Dissociation Between Speech and Emotion Effects in Short-Term Memory: A Data Reanalysis.

Stefan Wiens

Department of Psychology, Stockholm University

Abstract

Performance in visual serial recall tasks is often impaired by irrelevant auditory distracters. The duplex-mechanism account of auditory distraction states that if the distracters provide order cues, these interfere with the processing of the order cues in the serial recall task (interference by process). In contrast, the unitary account states that distracters capture only attention on a general level (attentional distraction) without interfering specifically with order processing. Marsh et al. (2018, *Journal of Experimental Psychology-Learning Memory and Cognition, 44*, 882-897) reported finding a dissociation between the effects of serial recall tasks and those of a missing-item task on the disruptive effects of speech and of emotional words, as predicted by the duplex-mechanism account. Critically, the reported analyses did not test specifically for the claimed dissociation. Therefore, I reanalyzed the Marsh et al. data and conducted the appropriate analyses. I also tested the dissociation more directly and added a Bayesian hypothesis test to measure the strength of the evidence for a dissociation. Results provided strong evidence for a dissociation (i.e., crossover interaction) between effects of speech and of emotion. Because the duplex-mechanism account predicts this dissociation between speech effects (interference by process) and emotion effects (attentional diversion) whereas the unitary account does not, Marsh et al.'s data support the duplex-mechanism account. However, to show that this dissociation is robust, researchers are advised to replicate this dissociation in an adversarial registered report.

Keywords: Short-term memory, Irrelevant speech, Serial recall, Auditory distraction

Marsh et al. (2018) study

The ability to remember the order of events, which is critical for short-term memory, is commonly tested with serial recall tasks. For example, participants are shown a series of digits at a rate of one digit per second, and afterward, they are asked to recall the correct order of the digits. The *irrelevant sound effect* refers to the observation that irrelevant speech or other sounds presented during the task impair recall performance (Beaman and Jones, 1997; Ellermeier and Zimmer, 2014; Jones and Macken, 1993). One prominent explanation

for this effect is the *duplex-mechanism account* (Hughes, 2014; Hughes et al., 2007), which proposes two separate mechanisms: *interference by process* and *attentional diversion*. Interference by process occurs because irrelevant sounds that change over time provide order cues that are processed automatically, and these order cues interfere with the processing of the order cues in the serial recall task. Attentional diversion occurs because irrelevant sounds capture attention, and this attentional diversion away from the serial recall task also impairs performance. According to the duplex-mechanism account, task impairment results from both processes. In

contrast, according to the *unitary account*, task impairment results only from attentional diversion (Bell et al., 2019; Körner et al., 2017; Röer et al., 2015).

In support of the unitary account, recent studies found that the content of speech (i.e., postcategorical properties such as meaning) can disrupt performance in serial recall. For example, recall performance is disrupted more by emotional words than neutral words (Buchner et al., 2006; Buchner et al., 2004), more by taboo words than neutral words (Röer et al., 2017), and more by participants' own names than control names (Röer et al., 2013). However, as argued by Marsh et al. (2018), these findings do not necessarily demonstrate that auditory distraction in serial recall can be caused only by attentional diversion. Both interference by process and attentional diversion may disrupt recall performance. If so, it should be possible to dissociate these disruptive effects.

In a clever study, Marsh et al. (2018) conducted two experiments intended to demonstrate this dissociation. In the first experiment, participants performed a serial recall task with eight digits (1 to 8 in random order, one per 900 ms) that were either easy to read (*low load*) or difficult to read (*high load*). In low load, the digits were clearly visible, whereas in high load, the digits were embedded in random visual noise. After each series of eight digits, participants had to recall the order of the digits. Recall performance was indexed by proportion correct (across serial positions). In the second experiment, participants performed a missing-item task: Participants were presented with series of eight different digits in random order (as in serial recall), and at the end of each trial, the participants had to report which digit from the 1-9 range was missing from the series. Thus, digit order did not have to be processed to perform the missing-item task. Performance was also indexed by proportion correct.

Each task comprised six conditions: quiet and five conditions with auditory distracters (15 trials per condition). The five distracter conditions were neutral words and two content categories (social and physical) each of positive and negative words. For each trial with auditory distracters, all distracters were drawn from the same condition. Also, each digit in a series was accompanied by an auditory distracter, and both had the same onset. Note that the two content categories (social and physical) were merged by Marsh et al. (2018) on the basis of preliminary analyses. Thus, there were four conditions in the final analyses: quiet, neutral, positive, and negative.

Because the neutral words were not emotional, they were used to capture the distracting effect of speech per se. Quiet (i.e., no sound) was the control condition.

Thus, the *speech effect* was the difference of quiet minus neutral words. Neutral, positive, and negative words were used to capture the distracting effect of emotion. Thus, the *emotion effect* was the difference of neutral words minus the mean of positive and negative words. Note that Marsh et al. (2018) referred to a valence effect, but this term implies a specific interest in the difference between positive and negative words. Instead, I prefer to refer to an emotion effect because the main interest was the difference between neutral words and emotional words (positive and negative combined).

Marsh et al. (2018) argued that according to the duplex-mechanism account, the effects of speech (i.e., interference by process) and of emotion (i.e., attentional diversion) should differ (i.e., be dissociated) between the serial recall task and the missing-item task. The authors argued that the following results would support a dissociation: First, in the serial recall task, attending to digits that are hard to read should decrease attentional diversion but not interference by process. Accordingly, high load (vs. low load) should decrease the emotion effect but not the speech effect. Second, performing the missing-item task (in which order cues are irrelevant) should decrease interference by process but not attentional diversion. Accordingly, the missing-item task (vs. low load in the serial recall task) should decrease the speech effect but not the emotion effect.

