Reviewer comments in black, *author responses in blue italics*

**Major Revision**

I have now received three very helpful and detailed evaluations of your submission. The reviews are overall positive and I believe your submission is a promising candidate for eventual Stage 1 in-principle acceptance. The reviewers offer a range of constructive comments and suggestions, which I won't attempt to exhaustively summarise, but I will highlight a few points that struck me as particularly salient based on my own reading of your submission.

First, it seems important in this case to ensure that the replication is as similar as possible to the original study in critical areas of the design that could explain any potential differences in outcomes. One example highlighted in the reviews is the choice of stimuli. This is an understandable (and justifiable) deviation considering the target sample but also raises issues to consider. One possibile solution would be to run two replications in parallel, one using original English stimuli and the other using German stimuli, perhaps validated for suitability as Reviewer 1 proposes. I will leave you to consider this possibility -- it is not a concrete requirement but a point that warrants careful consideration. I would also suggest including an additional table in the manuscript that highlights key similarities and differences in the methods of the replication and the original study.

Other points of note include clarifying (and further justifying) the rationale for adjudicating between the different theoretical accounts, clarifying the justification of the smallest effect size of interest in the sampling plan (Reviewer 2), and making sure the methods are as reproducible as possible (Reviewer 3, Massimo Grassi). I realise not all researchers use R (and it is not mandatory to do so), but whatever approach is used, the goal expressed by the reviewer should be met to ensure that the sampling plans and analyses are as scripted as much as they can be, fully explained, and computationally reproducible.

*We thank the editor for the invitation to revise the manuscript as well as for highlighting the crucial points in the reviewers‘ comments to be addressed in the revision.*

*We can see that both the editor and three reviewers provided very thorough and extensive comments on our manuscript, and we hope these will be instrumental in improving the quality of our paper. We have taken their suggestions into careful consideration and successfully incorporated most of them into the revised version. These changes have been clearly highlighted throughout the manuscript for easy reference.*

*In the following sections, we present our point-by-point responses to address all the reviewers' concerns and queries thoroughly.*

*As to contrasting similarities and differences between our replication and the original study: Based on your and the reviewers’ concerns (in particular Reviewer 3), we have adopted the strategy to reduce the differences between our planned replication and the original to a minimum: (1) We are now using the original distractor sounds with English pronunciation of the letters V, J, X. (2) We have removed the feedback we were going to implement for motivation. (3) We have adjusted the timing of presentation and retention intervals to what most likely was what Jones & Macken (1995) used. The only remaining*

*difference is that we ask subjects to click on a numerical array to enter the digits in the order they recall, rather than having them write down numbers (see our response below).*

**Review by anonymous reviewer 1, 30 Jun 2023 16:35**

This paper proposes to replicate the study of Jones and Macken (1995) which showed that the disruptive effect of changing-state sound on serial recall is attenuated when the changing-state stimuli are presented in the spatial domain in such a way as to try to induce the perception of a number of 'steady-state' streams. The Jones and Macken study was quite ingenious and in general I welcome the present authors' proposal to attempt to replicate their Exp 1c (which had the most comprehensive set of conditions). However, I have a number of concerns about the current proposal.

Main points:

1. I welcome places like PCI RR as (partly) an outlet for publishing replications (or replication attempts) even if that is the main or sole purpose of the study in question. However, I believe the methodology of such studies should adhere as closely as is reasonably possible to the original study. And indeed, the present authors state early in the paper that that is their intention here: "Jones and Macken's (1995a <<should be 'b' by the way) Experiment 1c, which contains all relevant control conditions including and beyond those used in Experiments 1a and 1b (and Jones et al., 1999), is to be replicated as closely as possible". However, when I read on, I found that this wouldn't in fact be the case if the authors were to proceed with the current proposal. There are three methodological changes from the original experiment, none of which I think is necessary:

*We thank the reviewer for the overall positive evaluation. The year "1995b" has been corrected in the citations throughout the manuscript. We have revised the manuscript to reduce the methodological discrepancies between the original study and our replication to a minimum. Our responses to these methodological changes and therevisions we made can be found below.*

i) The authors propose to use (computer-generated) German-pronounced versions of "V", "J", and "X" instead of the English-pronounced versions of the same letters used by Jones and Macken (1995). This is understandable in one sense because the authors will draw on a pool of participants in Germany and want the speech to be in the participants' first/dominant language just as it was in Jones and Macken's (1995) study (conducted in the UK). However, on balance, I think that it would be 'safer' to use the same English pronunciations as used by Jones and Macken (1995) (my understanding is that these three same letter-names are pronounced quite differently in the two languages). Of course, I don't mean the exact same sound recordings but rather that the authors use a computer-generated (native English) voice to produce the speech tokens with (broadly) the same pronunciation as those used by Jones and Macken. The danger with using different auditory tokens is that one set of auditory tokens, or combination thereof, may not stream apart as well as another set of auditory tokens. There may, for example, be higher acoustic or post-categorical transitional probabilities between certain speech sounds serving to inhibit the fission of successive tokens into separate streams (Bregman, 1990; Warren, 1999). Moreover, we know that the

changing-state effect is not modulated by linguistic factors: the effect, as well as, importantly, its modulation by streaming cues is also found with pure tones (Jones, Alford, et al., 1999, JEPLMC). Thus, on balance, I think it would be better to match the Jones and Macken experiment in terms of the letter-names used rather than match it in terms of the language of the speech being the same as the dominant language of the participants.

An alternative solution (though not my preference) would be to conduct some sort of independent empirical check on whether and how strongly and how consistently the (German-spoken) tokens stream apart, ideally involving a comparison with English-pronounced “V”, “J”, and “X”. I noticed that Jones and Macken did such a check in relation to their mono vs stereo CS conditions and so something similar could be done here re: the German tokens vs the English tokens.

*Thanks for pointing out that deviation from the original protocol: Actually, the German pronunciation of the letter sequence V, J, X, i.e. [faʊ], [jɔt], [ɪks], is quite similar to the English sequence of letter words, in that each letter word contains a unique and distinct combination of phonemes. Thus we did not expect them to be perceptually segregated / streamed differently a priori. Nevertheless, given that we don’t know the acoustic or post-categorical transitional probabilities of these sounds, and to accommodate the reviewer’s legitimate concern, we decided to use the English pronunciation of the letter words V, J, X, i.e., /viː/ /d͡ʒeɪ/ /ɛks/ using also a female voice with a British accent as in Jones and Macken’s (1995) experiment 1C for the replication; particularly as the bulk of research on the irrelevant speech effect dating back to Colle and Welsh’s (1976) or Salamé and Baddeley’s (1982) seminal work has shown that interference by utterances in a foreign language or by nonsense syllables is comparable to that by one’s own native language (although we found Japanese to be slightly more disruptive to Japanese participants compared to other languages, but this may be due to ideosynchracies of the Japanese language; Ellermeier et al., 2015).*

ii) The authors propose to use an order reconstruction response mode where participants are shown an array of the to-be-remembered letters and the participant must click them in the order in which they were just presented whereas Jones and Macken (1995) had participants write down their responses. Whilst transcribing written responses is more time-consuming, a written response mode should nevertheless be used in my view.

*We see the reviewer’s concern with aiming for a literal replication, but on the other hand we would prefer to use a modern, computer-based acquisition method to reduce potential errors occurring in transcription and entering data. We don’t think that such a change from written to a clicked response will threaten the interpretability of the outcome of the replication, given that both “writing the sequence down on paper” and “sequentially clicking on a number array” are order reconstruction responses. Thus, we argue that a clicking response should equally rely on serial-order processing and in particular the retrieval processes from short-term memory should be identical (mental order reconstruction). In line with what Reviewer 3 suggests, we believe that this is still a “direct replication”, even if we use a modern response collection technique.*

iii) The present authors propose to provide feedback on the participant's performance after each trial whereas this was not the case in Jones and Macken. Again, please remove this unnecessary addition to the original experiment.

*We agree. The feedback has been removed to be consistent with the original study.*

Whilst I admit that it is not obvious why any of these methodological features--either alone or in some combination--should make a difference to the outcome, if a different outcome is indeed observed, readers may well wonder whether one or more of them did indeed have an effect, which would not be satisfactory when the sole purpose of a study is to replicate a previous one.

