Skilled musicians are indeed subject to the McGurk effect

Stephen Politzer-Ahles* & Lei Pan

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong

Address correspondence to:

Stephen Politzer-Ahles

Department of Chinese and Bilingual Studies

The Hong Kong Polytechnic University

Hong Kong

E-mail: stephen.politzerahles@polyu.edu.hk

Abstract

The McGurk effect is a perceptual illusion whereby sounds are often mis-perceived when the auditory cues in the stimulus conflict with the visual cues from the speaker's face. A recent study[1] claims that "skilled musicians are not subject to" this effect. It is not clear, however, if this is intended to mean that skilled musicians do not experience the McGurk effect at all, or if they just experience it to a lesser magnitude than non-musicians. The study also does not statistically demonstrate either of these conclusions, as it does report a numerical (albeit non-significant) McGurk effect for musicians, and does not report a significant difference between musicians' and non-musicians' McGurk effect sizes. The present article reports a pre-registered, higher-power replication of that study (using twice the sample size and changing from a between- to a within-participants manipulation). Contrary to the original study's conclusion, we find that musicians do show a large and statistically significant McGurk effect, and that their effect is no smaller than that of non-musicians.

Introduction

When people comprehend physical stimuli, they integrate information from multiple sensory modalities to generate a psychological percept. For example, the way people perceive a sound can be modulated by visual sensory information accompanying the sound. Perhaps the most famous example of this is the McGurk effect[2], whereby people tend to mis-perceive sounds that are dubbed onto video of people pronouncing a different sound. For example, people might accurately perceive an audio recording of "ba" as being the sound /ba/ (slashes indicate representations in the International Phonetic Alphabet), but when that same sound is dubbed over a video of a person saying "ga" then people often perceive the sound as being something other than "ba". This effect is strong evidence that information from one modality (visual) can influence the perception of information in another modality (auditory).

Proverbio and colleagues[1] argue that the McGurk effect is absent in highly trained musicians. This observation is important because it suggests that their experience changes some mechanisms of basic sound perception, such as, for example, the relative weighting of different types of cues (acoustic, phonetic, visual) when speech is difficult to perceive.

Some aspects of the study, however, limit the conclusiveness of this finding. While the experiment did not reveal a significant difference between musicians' accuracy in audiovisual "McGurk" stimuli and audio-only stimuli, they nevertheless showed a large numerical effect in the direction of a typical McGurk effect (higher accuracy in audio-only than audiovisual stimuli). The experiment likely had low power to detect a significant effect, as this critical comparison was between only 10 participants

who heard audio-only stimuli and 20 who saw audiovisual stimuli. Furthermore, the experiment also tested a control group of participants without musical experience, and there was not a significant interaction between the groups and the type of stimuli perceived. Without a significant interaction, the conclusion that non-musicians had a McGurk effect and musicians did not is not necessarily justified: the fact that one group shows a significant effect and another group does not show a significant effect is not in of itself sufficient evidence that the two groups are significantly different from one another[3]. Another limitation is that the experiment included many stimuli that would not necessarily be expected to elicit McGurk effects even in typical participants: for example, while an audio recording of "ba" dubbed over a video of a person saying "ga" presents the listener with visual cues that conflict with the auditory input, a recording of "ba" dubbed over a video of a person saying "pa" does not, as these two sounds share the same place of articulation and the main difference between them (the duration between the release of the consonant and the onset of vocal fold vibration) is not easily visible. Proverbio and colleagues[1] classified both types of stimuli as incongruent/McGurk audiovisual stimuli; the inclusion of stimuli like "ba" audio dubbed over a "pa" video may have caused McGurk effects to be underestimated. In fact, if their data are re-analyzed with only cases of different place of articulation being included in the McGurk condition, the difference between musicians and non-musicians is reduced (musicians are 3.1% more accurate in the original analysis, but only 1.9% more accurate in this analysis). A final limitation is that the results reported in the paper do not match (i.e., cannot be replicated from) the data that are provided in the paper, as shown in comments on the online version of the paper.

