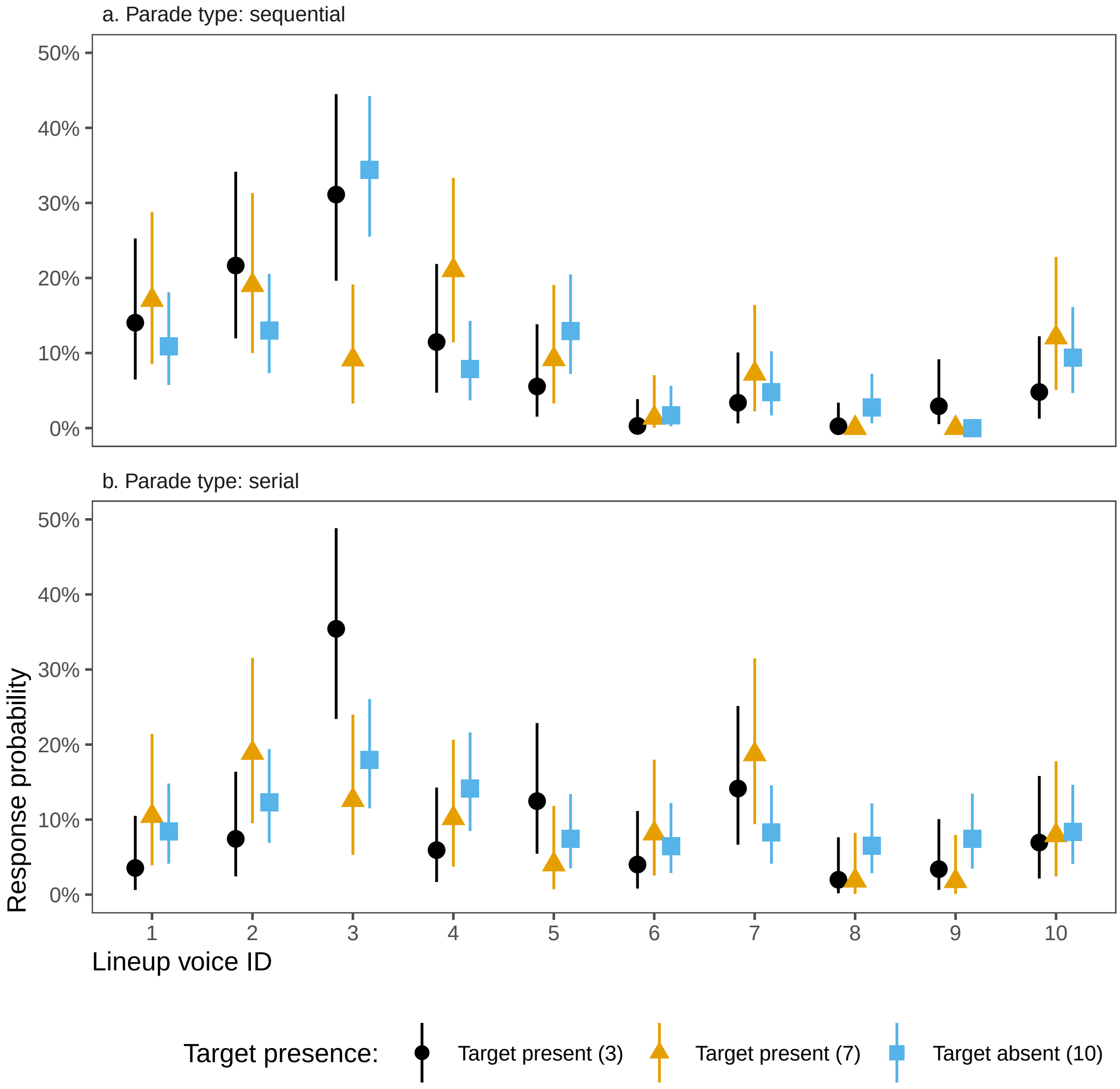

Figure 4. Probability of responses modelled in ordinal models. Errorbars show 95% PIs, the range containing the majority of the probability mass. Target absent responses were labelled 10.

present and target absent lineups. The target voices (3, 7) have a slightly higher probability than others voices to be chosen. Target absent decisions do not differ for target present and target absent trials.