2. P. 2 "Jones and Macken's study, which has never been replicated by a different laboratory, to our knowledge,…"

Whilst it's true (to the best of my knowledge too) that Jones and Macken's (1995) streaming-by-location study of the changing-state effect hasn't been replicated by a different laboratory, there are some highly relevant demonstrations of effects of streaming on auditory distraction during serial recall (one of them from a separate lab) that currently go unmentioned in the manuscript:

i) Jones, Alford, Bridges, Tremblay, & Macken (1999, JEPLMC, Vol. 25, No. 2, 464-473) showed that having a particularly large difference in pitch between successive tones—again the idea being to induce their perceptual segregation into two steady-state streams—attenuates the changing-state effect.

ii) Macken et al. (2003, JEPHPP, Vol. 29, No. 1, 43–51) showed the attenuating effect on the changing-state effect of streaming induced by increasing the rate of presentation (whilst keeping the content, including the pitch, of the tokens constant).

iii) Although not quite as relevant to the changing-state effect per se, Hughes and Marsh (2017. JEPHPP, Vol. 43, No. 4, 537–551) showed that the order incongruence effect—the particularly pronounced disruption to serial recall caused by sound tokens post-categorically identical to the to-be-remembered items but only when they are presented in an incongruent order with the to-be-remembered items (Hughes & Jones, 2005)—disappears when the successive sounds in an objectively order-incongruent sequence are presented from two different 'locations' (presented to each ear in an alternating fashion) and in two different voices so as to render it, perceptually, no longer order-incongruent.

One general point that arises from the fact these relevant studies appear to have been missed by the current authors is that rather too much is made of the notion that Jones and Macken (1995; and Jones et al. (1999) is the only demonstration of the key phenomenon of interest. Whilst it is true that these other studies didn't focus on spatial location cues (though Hughes and Marsh, 2017 did include spatial location cues), I don't think the fact that Jones and Macken (1995) used location cues specifically is as important as the authors make out – the key question is whether streaming cues (whatever their nature) modulate auditory distraction effects in serial recall. The findings of some of these other studies are also relevant to some of the other concerns raised below.

*Indeed, as Reviewer 1 is pointing out, there have been other demonstrations of the effects of auditory streaming on the magnitude of the ISE, namely (a) streaming by*

*frequency separation (Jones et al., 1999) and (b) streaming by (increased) tempo (Macken et al., 2003), suggesting that perceptual organization plays a significant role. Nevertheless, we think that streaming-by-location is an interesting case deserving study (and replication) in itself, particularly, since the spatial arrangement of the distracting sound may be important for practical applications, such as the acoustical layout of open-plan offices or the rendering of multiple sound sources in telecommunications environments.*

*The theoretical significance of the studies the reviewer pointed to for the central concept of auditory streaming (points i) to iii) above), however, led us to incorporate two of these references into the intro section of the present revision (p. 9). The third study iii) by Hughes and Marsh (2017) is of particular interest, because it makes use of spatial streaming in Experiment 1 to test a particular account of auditory interference, but since it is hard to integrate into our rationale for the present replication (the reviewer characterizes it as "not quite as relevant"), we would rather refer to it in the discussion section of our article.*

3. P. 7. "In our view, these steady-state control conditions from Experiment 1c are crucial as they constitute the true "steady-state" reference and allow to tease apart the irrelevant speech effect (comparison with silence) into a changing-state effect (changing condition vs. steady condition) and a steady-state effect (steady condition vs. silence). Note that this kind of reference is needed to be able to assess whether the release from interference caused by spatial streaming (in the "V-J-X" stereo condition) reduces to a perceptual steady-state condition (or three steady-state streams, according to Jones and Macken's reasoning), or whether some sort of changing-state effect remains (i.e., changing-state stereo being slightly more disruptive than steady-state stereo)."