Critically, Marsh et al. (2018) analyzed performance data (proportion correct) with ANOVAs and *t* tests, but these analyses did not test specifically for a dissociation between task effects on speech and on emotion. To address this issue, I reanalyzed the Marsh et al. data, which the authors kindly shared with me. Below, I review the original analyses and discuss their problems, simulate hypothetical data to illustrate these problems, conduct the appropriate analyses, test the dissociation more directly, add a Bayesian hypothesis test to measure the strength of evidence for a dissociation, and discuss theoretical implications. In closing, I discuss some meta-scientific concerns.

All scripts, analyses, figures, and additional material are available at Open Science Framework. To facilitate open and reproducible science (Munafò et al., 2017, this material includes a complete *R-markdown* script (Baptiste, 2017; Bates et al., 2015; Lawrence, 2016; Lüdecke, 2021; Müller, 2020; R Core Team, 2016; Singmann et al., 2020; Team, 2020; Wickham et al., 2019; Wiens, 2017; Zhu, 2020). For example, recall performance was measured as proportion correct; thus, it may violate assumptions for ANOVAs (e.g., normality). However, because additional analyses suggested that results were unaffected (see R-markdown script), the simpler analyses with proportion correct are

reported below.

Original analyses

Figure 1a shows the mean proportion correct for the four sound categories for both low and high load in the serial recall task of the first experiment and for the missing-item task of the second experiment.

Marsh et al. (2018) conducted two main analyses to test for the expected pattern of results. Within the framework of null hypothesis significance testing (Perezgonzalez, 2015; Szucs and Ioannidis, 2017; Wiens and Nilsson, 2017), the authors interpreted a significant finding ($p < .05$) as evidence for an effect (thus, significant indicates statistically significant).

The first main analysis compared low and high load in the serial recall task. Proportion-correct data were analyzed in a 2×4 mixed ANOVA with load (low and high load) as a between-subjects variable and sound (quiet, neutral, positive, and negative) as a within-subjects variable. Because the overall interaction was significant ($p = 0.031$ after Greenhouse-Geisser correction), Marsh et al. (2018) interpreted this as evidence that the effect of load (low vs. high load) was different on the speech effect than on the emotion effect. In follow-up analyses, Marsh et al. separately assessed the effect of load (low vs. high) on the speech effect and on the emotion effect. For speech, results showed no significant interaction between load (low and high) and sound (quiet and neutral), $p = .90$. For emotion, results showed that for low load, proportion correct decreased significantly ($p < .05$) from neutral to positive and from positive to negative, whereas for high load, proportion correct did not differ significantly among the sound conditions. According to the authors, these results support a dissociation: Whereas high load had no effect on the speech effect (quiet vs. neutral), it reduced the emotion effect (performance difference among neutral, positive, and negative).

The second main analysis compared the missing-item task with low load in the serial recall task. Notably, Marsh et al. (2018) did not conduct a 2×4 ANOVA (as in the first analysis) but analyzed the effect of task (missing item vs. low load) on the speech effect and emotion effect separately. With regard to the speech effect, an ANOVA of proportion correct with task (missing-item task and serial recall task with low load) as a between-subjects variable and sound (quiet and neutral) as a within-subjects variable showed that the interaction in this 2×2 ANOVA was significant, $p = 0.008$. With regard to the emotion effect, a similar ANOVA with task and sound (neutral, positive, and negative) showed that the interaction in this 2×3 ANOVA was not significant, $p = 0.076$.

Problems with original analyses

Although ANOVAs are commonly used and have intuitive appeal, the analyses reported by Marsh et al. (2018) were not specific enough to support a claim for a dissociation. The critical question is whether the speech and emotion effects differ between the tasks, that is, high versus low load in the first analysis and missing-item versus low load in the second analysis. With regard to the first analysis, high load (vs. low load) should reduce the emotion effect more strongly than the speech effect. The overall interaction in the 2×4 ANOVA should be sensitive to this difference. However, this interaction is unspecific: With its 3 *df*s, the interaction can be conceptualized as representing a combination of three orthogonal (i.e., statistically independent) contrasts at the same time (Wiens and Nilsson, 2017). Therefore, the interaction may be significant because of effects that are irrelevant to the critical question.

Further, in both analyses, Marsh et al. (2018) conducted separate tests of the speech and emotion effects, but the results do not resolve whether the two effects differed from each other. In fact, a nonsignificant effect in one condition and a significant effect in another condition does not imply that the two conditions differ significantly from each other (Gelman and Stern, 2006; Makin and Orban de Xivry, 2019; Nieuwenhuis et al., 2011). Accordingly, the difference between a nonsignificant effect and another, significant effect needs to be tested explicitly.