References

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.

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

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. (2019). Bayesian item response modelling in R with brms and Stan. *arXiv Preprint arXiv:1905.09501*.

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.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *20*.

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.

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

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

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

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

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

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

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.

Liu, Y., Gelman, A., & Zheng, T. (2015). Simulation-efficient shortest probability intervals. *Statistics and Computing*, *25*(4), 809–819.

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

Tutz, G. (1997). Sequential models for ordered responses. In *Handbook of modern item* *response theory* (pp. 139–152). Springer.

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.

Appendix A

Results for native English speakers only

In this section we present the results for those participants in our sample who were native speakers of English.

Response accuracy

The accuracy data were analysed as binary responses (0 = incorrect, 1 = correct) in Bayesian linear mixed effects models with binomial link function. Figure A1 illustrates the response-accuracy probability distributions for all conditions. Target absent and target present lineups indicate the bimodal distributions displayed in grey and yellow, respectively. Chance-level is indicated as vertical dashed line.

The results indicates a high probability of incorrect decisions with a more diffuse probability density distribution in target present lineups indicating a large uncertainty about the true parameter value. The presence of the target voice in lineups lead to a higher accuracy of $\hat{\mu}$=29.91% (95% HPDI [-1.94%, 87.84%]) for serial parades when standard FA warnings were provided. This difference was negligible for strong FA warnings ($\hat{\mu}$=1.48%, 95% HPDI [-43.54%, 86.60%]). In sequential parades, target presence did not substantially increase response accuracy neither when strong ($\hat{\mu}$=1.75%, 95% HPDI [0%, 91.52%]) nor when standard FA warnings ($\hat{\mu}$=0.54%, 95% HPDI [-22.47%, 61%]) were provided.

Figure A1 shows that for target absent trials we observe a large amount of posterior probability mass either centered around or concentrated below chance-level indicating a high probability of incorrect identifications. From these distributions we inferred the posterior probability of below chance-level response accuracies. The resulting values are illustrated in Figure A2.

Figure 2 shows that target absent trials are associated with a higher probability of