And, relatedly, on p. 10 "…the results are not as clear-cut as Jones and Macken (1995a)'s interpretation suggests: While the critical "stereo" condition affording spatial streaming into three steady-state sources should completely abolish the changing-state effect (due the eliminated interference with serial-order processing)…"

I don't agree with the authors here. Whilst it was/is ideal to include the SS conditions (for one thing, their inclusion showed the effect wasn't driven merely by the spatial changes in the CS-stereo vs CS-mono conditions, as argued by Jones and Macken), they were/are not crucial for the theoretical conclusions of Jones and Macken (1995). As the present authors say in their previous paragraph, "the crucial prediction made by Jones and Macken (1995b) is that the changing-state mono condition should be significantly more disruptive…than the stereo condition..". Indeed, Jones and Macken made no prediction regarding a possible ('residual') difference between CS-stereo and SS-stereo, no doubt because they were well aware that streaming is an unstable and not an all-or-none phenomenon. The strength of streaming could thus easily vary between trials or/and individuals. Indeed, Jones, Alford et al. (1999), using pitch as a streaming cue, did find a 'residual' effect in a streamed CS condition vs a 'true' SS condition and stated that: "This may be because the effect of having two steady-state streams is more disruptive than having one steady-state stream. *More likely, as studies of the phenomenology of streaming suggest, the process of stream separation is relatively unstable, so that the two channels are not rendered consistently*. In any case, whatever the precise level of performance in relation to the repeated-token conditions, analytically, the significant improvement in going from 5 to 10 semitones seems to suggest quite

strongly that streaming processes were at work and that these modulated the disruptive effect of irrelevant sound." (p. 471, asterisks added).

Thus, I don't believe that a residual CS effect would speak to the veracity of the interference-by-process account. It is certainly not the case in my view, therefore, that "If the spatial separation of sound sources (via stereo) does not reduce distraction to the level of "steady-state" speech, the strict interpretation from the original study is refuted, thus challenging the interpretation given by Jones and Macken (1995)." (from Table on p. 21).

I also don't really see why being able to decompose the disruption into a 'changing-state effect' and 'steady-state effect' is important, at least in the present context. It has already been demonstrated that a steady-state effect can be detected with enough statistical power (Bell et al., 2019); examining whether that effect is observed again here seems tangential to the main point of the proposed experiment.

In sum, I agree that the two SS conditions should be included in the proposed experiment (as they were in Jones and Macken's Exp 1c) but I'm not at all convinced by the current authors' justifications for why they are necessary; they are worthwhile additions but not "crucial" for the theoretical conclusions of Jones and Macken (1995) or those drawn in the related papers by Jones, Alford et al. (1999), Jones et al. (1999), and Macken et al. (2003).

*Thank you for this valuable comment clarifying the theoretical implications of the inclusion of the steady-state control conditions in Jones and Macken's (1995b) Experiment 1C. We agree that neither a residual changing-state effect in the stereo condition nor the observation of a steady-state effect would challenge the interpretation of the results in terms of steaming-by-location.*

*We now included the interpretations given in the original article and by Jones, Alford et al. (1999) for the residual steady-state effect: (a) additive disruption produced by 3 steady-state streams and (b) unstable streaming (on p. 8 of our manuscript). We also toned down our assertion that showing a null effect for the steady-state conditions is crucial (on p. 11 of the manuscript and in the design table). Nevertheless - as agreed by Reviewer 1 - we maintain that these control conditions should be included in the replication, since they were only part of one of three experiments addressing the spatial-streaming hypothesis.*

4. The authors seek to pit the interference-by-process account against the attentional account (Bell, Roer et al.). However, I don't think the attempt works. There are, for example, some obvious difficulties with the attentional account of the changing-state account—both in general and in relation to the effects of streaming in particular—that go unmentioned.

For example, the authors reiterate the attentional account's proposition that "the changing-state effect may arise because a sequence of changing sounds is less predictable than a steady-state sequence" (p. 4) and yet the degree to which a changing-state sequence disrupts performance is not a function of its predictability (e.g., Hughes & Marsh, 2020; Tremblay & Jones, 1998; Jones et al., 1992).

Furthermore, an (unpredictable) changing-state sequence does not disrupt non-seriation based processing (e.g., Hughes & Jones, 2020).

In relation to streaming studies in particular, it is difficult to see how the attentional account can explain the fact that having more tokens per unit of time can *reduce* the degree of disruption (Macken et al., 2003): having more tokens per unit of time should surely require more attentional resources for monitoring and hence increase the amount of disruption?