Simulation

To illustrate that an omnibus interaction in an ANOVA may be significant even though specific effects are not, I simulated data for a 2×4 factorial design (modelled after the first analysis in Marsh et al., 2018). Figure 2a shows the means for the simulated data, and Figure 2b shows difference scores that isolate the speech and emotion effects (i.e., 2×2 design). The speech effect is the difference of quiet minus neutral, and the emotion effect is the difference of neutral minus combined positive and negative. As shown in Figure 2b, low load had the same (large) effect on speech and emotion, whereas high load had the same (small) effect on speech and emotion. Thus, in this 2×2 design, there is a main effect of load. Critically, because the difference between speech and emotion was identical for both loads, there is absolutely no evidence for a dissociation of the effects of load on speech and emotion. That is, there is no interaction between load and effects on speech and emotion. In support, the overall (3-*df*) interaction in the 2×4 ANOVA of the means was significant ($p < .001$), but the specific (1-*df*) interaction in the 2×2 ANOVA
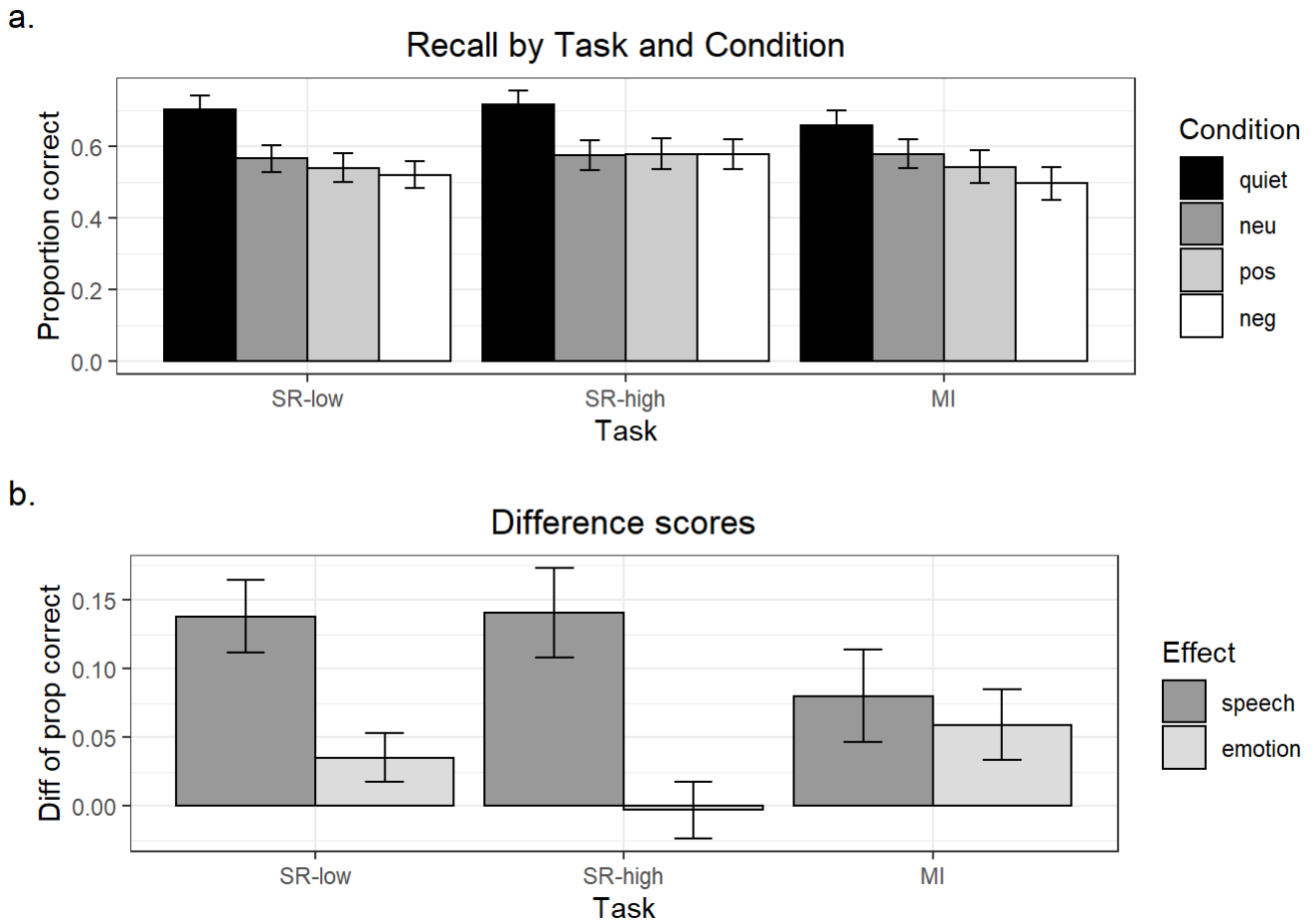
a.

b.

Figure 1. Mean proportion correct of the Marsh et al. (2018) data for the serial recall task with low load (SR-low), the serial recall task with high load (SR-high), and the missing-item task (MI) for the four sound conditions (a) and for effects of speech and of emotion (b). In (b), the speech effect was the difference of quiet minus neutral (neu), and the emotion effect was the difference of neutral minus combined positive (pos) and negative (neg). The error bars denote the 95% CI for each individual mean.

of the difference scores was not ($p = 1$). Note that exact p values (and 95% CI) are not informative because they depend on the noise in the simulated data. Nonetheless, the p values illustrate that an overall interaction in the 2×4 ANOVA does not necessarily support the claim that the load effect differs between speech and emotion. Therefore, it is unclear whether the significant interaction in the first analysis by Marsh et al. shows that the load effect differed between speech and emotion.

Further, because in both analyses by Marsh et al. (2018), a significant effect in one condition and a non-significant effect in another condition does not necessarily imply a significant difference between the two conditions (Gelman and Stern, 2006; Makin and Orban de Xivry, 2019; Nieuwenhuis et al., 2011), the Marsh et

al. analyses did not reveal whether there was a dissociation.