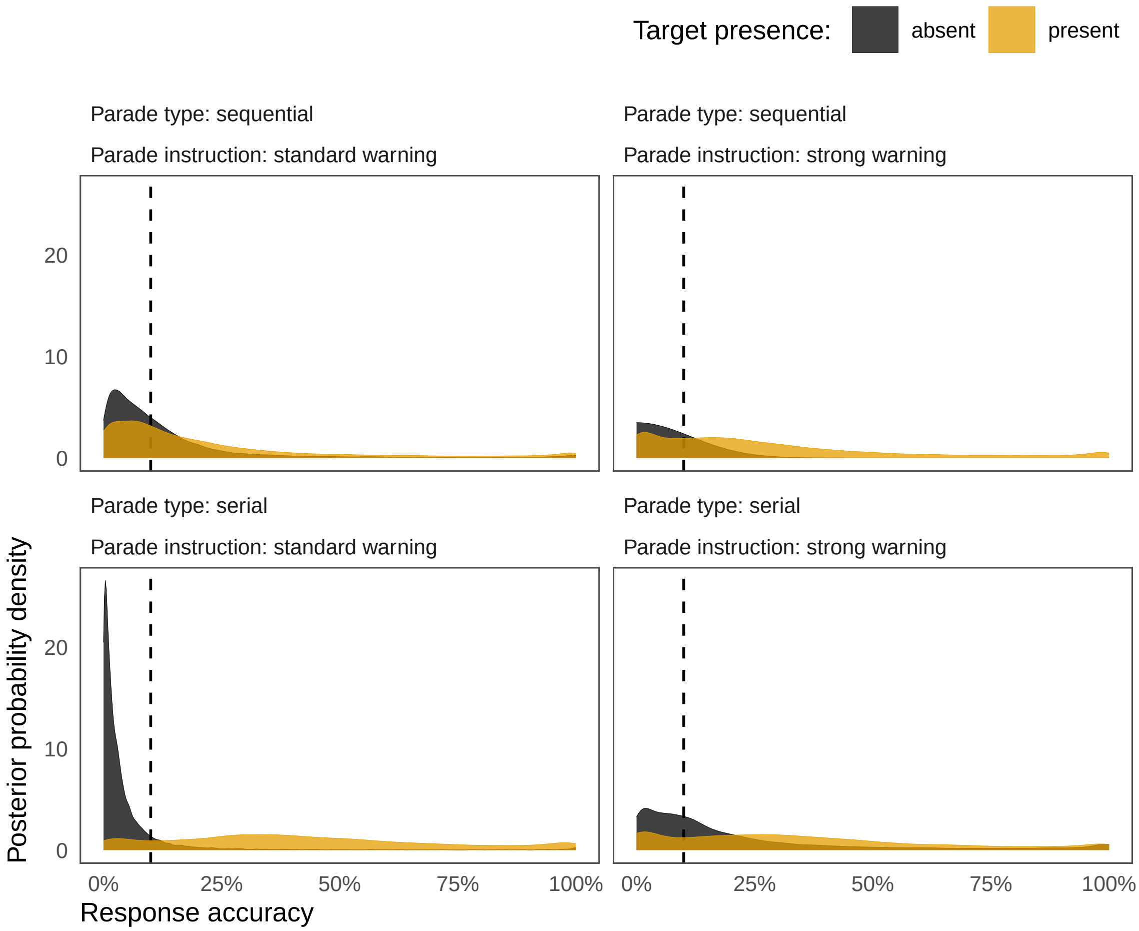

Figure A1. The posterior probability density of the response accuracy for target absent and target present lineups displayed in grey and yellow, respectively for L1 English participants. The posterior probability is shown by Parade type × FA warning to illustrate the performance against chance-level (10%). The dashed line indicates chance-level performance.

below chance-level responses compared to target present trials. The probability of below chance-level responses was highest for target absent trials when strong FA warnings were presented in sequential parades and for standard FA warnings in serial parades. In particular, in serial parades the probability of below chance-level responses was 44.74% lower but 41.73% higher in sequential parades when strong FA warnings were provided. In other words, FA warnings have opposite effects on the probability of observing false-positives for serial compared to sequential parades. This difference for target present lineups was