Overall, the findings of Proverbio and colleagues[1] seem indeterminate: they are consistent with the presence of a McGurk effect for musicians and they are also consistent with the absence of such an effect. For these reasons, it is valuable to conduct a close replication of this study to assess whether skilled musicians really are not subject to the McGurk effect. For a stronger test of this question, we make several changes to the study design:

1) We manipulate the crucial comparison (whether participants hear audiovisual stimuli or audio-only stimuli) within subjects, rather than between subjects, to have greater statistical power.
2) We double the sample size.
3) In an additional analysis, we quantify McGurk effects by classifying only stimuli in which the visual and audio place of articulation mismatch (i.e., a recording of "ba" dubbed on a video of "pa") as McGurk stimuli, and classifying others as congruent audiovisual stimuli.
4) We improve the statistical analysis by using generalized linear mixed models[4], rather than analysis of variance, which Proverbio and colleagues[1] used but which are inappropriate for binomial data of this sort[5].
5) We test two hypothesis of relevance. While Proverbio and colleagues[1] focus on whether musicians show a McGurk effect at all, the conceptual conclusions

of their paper would also be supported if musicians just showed a smaller (but still significant) McGurk effect than non-musicians did. We thus report two comparisons: whether musicians' McGurk effect is greater than zero, and whether it is smaller than non-musicians'. Possible outcomes of the experiment are a) musicians' McGurk effect is non-significant, and smaller than that of non-musicians [consistent with the conclusions Proverbio and colleagues[1] made, although not fully consistent with their results]; b) musicians' McGurk effect is significantly greater than zero, but smaller than that of non-musicians [consistent with the conclusions and results of Proverbio and colleagues[1]]; c) musicians' McGurk effect is significantly greater than zero and not significantly smaller than that of non-musicians [inconsistent with the conclusions of Proverbio and colleagues[1], but consistent with their results]; or d) musicians' McGurk effect is not significantly different either from zero or from non-musicians.

Methods

All experimental methods were preregistered at https://osf.io/cuzax/register/565fb3678c5e4a66b5582f67.

Participants

62 skilled musicians and 62 non-musicians were recruited in Hong Kong and Shenzhen. (As Proverbio and colleagues[1] had 30 musicians and 30 non-musicians in the critical conditions, we set a goal of 60 musicians and 60 non-musicians; as we ran the participants in groups, we scheduled slightly more participants than necessary in case of no-shows.) Demographic details for the participants are available at https://osf.io/5ezcp/. Following Proverbio and colleagues[1], we recruited musicians who had at least 13 years' training in musical instruments, and were not singers. The non-musicians were participants who reported that they did not listen to music for more than an hour per day, and either had no music training at all or had not had music training within the past ten years. All participants were native speakers of Mandarin or Cantonese. Experimental procedures were approved by the Human Subjects Ethics Sub-Committee at the Hong Kong Polytechnic University. Participants provided informed consent and were reimbursed with cash for their participation.

Stimuli

All stimuli used in the experiment are available at https://osf.io/5ezcp/. Following Proverbio and colleagues[1], we chose eight consonants to use for the experiment: /b, p, m, f, t, d, l, k/. We limited ourselves to consonants that form existing morphemes in both Mandarin and Cantonese in the frame /_a/. We did not use /n/ because /n/ and /l/ are merged in many southern dialects of Mandarin and

Cantonese. We also did not use /g/ because, while both /ka˥/ and /ga˥/ are not very meaningful in Mandarin (they are mainly used in phonetic borrowings, like 咖啡 /ka˥ fei˥/ "coffee"), /ka˥/ is much more frequent.

We recorded video and audio of each sound being produced by one female native speaker of Mandarin (the second author of this paper, who also served as the experimenter collecting the data) in front of a dark blue background. Each sound was produced three times with an approximately one-second stimulus onset asynchrony (realized by showing the speaker prompts with that timing), and several seconds of silence before and after; four such trios of each sound were recorded, and the best trio (with the clearest productions and least background noise) for each sound was selected. The videos with sound were edited to have one second of silence before and after each trio. Then, incongruent McGurk audiovisual stimuli were created by replacing each video's soundtrack with each other sound's soundtrack (we used Praat[6] to adjust the timing of sounds such that the onset of each sound aligned with the onset of articulation in the video). Microsoft Movie Maker was used to replace soundtracks. This procedure yielded 64 videos (eight sounds times eight videos), eight of which were completely congruent (original, un-edited videos) and 56 of which were edited to replace the original sound with a new sound. Finally, the videos were edited to include a fixation cross located approximately at the tip of the speaker's nose.

The same procedure was used to create practice stimuli, using the consonants /w/ and /j/.