Reanalyses

Because the analyses by Marsh et al. (2018) did not resolve whether there was a dissociation, I reanalyzed the Marsh et al. data to conduct the critical analyses. To simplify them, I computed difference scores to capture the speech effect and the emotion effect, as described above. Figure 1b shows the difference scores for low and high load of the serial recall task and for the missing-item task.

With regard to the first main analysis in Marsh et al. (2018), I conducted a 2×2 ANOVA of proportion correct with load (low and high) as a between-subjects
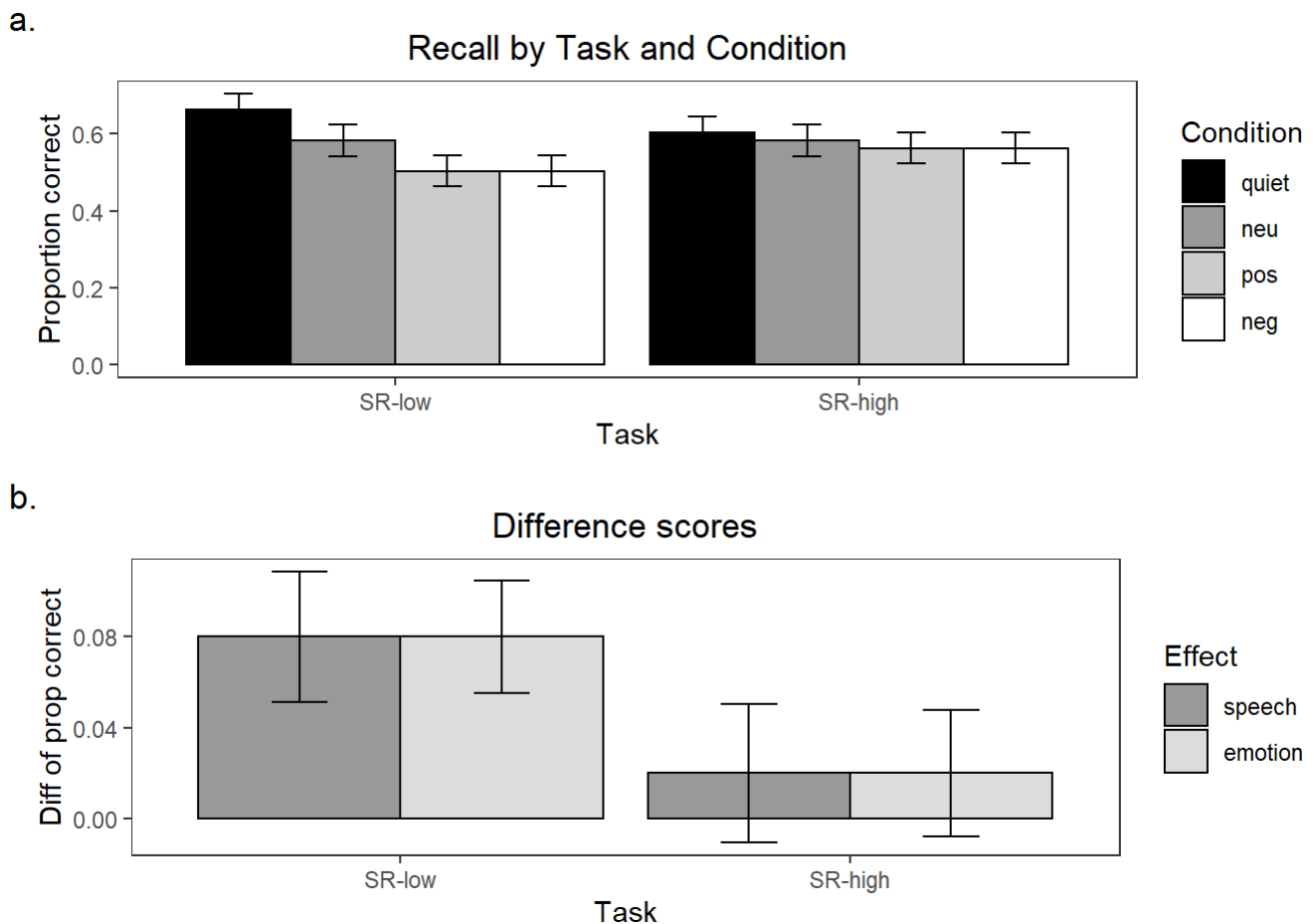
a.

b.

Figure 2. Mean proportion correct of simulated data for the serial recall task with low load (SR-low) and the serial recall task with high load (SR-high) for the four sound conditions (a) and for speech and emotion (b). In (b), the speech effect was the difference of quiet minus neutral (neu), and the emotion effect was the difference of neutral minus combined positive (pos) and negative (neg). The error bars denote the 95% CI for each individual mean.

variable and effect (speech and emotion) as a within-subjects variable. According to Marsh et al., the difference of high minus low load for speech minus emotion should be positive; however, this interaction was not significant, $M = 0.04$, 95% CI [–0.02, 0.10], $p = 0.160$. This result does not support the claim of a dissociation between the effects of low and high load on the disruptive effects of speech and of emotional words. Specifically, it does not provide evidence for a larger effect of load on emotion than on speech.