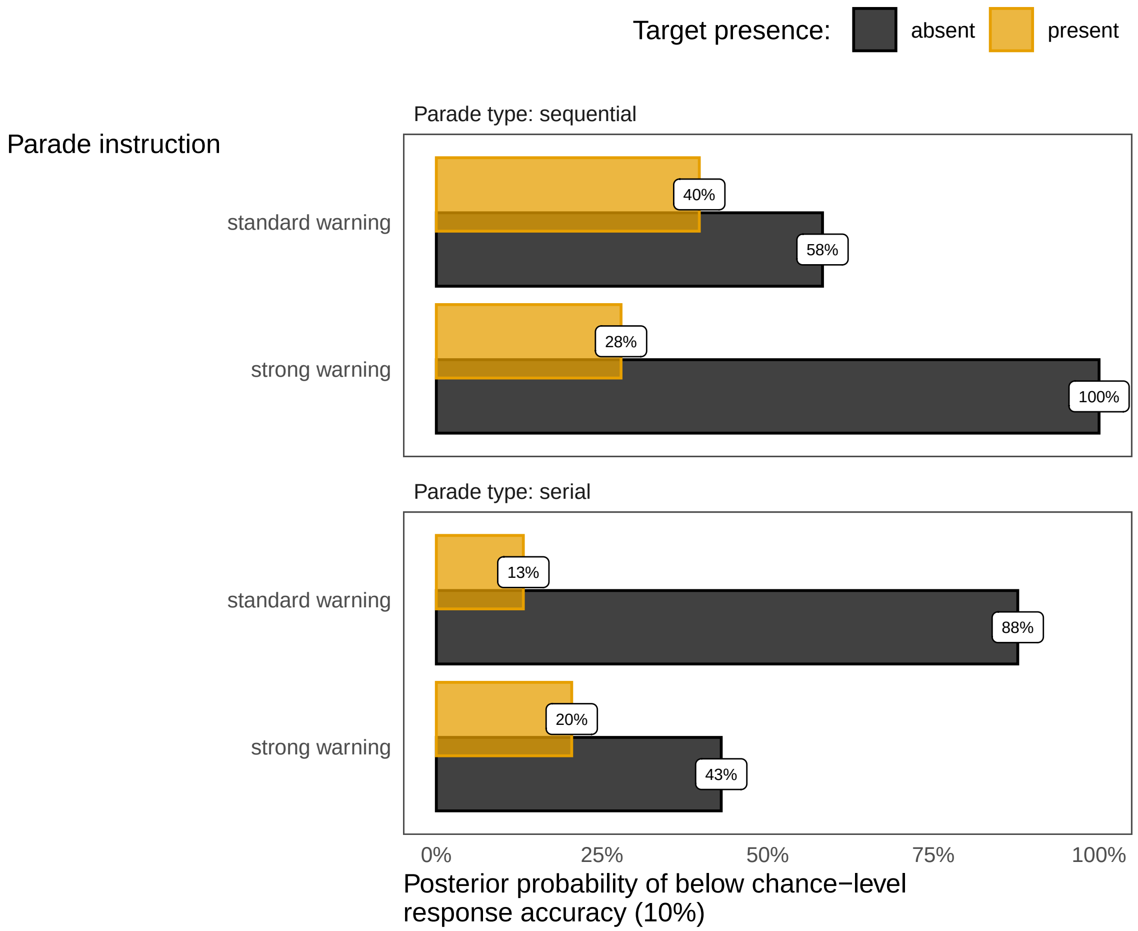

Figure A2. The posterior probability of below chance-level (10%) responses for target absent and target present lineups displayed in grey and yellow, respectively (L1 English participants).

substantially smaller, i.e. 11.81% for sequential parades and 7.30% for serial parades.

Confidence ratings. Confidence ratings (0 - 10) were analysed in cumulative mixed effects models for ordinal data. Accuracy was included as a nonlinear predictor with adjustments for chance to prevent possible biases in the estimator. The posterior ratings show an overall high confidence. There was no evidence for any effects associated with response accuracy ($\hat{\mu}$=1.96, 95% PI [-3.13, 7]). The posterior confidence ratings can be found in Table A1. The evidence for differences between conditions is negligible.

Table A1

L1 English participants. Posterior confidence ratings with most probable parameter value and 95% probability intervals in brackets for serial parades and sequential parades by factors Parade instruction (warning, no warning) and Target presence (present, absent).

Parade instruction	Target presence	Parade type: sequential	Parade type: serial
standard warning	absent	7.97 [7.44, 8.39]	7.93 [7.26, 8.44]
	present	8.03 [7.3, 8.54]	7.9 [7.2, 8.46]
strong warning	absent	8.02 [7.33, 8.51]	7.78 [7.13, 8.37]
	present	8.1 [7.42, 8.57]	7.58 [6.98, 8.27]

Appendix B

First language

The response accuracy and confidence ratings might be affected by whether or not the participant was a native speaker of English. To account for this possibility we modelled the response accuracy and confidence ratings in Bayesian models as described in the results section. The predictor variable was L1 (1 = English, 0 = Others). Figure B1 shows the outcome of both models. Figure B1a shows the posterior distribution of response accuracies, indicating that response accuracy for other first languages than English was distributed tighter around chance level. Figure B1b shows the confidence ratings as a function of response accuracy by first language. The results show that participants with a different first language than English responded with a slightly lower confidence, in particular for incorrectly judged lineups. Overall there is negligible evidence for systematic differences between native and non-native speakers of English.