Procedure

Stimulus presentation and response logging were controlled using DMDX[7] (stimuli and scripts available at https://osf.io/5ezcp/). The 64 audiovisual stimuli and 64 audio-only stimuli were arranged in random orders (62 different stimulus lists with their own randomized orders were created, and each list was used for one musician and one non-musician); after each stimulus, the participant was prompted to use the keyboard to enter their transcription of what sound they believe they heard. Participants were instructed to focus on the fixation cross on the screen. Prior to the main experiment, the four audiovisual practice trials and four auditory-only practice trials were presented in a fully random order.

Analysis

Data were automatically coded as correct or incorrect based on whether the first character of the response (forced to uppercase, and with leading whitespaces trimmed) matched the first character of the expected correct response (e.g., for a stimulus whose auditory sound we coded as "KA", responses of "KA", "ka", "KAA", or "KO" would all be marked as correct, but "GA" or "TA" would not). (This deviates

from the pre-registered plan, where we stated that the responses would be manually coded as "correct" or "incorrect" by the authors. We opted for automatic coding instead because it is far more efficient and, based on cursory reviewing of several responses, accurate.)

Statistical analysis was conducted using generalized (logistic) linear mixed-effects models[4] with random effects for participants; maximal random slopes justified by the design were used[8]. The models were fitted using the {lme4} package[9] of the R statistical computing environment[10]; all analysis code is available at available at https://osf.io/5ezcp/. We conduct two analyses: a replication analysis meant to closely replicate the comparisons made by Proverbio and colleagues[1], and a targeted analysis focusing on comparisons where there was a stronger *a priori* expectation of observing McGurk effects (as described in point (3) above). The details of the implementation of each analysis are described in the Results.

Results

All data files are available at https://osf.io/5ezcp/.

For each analysis, we used mixed-effects models to compare the likelihood of correct responses across conditions and groups. The difference between analyses lies in what trials are assigned to which condition: for the replication analysis, trials with the same place of articulation for the audio and visual stimuli but different voicing or manner (e.g., audio "ba" with visual "pa") were treated as incongruent trials, whereas for the targeted analysis these were treated as congruent trials and thus were not analyzed (see below).

Each analysis compared accuracy in the audio-only and audiovisual incongruent conditions, ignoring the audiovisual congruent condition. The McGurk effect was quantified as the difference in accuracy between the audiovisual incongruent condition and the audio-only condition. In each analysis, we regressed accuracy on condition (audio-only vs. audiovisual incongruent), group (musicians vs. non-musicians), and their interaction, as well as random intercepts for participants and random by-participant effects of condition (including correlations between these slopes and intercepts); full analysis code and model specifications are available at https://osf.io/5ezcp/. For each analysis, we focus on two comparisons, corresponding to the two hypotheses listed in point (5) in the introduction: the McGurk effect for musicians (simple effect comparing accuracy for musicians in the audio-only condition to accuracy for musicians in the audiovisual incongruent condition), and the difference between musicians' and non-musicians' McGurk effects (the interaction coefficient). Group was dummy-coded with musicians as the baseline level, and condition dummy-coded with audio-only as the baseline level. Therefore, the coefficient for the condition factor represents the McGurk effect for musicians, and will be negative if audiovisual incongruent trials have lower accuracy

than audio-only trials; and the interaction coefficient will be negative if musicians have a smaller McGurk effect than non-musicians.