In relation to the proposed experiment, the authors state that, on the one hand, the attentional account predicts greater disruption in the CS-stereo condition than the CS-mono condition because in the former condition there are changes occurring in terms of both spectral content and spatial location (see p. 11). That is, it predicts the opposite result to that found by Jones and Macken (1995). So far, so good: the accounts make opposite predictions (though it doesn't bode well for the attentional account given Jones and Macken's, 1995, results). However, in the next sentence, the authors state that "Only if some kind of auditory stream segregation - or, more akin to the model, divided attention - is postulated, will Jones and Macken's (1995b) "stereo" condition boil down to three steady-state streams.". Thus, the authors seem to be suggesting here that the attentional account can also explain the result obtained by Jones and Macken (1995b). This is a problem in itself in that the attentional account is clearly too ill-specified and hence too flexible to serve as a theoretically useful counterpoint to the interference-by-process account, at least in the present context. This is especially the case given that we are told that the account would have to appeal (and in an ad-hoc way it seems) to the same mechanism (auditory stream segregation) as the interference-by-process account. I didn't understand what the authors meant by "or, more akin to the model, divided attention": Given that the irrelevant sound effect itself is explained on the attentional account in terms of attentional resources being divided between the focal task and the sound, wouldn't having to divide attention further 'within' the irrelevant sound again be expected to increase the level of disruption still further, not decrease it?

The authors go on to state that "The stereophonic control condition of Jones and Macken's Experiment 1c, where one and the same letter alternates between three locations, might be thought to require slightly more attentional resources than the monophonic steady-state condition (repetitions of the same letter at a single location), since the changing spatial position of that letter (driven by the interaural level differences) constitutes a changing sound feature."

If this is the case, then surely it's the first prediction above that must hold in relation to the CS-stereo vs CS-mono conditions (i.e., greater disruption in the CS-stereo condition than the CS-mono condition because in the former condition there are changes occurring in terms of both spectral content and spatial location) and not the second prediction?

Perhaps it is because clear predictions cannot be derived from the attentional account that the implications for this account of various possible outcomes are not included in the table on p. 21?

One general observation arising from main points 2-4 above, together with minor point ? below is that none of the 'added' justifications offered by the current authors for the replication attempt are convincing. But then such additional justifications may not be needed; a simple (faithful) replication may well be of sufficient value in and of itself.

*We thank the reviewer for pointing out that our predictions derived from the "attentional account" were somewhat contradictory. We have tried to remedy that in three ways without going as far as entirely deleting the reasoning based on that account:*

1. *We have tried to support the theoretical assumptions made by the attentional account by additional references (on p. 4).*
2. *We have entirely scrapped the sketchy explanation (2) recurring to unsubstantiated concepts of 'divided attention' or 'streaming' which actually resulted in the same predictions as the interference-by-process account.*
3. *We added a speculation about the predictability of the irrelevant-sound sequences, which should be perfect after less than a second of irrelevant-sound playback, given that the sequence just cycles through 3 letters in a fixed order.*
4. *We have added a disclaimer, saying that it is hard to derive stringent predictions for this paradigm from the attentional capture account (on p. 13).*

Minor points

1. There is some mixing up of the citations to Jones and Macken (1995a) and Jones and Macken (1995b) at various points through the manuscript and there are also occasions on which it is ambiguous (i.e., "Jones and Macken, 1995", especially as there was a third 1995 paper by Jones and Macken!)

*Thank you for this comment. We have corrected all the ambiguous or incorrect citations.*

2 P. 3. "While the earliest theoretical explanation as to why irrelevant speech effects occur, focused on interference-by-content between articulatory rehearsal of the to-be-remembered verbal material and the automatically encoded phonological elements of irrelevant speech in working memory (Baddeley & Hitch, 1974),…"

First, I think Salamé and Baddeley (1982) should be cited here rather than (or as well as) Baddeley and Hitch (1974) as the latter paper predates the discovery of the irrelevant speech effect. Second, the description of the theory isn't quite right (at least as of Baddeley, 1986, onwards): An important discriminating feature of the phonological loop-based account of the ISE is that the irrelevant speech disrupts the passive phonological store and not articulatory rehearsal (e.g., Baddeley, 2007; Jones et al., 2004, JEPLMC).