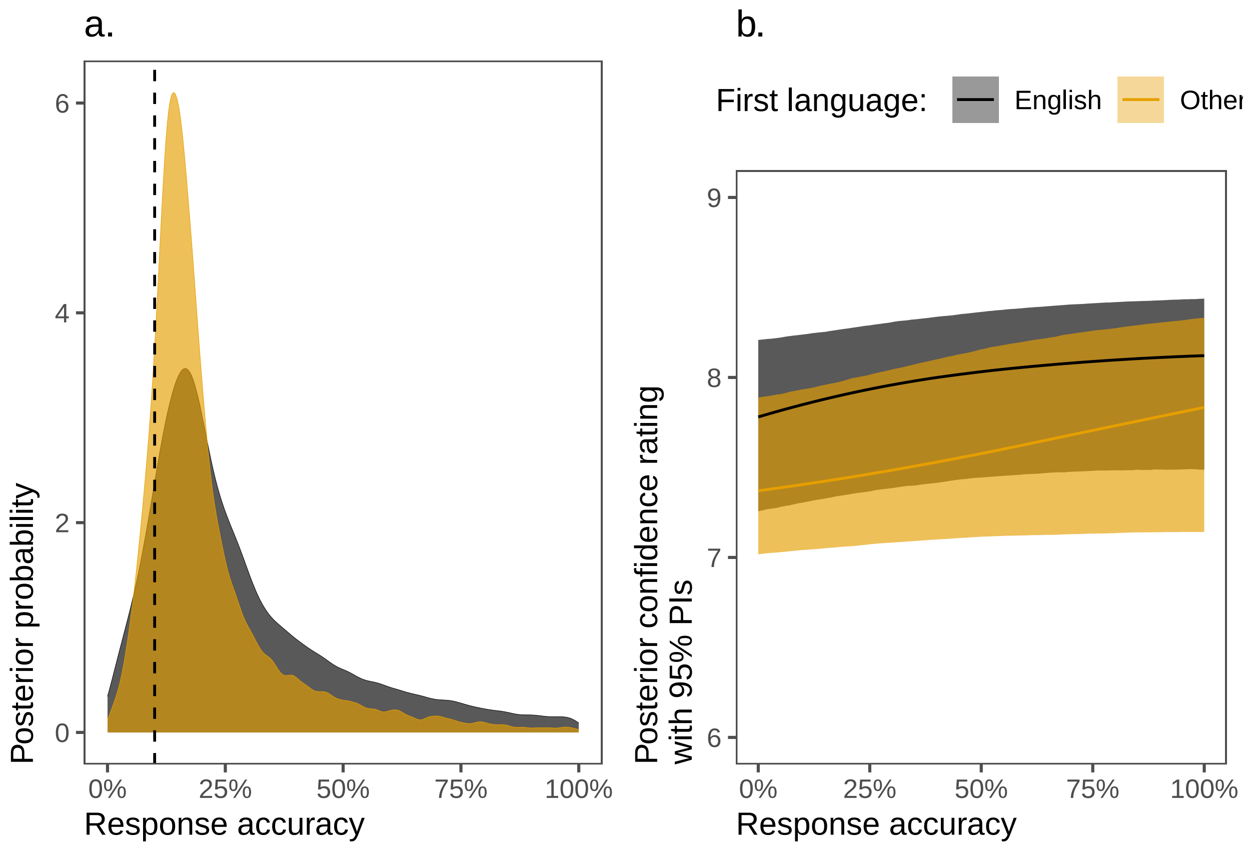

Figure B1. Posterior distribution by first language. Panel (a) shows the distribution of the response accuracy relative to chance-level indicated by the dashed line. Panel (b) shows the confidence ratings as a function of response accuracy. The lines indicate $\hat{\mu}$, the most probable parameter value, and error ribbons show 95% PIs.