With regard to the second main analysis in Marsh et al. (2018), I conducted a 2 × 2 ANOVA of proportion correct with load (low load in the serial recall task and missing-item task) as a between-subjects variable and effect (speech and emotion) as a within-subjects variable. According to Marsh et al., the difference of

missing-item task minus low load for the difference of speech minus emotion should be negative; indeed, this interaction was significant, $M = -0.08$, 95% CI [–0.15, –0.02], $p = 0.011$. This result is consistent with the claim for a dissociation between task effects on speech and on emotion. Specifically, task effects (missing-item vs. low load) were larger on speech than on emotion.

Taken together, however, it is unclear whether these results support the claim of a dissociation by Marsh et al. (2018). On the one hand, there was no significant difference in the load effect (low vs. high load) on speech and emotion. On the other hand, there was a significant difference in the task effect (missing-item task vs. low load) on speech and emotion. To resolve this issue, I propose a direct test of the dissociation.

Direct analysis

According to the duplex-mechanism account, there should be a strong dissociation between high load in the serial recall task and the missing-item task in their effects on speech and emotion. Because interference by process should occur mainly in the serial recall task, the speech effect should decrease from the serial recall task with high load to the missing-item task. Conversely, because attentional diversion should be reduced during high load, the emotion effect should increase from the serial recall task with high load to the missing-item task. Thus, there should be a qualitative (crossover) interaction in that task effects should differ in their direction (Berrington de González and Cox, 2007; VanderWeele, 2015). Indeed, in a 2 × 2 ANOVA of proportion correct with load (high load and missing-item task) as a between-subjects variable and effect (speech and emotion) as a within-subjects variable, the interaction was significant, $p < 0.001$. That is, with regard to the difference of high load minus missing-item task for the difference of speech minus emotion, $M = 0.12$, 95% CI [0.05, 0.19], $p = .001$. Follow-up t tests confirmed that the speech effect decreased from the serial recall task with high load to the missing-item task (mean difference of high load minus missing-item task = 0.06, 95% CI [0.01, 0.11], $p = .011$). Conversely, the emotion effect increased from the serial recall task with high load to the missing-item task (mean difference of high load minus missing-item task = –0.06, 95% CI [–0.09, –0.03], $p < .001$). The results of this additional analysis clearly support the claim of a dissociation between the effects of the serial recall task with high load and the missing-item task on speech and on emotion. Specifically, the two tasks' effects were in opposite directions for speech and emotion.

Although these results from null hypothesis significance testing provide evidence against the null hypothesis, they are limited because they do not measure the strength of the evidence for the alternative hypothesis (Dienes and McLatchie, 2018; Szucs and Ioannidis, 2017; Wagenmakers, 2007). That is, because the alternative hypothesis is not made explicit, a statistically significant finding does not necessarily imply that the data support the alternative hypothesis (Dienes, 2008, 2016; Wagenmakers, Love, et al., 2018; Wagenmakers, Marsman, et al., 2018; Wagenmakers et al., 2016; Wiens and Nilsson, 2017). For example, a hypothetical study may find that the task effect is significantly smaller on emotion (0.15) than on speech (0.18). Although this implies that the task effects differ (i.e., the difference is not nil), the difference may seem too small to be theoretically important. In contrast, Bayesian hypothesis testing requires an explicit alternative hypothesis

and allows one to distinguish among data that support the alternative hypothesis, support the null hypothesis, or are inconclusive (Dienes, 2008, 2016; Wagenmakers, Love, et al., 2018; Wagenmakers, Marsman, et al., 2018; Wagenmakers et al., 2016; Wiens and Nilsson, 2017).

In Bayesian hypothesis tests, the *Bayes factor* (BF) compares the likelihood of the data given the null hypothesis with the likelihood of the data given an alternative hypothesis. Because the BF provides a continuous measure of the strength of evidence, I computed the BF to measure the evidence for or against a dissociation between task effects on speech and on emotion. I also used a suggested interpretation scheme to represent the values with a verbal label: $3 > BF > 1$ is considered anecdotal (or inconclusive) evidence, $10 > BF > 3$ is considered moderate evidence, and $30 > BF > 10$ is considered strong evidence (Wagenmakers, Love, et al., 2018).

In the direct analysis, the mean difference of high load in the serial recall task minus the missing-item task was 0.061 for speech (i.e., the distracting effect of speech was larger during high load than during the missing-item task) and was in the opposite direction for emotion (–0.062). To compute the BF, the observed speech effect (i.e., 0.061) was used to define the alternative hypothesis, that is, the task effect on speech was used as a reasonable estimate of the task effect on emotion.

I used three different alternative hypotheses to assess the robustness of the results, as recommended (Dienes, 2014; Dienes and McLatchie, 2018) and as used in previous research in my lab (Ströberg et al., 2017; Wiens et al., 2019). For the *uniform* distribution, the true effect was supposed to fall between 0 and 0.061, and all values were equally likely. For the *half-normal* distribution, the true effect was modeled as a half-normal with a mean of zero and a standard deviation of 0.061. Accordingly, the true effect was supposed to be greater than zero and more likely to be less than 0.061 than to be greater. For the *data-driven t* distribution, the true effect was modeled as a t distribution as defined by the observed speech effect (Dienes and McLatchie, 2018). For the three alternative hypotheses, the BF01 ranged between 12.5 (for uniform) to 16.7 (for half-normal). Because these results provide strong support for the null hypothesis (Wagenmakers, Love, et al., 2018), they support the claim for a dissociation: the effect of the serial recall task with high load versus the missing-item task differed between emotion and speech.