Replication analysis

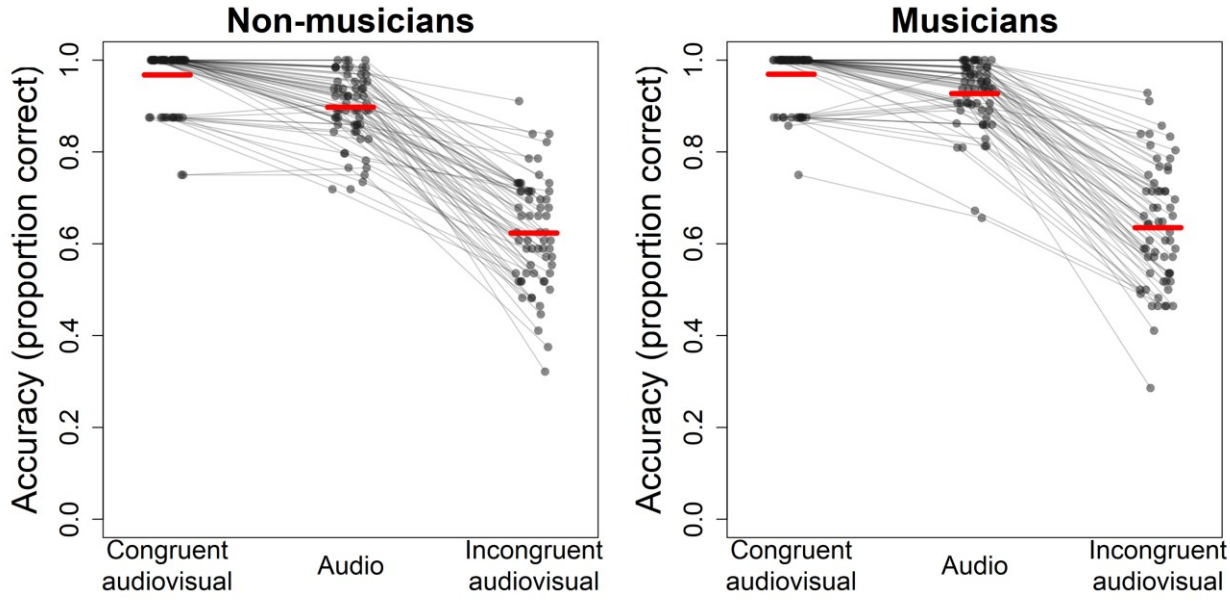

Figure 1. Results for each participant, with conditions coded according to Proverbio and colleagues'[1] scheme. The left panel of the figure shows the results for musicians, and the right panel results for non-musicians. In each panel, accuracy for congruent audiovisual stimuli is shown on the left, followed by accuracy for auditory-only stimuli in the middle, and accuracy for incongruent audiovisual stimuli on the right. The y-axis shows accuracy ranging from 0% correct at the bottom to 100% correct at the top. In each panel, each participant's accuracy on the three conditions is represented by a series of three dots connected by lines. Superimposed over the dots in each condition is a thick red line, one per condition, showing average accuracy across participants. For each group, accuracy in the congruent audiovisual condition is clustered near the maximum, accuracy in the audio-only condition is similar or slightly lower for most participants, and accuracy in the audiovisual incongruent condition is slightly or substantially lower than accuracy in the audio-only condition for every participant.

A summary of the results is shown in Figure 1. It is clear that both musicians and non-musicians had robust McGurk effects: for non-musicians, the audiovisual incongruent condition is 27.4 percentage points less accurate than the audio-only condition, and for musicians it is 29.2 percentage points less accurate (95% two-tailed percentile bootstrap confidence interval: 27.7%...33.4%). The McGurk effect for musicians is numerically *larger* than that for non-musicians (95% CI of difference: -2.5%...5.7%). The statistical analysis confirmed that musicians have a

highly significant McGurk effect (b=-2.34, z=-18.37, p<.001), and that it in fact is significantly larger than the non-musicians' McGurk effect (b=0.46, z=2.72, p=.007).