*We thank the reviewer for this important point. The explanation in line with the phonological-loop account has been corrected, now referring to interference in the phonological store rather than with the rehearsal process. Furthermore, Salamé and Baddeley's (1982) study is being cited now.*

3. p. 8 "This result is (1) significant in the web of empirical results determining what factors modulate memory disruption by irrelevant speech, particularly, since, among the many acoustical factors modulating the irrelevant speech effect (Ellermeier & Zimmer, 2014), it is the first to demonstrate an effect of the spatial layout of the irrelevant sound background in affecting the amount of auditory distraction".

It's not entirely clear what these 'many acoustical factors' are? The key factor is changing-state/acoustic variability, which is what Jones and Macken (1995) were examining.

*We added examples of these acoustical factors in parentheses, and further characterized them as psychoacoustical. This is alluding to a research tradition trying to find physical metrics predicting the magnitude of the ISE; and reiterating the point that Jones and Macken (1995b) is the first instance to investigate the effects of* spatial *audio on the ISE.*

4. p. 9 “Though it does so with only minimal, highly artificial means to “spatialize” the sound image, i.e. by using monotic (left, right) vs. diotic (center) headphone presentation, resulting in a lateralized or central stream in the listener’s head; no other study to date has replicated this classic study with more sophisticated means of generating spatially localizable sound sources, such as loudspeaker arrangements or headphone-based binaural techniques (Hammershøi & Moller, 2005), thus generating realistic externalized sound images rather than producing an “in-the-head” localization.”

When I read this, I feared that the present authors’ proposed experiment was going to deviate from using the same (“highly artificial”) method. Gladly, I saw from the Method section that it is not but this paragraph may mislead other readers too.

*We just tried to put into perspective how impoverished the spatial audio is with in-the-head-localization. Nevertheless, we now deleted all reference to other, potentially more natural, means of spatialization, such as loudspeaker arrays or binaural technique. As you say: That will only mislead the reader. That point is reinforced by Reviewer 3.*

5. p. 10 “Second, Jones and Macken’s (1995b) study is statistically underpowered: Even though they obtain a spatial-streaming effect (statistically significant difference between their “mono” and “stereo” conditions) in three separate experiments (1a-c), each of them is based on data from twenty participants only”

I didn’t find this argument very convincing – the fact that three separate experiments obtained the same reliable effect makes it very likely that a reliable ‘overall’ effect would have been found in a cross-experiment analysis (i.e., with an n = 60, which is similar to the n of 54 proposed for the current experiment). Moreover, the effect was replicated again (as the authors note) in Jones et al. (1999).

*We agree that there is definitely more evidence than a single low-powered experiment, and we have changed the argument. Although there are 3-4 separate low-powered experiments by Jones et al., in which the effect has been observed (all using slightly different design, stimuli or procedures), we believe that an independent replication of Exp. 1C with (1) enhanced power within a single experiment and (2) conducted by a different laboratory and with a new, independent sample is important given the theoretical significance of that study.*

6. p. 10 “That might suggest that this is not just a steady-state effect (which typically is hardly measurable with so few participants, see Bell et al. 2019) but a residual changing-state effect, potentially due to the spatial switching between locations”

First, as noted in main point 3, any ‘residual changing-state effect’ could be due to imperfect/unstable streaming. Second, Jones and Macken’s (1995b) Exp 1c was specifically designed to examine whether spatial switching per se could produce some disruption and they found no support for that.

*We thank the reviewer for this comment and we added the information and state that the steady-state effect is more likely to be due to imperfect or unstable streaming.*

7. I don't think the authors mention the intensity level at which the sounds will be presented but presumably this will be 65dB(A) as in Jones and Macken's study?

*Thanks. We added the information on the sound pressure level.*

Very minor points:

1. Abstract "…that provided evidence for the spatial separation of sound sources to reduce the detrimental effects of irrelevant speech on short-term memory"

The wording is awkward here (and it doesn't quite work). I'd suggest: "…that showed that the separation of successive sound stimuli to distinct spatial sources reduced the detrimental effect of irrelevant speech on short-term memory"

*We rephrased the sentence, and we used "locations" rather than "sources" to avoid suggesting that the sound images were externalized.*

2. The authors use the term 'unitary attentional account' but the term 'unitary' won't make sense to an uninitiated reader unless the contrasting 'duplex' account is also introduced. I think 'attentional account' is sufficient in any case (or 'attentional-capture account' would be even better as the interference-by-process mechanism can also arguably be couched in 'attentional' terms, e.g., Hughes & Marsh, 2017).