Implications

These reanalyses of the Marsh et al. (2018) data constitute critical tests of the claim that task effects on speech and on emotion are dissociated. The most important analysis is that of the differences between the serial recall task with high load and the missing-item task. The duplex-mechanism account predicts a clear dissociation (i.e., crossover interaction) between the task's effects on speech and on emotion. Because interference by process should be greater during the serial recall task with high load than during the missing-item task, the speech effect should be larger during high load than during the missing-item task. Conversely, because attentional diversion should be less during the serial recall task with high load than during the missing-item task, the emotion effect should be smaller during high load than during the missing-item task. In support, null hypothesis significance tests suggested that the task effect differed between speech and emotion (as $p < .001$), and that the task effect on speech ($p = .011$) was opposite to the task effect on emotion ($p < .001$). Further, Bayesian hypothesis tests provided strong evidence ($16.7 > BF > 12.5$) that the task effect differed between emotion and speech. Taken together, the present reanalyses confirmed that Marsh et al. were correct in their initial claim: Their data provide strong evidence for a dissociation between the task effect on speech and the task effect on emotion.

The reanalyses of the Marsh et al. (2018) data confirm and extend previous reports of an apparent dissociation between interference by process and attentional diversion (Elliott et al., 2016; Hughes et al., 2013; Kattner and Ellermeier, 2018). However, these studies relied on the framework of null hypothesis significance testing, and a statistically significant finding does not necessarily imply that the data support the alternative hypothesis, because the alternative hypothesis is not made explicit (Dienes, 2016; Wagenmakers, Marsman, et al., 2018). The present findings extend previous reports because they provide a direct measure of the strength of the evidence for the idea that interference by process can be dissociated from attentional diversion.

Although results from hypothesis testing are useful in determining whether there is an effect per se (Haaf et al., 2019; Wagenmakers, Marsman, et al., 2018), a complementary approach is to estimate the effect size (Calin-Jageman and Cumming, 2019; Cumming, 2014; Wasserstein et al., 2019; Wiens and Nilsson, 2017). At face value, the 95% confidence intervals (if viewed as likelihood intervals) suggest that the true effect sizes may be rather small (e.g., between 0.01 and 0.11 for the speech effect), but because current theories do not make quantitative predictions, it cannot be resolved whether these effect sizes are theoretically important.

Meta-thoughts

In response to some concerns raised by the reviewers, I would like to discuss a few meta-psychological issues, which seem fitting for the present journal. First, Dr. Marsh introduced me to his article when he visited the department. Although his research is outside of my area, Dr. Marsh encouraged me to submit the reanalysis to the original journal. The three reviewers (Dr. Marsh among them) were positive, but the associate editor of *JEP:LMC* rejected the submission: "To be clear the reanalyses are certainly important, but they are of limited scope. Perhaps with an additional experiment that replicates and extends the findings the current paper would make more of an independent contribution" (2019). This view is problematic because the reanalysis of the data by Marsh et al. concerns directly the validity of the authors' claim for a dissociation. Importantly, because this view prioritizes the novelty of a claim over the truth of the claim (Nosek et al., 2012), it hinders the critical process of self-correction in science (Ferguson and Heene, 2012).

Second, I might not have been able to conduct this reanalysis if John Marsh had not shared his data willingly. Therefore, any data should be readily available for reanalyses (as in the present case) and future meta-analyses. For example, previous research on differences between interference by process and attentional diversion may have been confounded because studies used different setups (Körner et al., 2017). Although the design by Marsh et al. (2018) avoids this confound, easy access to the data from previous studies would allow an exploratory meta-analysis to study whether effects appear to differ. Therefore, scientists should consider a publication without shared material and raw data as incomplete.

Third, because Bayesian analyses refer to terms such as strength of evidence, results may appear to be more robust than those from null hypothesis significance testing. Although it is true that Bayesian results are more robust than p values (Dienes, 2008; Wagenmakers, 2007; Wagenmakers, Marsman, et al., 2018), they can be "b-hacked" nonetheless (Savalei and Dunn, 2015). To illustrate, the direct analysis may seem optimal, but this may simply be an illusion driven by hindsight bias. Similarly, I could have computed many Bayes factors for various alternative hypotheses, picked the largest one, and come up with a convincing post-hoc rationale (assisted by cognitive biases) regarding why this is the optimal alternative hypothesis. Accordingly, Bayesian analyses are not immune to questionable research practices (John et al., 2012; Simmons et al., 2011; Wicherts et al.,

2016).

Fourth, although the direct analysis supports the idea of a dissociation (Marsh et al., 2018), the robustness of this dissociation is unresolved. Progress in science may be described in terms of a cycle of creativity and verification (Wagenmakers, Dutilh, et al., 2018). In exploratory research, creativity is needed to aggregate current knowledge and data into theories. From these theories, hypotheses with specific predictions are derived. In the next step, hypothesis-testing (i.e., confirmatory) research tries to verify these predictions by comparing the predictions with independent data (i.e., data that were not used when developing the theory). Although both processes of the cycle are important, it is critical to distinguish between postdiction (exploratory research) and prediction (confirmatory research) to avoid biases, and this is easily done with preregistration (Nosek et al., 2018). Therefore, researchers in this area should embrace preregistering hypotheses and method to strengthen any claims for confirmatory research.