Targeted analysis

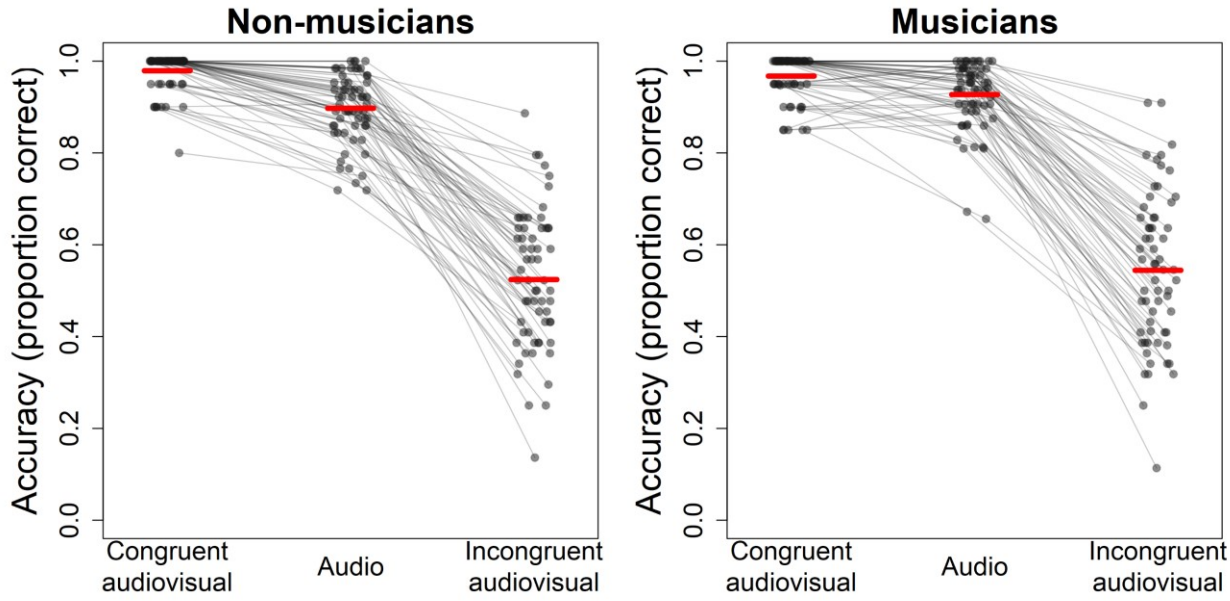

Figure 2. Results for each participant, with conditions coded in our new method (based only on whether place of articulation is incongruent between audio and visual information). The figure is lain out in the same way as Figure 1, and the pattern of results is the same, except that the accuracy for congruent audiovisual trials is generally higher and the accuracy for incongruent audiovisual trials generally lower.

A summary of the results is shown in Figure 2. The overall pattern is similar to that of the replication analysis (Figure 1), except that accuracy on the incongruent audiovisual conditions is visibly lower since this condition no longer includes stimuli with consistent place of articulation between video and audio. Non-musicians have a McGurk effect of 37.3 percentage points, and musicians 38.3 percentage points; again the effect for musicians is numerically larger than that for non-musicians. The statistical analysis confirmed that musicians have a highly significant McGurk effect (b=-2.73, z=-20.89, p<.001) and that their McGurk effect is significantly higher than that for non-musicians (b=0.42, z=2.44, p=.015).

Discussion

In a pre-registered study with more than double the sample size of the original, we found that musicians are subject to the McGurk effect, and that their susceptibility

to the McGurk effect is not significantly less than that of non-musicians—if anything, it is greater. These results are in fact broadly consistent with the results, but not the conclusions, reported by Proverbio and colleagues[1]. They also found that musicians were numerically more accurate on audio-only than incongruent stimuli, and although this effect was not significant, the low sample size and wide error bars in their Figure 1 suggest that this effect likely has a wide confidence interval, making it also not inconsistent with a McGurk effect. Likewise, they did not find a significant interaction between participant group and stimulus condition, but the likely wide confidence interval of this effect in their results makes it not inconsistent with our results, in which musicians' McGurk effects were numerically quite close (albeit statistically significantly larger than) non-musicians'. Thus, both their results and ours are consistent with the conclusion that musicians are subject to the McGurk effect at least as much as non-musicians are; our study provides additional evidence for this conclusion with higher power and more precise statistical estimates of the effect.

For comparing replication studies to original studies, comparing the pattern of significance is not the best approach[11]; it is better to perform some kind of direct statistical comparison between the two studies (ideally, between the effect observed in one study and a hypothetical effect size that the other study would have had sufficient power to detect). Simonsohn[11] outlines several ways this can be done, as well as their limitations. In the present case, how a direct comparison between studies would be conducted is not straightforward. Our statistical results are based on mixed-effect models, and it is not possible to re-analyze Proverbio and colleagues'[1] results with this approach using the data available: while they do provide fairly detailed tables summarizing the data (Proverbio et al.[1], Tables 1-2), these tables do not indicate which responses come from which participants, and without that information it is not possible to model these data with mixed effects. On the other hand, we also cannot analyze our data using the exact same method they used, as their analysis is not replicable (online comments on that article point out discrepancies between the values reported in the article and the values calculated from the data provided according to the methods described in the article). Thus, we are not aware of a way to perform a direct statistical test between the studies. However, we can make rough comparisons between the raw values reported in the different studies. According to the prose results and Tables 1-2 in Proverbio et al.[1], musicians were 90% accurate on audio-only stimuli and 92% accurate on incongruent audiovisual stimuli, yielding a 2% effect in the opposite direction of a typical McGurk effect (whereas their Figure 1, showing arcsin transformed data, shows an effect numerically consistent with a McGurk effect; it is not clear how arcsin transformation would reverse the sign of the effect, and it is possible that one or more of these values may be inaccurate, given other inconsistencies in the data reporting, e.g. the fact that they report 94% accuracy for musicians in congruent audiovisual stimuli whereas the congruent audiovisual cells in Table 1 average out to 97% accuracy). Non-musicians, on the other hand, were 94% accurate on audio-only stimuli and 89% accurate on incongruent