Fifth, because the present results of a dissociation are encouraging, it seems worthwhile to show that the dissociation is robust. Although a single study may claim an effect, several independent studies need to replicate the effect before it can be considered a scientific fact (Chambers, 2017; Zwaan et al., 2018). The most promising approach for a replication is an *adversarial registered report* (Nosek and Errington, 2020a). Researchers with one theoretical perspective invite researchers with an opposing theoretical perspective to collaborate on a study (or to serve as reviewers). The researchers from both camps need to agree on study design, method, and analyses. Although the researchers do not have to agree on hypotheses (e.g., direction of predicted effect), they have to agree on method and analyses and that results will be informative no matter their outcome (Nosek and Errington, 2020b). Then, a manuscript with introduction, method, and analyses is submitted as a registered report to a journal (Chambers, 2015, 2017). Because the submission does not contain results (as they are unknown), the reviewers evaluate the merits of the idea and the method. After peer review, the submission is locked (preregistered) and receives an in-principle acceptance: If the researchers conduct their study according to the preregistration and interpret the results sensibly, then the final paper will be accepted no matter the results. This approach minimizes biases such as hindsight bias, confirmation bias, and CARKing (critiquing after results are known; Nosek and Lakens, 2014) and promotes a productive research process (Chambers, 2017). For example, if researchers think of alternative explanations upon viewing the re-

sults, then these become hypotheses for follow-up research and are not considered as actual explanations for unexpected outcomes (Nosek and Errington, 2020a). Researchers in this area (and others) are encouraged to embrace this approach because it emphasizes the important role of replication: confronting our current theoretical understanding with new evidence (Nosek and Errington, 2020b).

Conclusion

Results provided strong evidence for a dissociation (crossover interaction) between speech effects and emotion effects. Because the duplex-mechanism account predicts this dissociation between speech effects (interference by process) and emotion effects (attentional diversion) whereas the unitary account does not, Marsh et al.'s (2018) data support the duplex-mechanism account. However, to show that this finding is robust, researchers are advised to replicate this finding in an adversarial registered report.

Author Contact

Stefan Wiens, Psykologiska institutionen, Stockholms universitet, 106 91 Stockholm, Sweden, SU website, +468163933, sws@psychology.su.se, ORCID: 0000-0003-4531-4313.

Author Contributions

Stefan Wiens: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization, writing-original draft, and writing-review & editing.

Open Science Practices

This article earned the the Open Materials badge for making the materials openly available. This is a commentary that focused on re-analyzing the findings of a published article, and as such there are no (new) data. It has been verified that the analysis reproduced the results presented in the article. This study was not pre-registered. The entire editorial process, including the open reviews, are published in the online supplement.

References

Baptiste, A. (2017). *Gridextra: Miscellaneous functions for \"grid\" graphics* (Version 2.3). https://CRAN.R-project.org/package=gridExtra

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Beaman, C. P., & Jones, D. M. (1997). Role of serial order in the irrelevant speech effect: Tests of the changing-state hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(2), 459–471. https://doi.org/10.1037/0278-7393.23.2.459

Bell, R., Röer, J. P., Lang, A.-G., & Buchner, A. (2019). Distraction by steady-state sounds: Evidence for a graded attentional model of auditory distraction. *Journal of Experimental Psychology: Human Perception and Performance, 45*(4), 500–512. https://doi.org/10.1037/xhp0000623

Berrington de González, A., & Cox, D. R. (2007). Interpretation of interaction: A review. *The Annals of Applied Statistics, 1*(2), 371–385. https://doi.org/10.1214/07-AOAS124

Buchner, A., Mehl, B., Rothermund, K., & Wentura, D. (2006). Artificially induced valence of distractor words increases the effects of irrelevant speech on serial recall. *Memory & Cognition, 34*(5), 1055–1062. https://doi.org/10.3758/BF03193252

Buchner, A., Rothermund, K., Wentura, D., & Mehl, B. (2004). Valence of distractor words increases the effects of irrelevant speech on serial recall. *Memory & Cognition, 32*(5), 722–731. https://doi.org/10.3758/BF03195862

Calin-Jageman, R. J., & Cumming, G. (2019). The new statistics for better science: Ask how much, how uncertain, and what else is known. *The American Statistician, 73*, 271–280. https://doi.org/10.1080/00031305.2018.1518266

Chambers, C. (2015). Ten reasons why journals must review manuscripts before results are known. *Addiction, 110*(1), 10–11. https://doi.org/10.1111/add.12728

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice / Chris Chambers*. Princeton University Press.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Palgrave Macmillan.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 72*, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

Dienes, Z., & McLatchie, N. (2018). Four reasons to prefer bayesian analyses over significance testing. *Psychonomic Bulletin & Review, 25*(1), 207–218. https://doi.org/10.3758/s13423-017-1266-z

Ellermeier, W., & Zimmer, K. (2014). The psychoacoustics of the irrelevant sound effect. *Acoustical Science and Technology, 35*(1), 10–16. https://doi.org/10.1250/ast.35.10

Elliott, E. M., Hughes, R. W., Briganti, A., Joseph, T. N., Marsh, J. E., & Macken, B. (2016). Distraction in verbal short-term memory: Insights from developmental differences. *Journal of Memory and Language, 88*, 39–50. https://doi.org/10.1016/j.jml.2015.12.008

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*(6), 555–561. https://doi.org/10.1177/1745691612459059

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Haaf, J. M., Ly, A., & Wagenmakers, E.-J. (2019). Retire significance, but still test hypotheses. *Nature, 567*(7749), 461. https://doi.org/10.1038/d41586-019-00972-7 461

Hughes, R. W. (2014). Auditory distraction: A duplex-mechanism account: Duplex-mechanism account of auditory distraction. *PsyCh Journal, 3*(1), 30–41. https://doi.org/10.1002/pchj.44

Hughes, R. W., Hurlstone, M. J., Marsh, J. E., Vachon, F., & Jones, D. M. (2013). Cognitive control of auditory distraction: Impact of task difficulty,

foreknowledge, and working memory capacity supports duplex-mechanism account. *Journal of Experimental Psychology: Human Perception and Performance, 39*(2), 539–553. https://doi.org/10.1037/a0029064

Hughes, R. W., Vachon, F., & Jones, D. M. (2007). Disruption of short-term memory by changing and deviant sounds: Support for a duplex-mechanism account of auditory distraction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(6), 1050–1061. https://doi.org/10.1037/0278-7393.33.6.1050

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Jones, D. M., & Macken, W. J. (1993). Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(2), 369–381. https://doi.org/10.1037/0278-7393.19.2.369

Kattner, F., & Ellermeier, W. (2018). Emotional prosody of task-irrelevant speech interferes with the retention of serial order. *Journal of Experimental Psychology: Human Perception and Performance, 44*(8), 1303–1312. https://doi.org/10.1037/xhp0000537

Körner, U., Röer, J. P., Buchner, A., & Bell, R. (2017). Working memory capacity is equally unrelated to auditory distraction by changing-state and deviant sounds. *Journal of Memory and Language, 96*, 122–137. https://doi.org/10.1016/j.jml.2017.05.005

Lawrence, M. A. (2016). *Ez: Easy analysis and visualization of factorial experiments* (Version 4.4-0). https://CRAN.R-project.org/package=ez

Lüdecke, D. (2021). *Sjplot: Data visualization for statistics in social science* (Version 2.8.7). https://CRAN.R-project.org/package=sjPlot

Makin, T. R., & Orban de Xivry, J.-J. (2019). Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife, 8*, e48175. https://doi.org/10.7554/eLife.48175

Marsh, J. E., Yang, J., Qualter, P., Richardson, C., Perham, N., Vachon, F., & Hughes, R. W. (2018). Postcategorical auditory distraction in short-term memory: Insights from increased task load and task type. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(6), 882–897. https://doi.org/10.1037/xlm0000492

Müller, K. (2020). *Here: A simpler way to find your files* (Version 1.0.0). https://CRAN.R-project.org/package=here

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience, 14*(9), 1105–1107. https://doi.org/10.1038/nn.2886

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Errington, T. M. (2020a). The best time to argue about what a replication means? before you do it. *Nature, 583*(7817), 518–520. https://doi.org/10.1038/d41586-020-02142-6

Nosek, B. A., & Errington, T. M. (2020b). What is replication? *PLOS Biology, 18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? a tutorial for teaching data testing. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00223

R Core Team. (2016). *R: A language and environment for statistical computing*. Retrieved August 1, 2019, from https://www.R-project.org/

Röer, J. P., Bell, R., & Buchner, A. (2013). Self-relevance increases the irrelevant sound effect: Attentional disruption by one's own name. *Journal of Cognitive Psychology, 25*(8), 925–931. https://doi.org/10.1080/20445911.2013.828063

Röer, J. P., Bell, R., & Buchner, A. (2015). Specific fore-knowledge reduces auditory distraction by irrelevant speech. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 692–702. https://doi.org/10.1037/xhp0000028

Röer, J. P., Körner, U., Buchner, A., & Bell, R. (2017). Attentional capture by taboo words: A functional view of auditory distraction. *Emotion*, *17*(4), 740–750. https://doi.org/10.1037/emo0000274

Savalei, V., & Dunn, E. (2015). Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00245

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). *Afex: Analysis of factorial experiments* (Version 0.27-2). https://CRAN.R-project.org/package=afex

Ströberg, K., Andersen, L. M., & Wiens, S. (2017). Electrocortical N400 effects of semantic satiation. *Frontiers in Psychology*, *8*, 2117. https://doi.org/10.3389/fpsyg.2017.02117

Szucs, D., & Ioannidis, J. P. A. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience*, *11*, 390. https://doi.org/10.3389/fnhum.2017.00390

Team, R. (2020). *Rstudio: Integrated development environment for R*. Boston, MA. http://www.rstudio.com

VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Dutilh, G., & Sarafoglou, A. (2018). The creativity-verification cycle in psychological science: New methods to combat old idols. *Perspectives on Psychological Science*, *13*(4), 418–427. https://doi.org/10.1177/1745691618771357

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169–176. https://doi.org/10.1177/0963721416643289

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician*, *73*, 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*. Retrieved February 1, 2021, from https://joss.theoj.org/papers/10.21105/joss.01686

Wiens, S. (2017). Aladins Bayes Factor in R. https://doi.org/10.17045/sthlmuni.4981154.v3

Wiens, S., & Nilsson, M. E. (2017). Performing contrast analysis in factorial designs: From NHST to confidence intervals and beyond. *Educational and Psychological Measurement*, *77*(4), 690–715. https://doi.org/10.1177/0013164416668950

Wiens, S., Szychowska, M., Eklund, R., & van Berlekom, E. (2019). Cascade and no-repetition rules are comparable controls for the auditory frequency mismatch negativity in oddball tasks. *Psychophysiology*, *56*(1), e13280. https://doi.org/10.1111/psyp.13280

12

Zhu, H. (2020). *Kableextra: Construct complex table with 'kable' and pipe syntax* (Version 1.3.1). https://CRAN.R-project.org/package=kableExtra

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*, e120. https://doi.org/10.1017/S0140525X17001972