audiovisual stimuli, leading to a McGurk effect of 5% (the same caveat about inconsistencies in the reported values for congruent audiovisual stimuli also applies here). These effects are far smaller than the McGurk effects of 27% and 29% observed in our study (see "Replication analysis" above), which may in part be due to our use of noisier recordings, moving the participants farther away from ceiling. If these results are taken at face value, it seems that our results are inconsistent with theirs: the McGurk effect observed in musicians is far smaller in their study than ours, and the difference between musicians' and non-musicians' McGurk effects in their study is larger than the corresponding difference in ours. Their negative-3-point difference between musicians' and non-musician's McGurk effects (2% minus 5%), however, is only just barely outside the confidence interval of the corresponding difference in our study, and the difference in our study is likely within the confidence interval of theirs (given the smaller sample size and between-participants comparison of their study, their confidence interval is likely wider than ours). Furthermore, given that our study is higher powered (based on the larger sample size and the use of within-subject comparisons) and pre-registered, and given the other improvements made in our study (listed in the Introduction), we believe our result—musicians showing a substantial McGurk effect, which is not significantly smaller than that shown by non-musicians—to be the more robust one.

Proverbio and colleagues[1] make several conclusions about how musicians' non-susceptibility to the McGurk effect suggests that their music experience may have re-organized the functional specification of several brain areas. The present study obviously rules out those conclusions: if musicians *are* indeed subject to the McGurk effect, then McGurk studies do not provide any evidence for brain-level functional reorganization as a result of musical training. The present results do not necessarily, however, challenge the overall notion of brain plasticity as a result of training or experience; as reviewed by Proverbio and colleagues[1] in their discussion, there are several other lines of research providing converging evidence for the notion of plasticity in musicians, and thus this study does not alone reject that entire notion. It does suggest, however, that McGurk effects are not a piece of evidence for that sort of plasticity.

In short, the most parsimonious conclusion of the results from Proverbio et al.[1] and the present study is that musicians *are* subject to the McGurk effect, and to at least as much extent (if not more) as non-musicians are.

Author contributions

SPA and LP conceptualized the experiment; LP prepared the stimuli; SPA and LP prepared the experiment protocol; LP collected the data; SPA analyzed the data; SPA and LP wrote the manuscript.

Competing interests

The authors declare no competing interests.

Data availability

All stimuli, experiment programs, data, and analysis code are available at https://osf.io/5ezcp/.

References

1. Proverbio, A., Massetti, G., Rizzi, E., & Zani, A. Skilled musicians are not subject to the McGurk effect. *Sci. Rep.* **6**, 30423 (2016).
2. McGurk, H., & MacDonald, J. Hearing lips and seeing voices. *Nature* **264**, 746-748 (1976).
3. Gelman, A., & Stern, H. The difference between "significant" and "not significant" is not itself statistically significant. *Am. Stat.* **60**, 328-331 (2006).
4. Baayen, R., Davidson, D., & Bates, D. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* **59**, 390-412 (2008).
5. Jaeger, T. Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* **59**, 434-446 (2008).
6. Boersma, P., & Weenink, D. Praat: doing phonetics by computer [computer program]. Version 6.0.30, retrieved 22 July 2017 from http://www.praat.org/ (2017).
7. Forster, K., & Forster, J. DMDX: a Windows display program with millisecond accuracy. *Behav. Res. Methods Instrum. Comput.* **35**, 116 (2003).
8. Barr, D., Levy, R., Scheepers, C., & Tily, H. Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* **68**, 255-278 (2013).
9. Bates, D., Maechler, M., Bolker, B., & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1-48 (2015).
10. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (2016).
11. Simonsohn, U. Small telescopes: detectability and the evaluation of replication results. *Psych. Sci.* **26**, 559-569 (2015).