*We agree, the term "unitary" has been removed.*

3. In a number of places, the authors refer to Jones and Macken's 'classical' experiment. I believe the authors mean 'classic' ('classical' means traditional) as indeed they do use 'classic' at other points.

*We now use "classic" instead of "classical".*

4. P. 5 "was that both a "mono" version of the irrelevant letter sequence was contrasted with a "stereo" version."

>>delete "both" (one cannot contrast both of two things; each is contrasted with the other).

*The word "both" was deleted.*

5. p. 7 "more disruptive to the serial recall of information stored in short-term memory"

The wording is a bit unfortunate here because the Jones and Macken (1995) study formed part of what was (or what certainly became) their view that there is no such thing as 'short-term memory' for things to be 'in'! I would suggest just saying '..more disruptive of serial recall than the stereo condition'

*Thanks. The reference to "short-term memory" has been removed.*

6. P. 13 "If, in fact, spatial streaming reduces the changing-state effect by producing all but a steady-state effect…"

*The wording has been corrected.*

I think the authors meant to say 'nothing but a…' not 'all but a…' here.

*Yes, it has been corrected.*

**Review by anonymous reviewer 2, 20 Jun 2023 13:17**

Let me first emphasize that I think that it is a great idea to perform a direct replication of the study of Jones and Macken (1995) with increased statistical power. As the authors write, the study has had a huge impact on theories on auditory distraction. The finding is cited to this day as a key finding that theories on auditory distraction have to explain. Considering the strong impact on the field, it is indeed noticeable that independent replications of this finding are still missing. As noted by the authors of the Stage-1 manuscript, there are replications of the same research group (Jones et al., 1999) but no independent replications of other research groups. I fully agree with the authors that a replication is needed based on the fact that the original study as well as the replication study had only small sample sizes and were thus seriously underpowered, combined with the fact that there are already failed replications in the literature of studies with similar characteristics (e.g., Kvetnaya, 2018). The study is a straightforward replication that will be highly useful for providing a better empirical basis for theories on auditory distraction regardless of the outcome of the statistical hypothesis test.

The State-1-manuscript is very well written, the literature review is up to date, and the theoretical reasoning is convincing. Overall, I think that this is a great proposal.

*Thank you for the positive overall evaluation of our manuscript.*

I fully agree with the authors that the original study by Jones and Macken (1995) as well as the replication study by Jones et al. (1999) were seriously underpowered and I think that it is great that the authors plan to repeat Jones and Macken's (1995) study with more than double the sample size. Nevertheless, I wondered whether the power calculations are still too optimistic. How did the authors arrive at an effect size estimation of f = .25? Did the authors use the default options of G-Power? Is the correlation among the levels of the repeated-measures factor included in the effect size measure? Furthermore, it seems to me that the sample-size planning did not consider that the authors intend to use Holm corrections of the alpha level for multiple comparisons which should decrease the sensitivity of the statistical test. As a side note, I found it somewhat confusing that the term "changing-state effect" (at least to my understanding) is used to refer to the comparison between changing-state speech and steady-state speech as well as to the comparison of steady-state speech between mono- and stereophonic presentation modes. I think that the term "changing-state" should be reserved to the comparison between changing-state speech and steady-state speech. As a possible way to move forward, the authors may consider highlighting one specific statistical test as the decisive test of the hypothesis instead of correcting for multiple comparisons. The other tests can then performed as part of supplementary exploratory tests.

*We thank the reviewer for this important point. We replaced the previous power calculations using G*Power with a simulation-based power estimation tailored to the specific experimental design and the planned analyses. That is, we ran 1000 iterations of the experiment with varying sample sizes and determined the statistical power to observe both a main effect of auditory condition and the crucial contrast between stereophonic and monophonic changing-state conditions (using Holm corrections), assuming means and standard errors as depicted by Jones and Macken (1995b, Fig. 3). We also made sure to use the term "changing-state effect" exclusively to refer to the contrast between changing-state and steady-state conditions.*

Minor issues

I think that Jones and Macken (1995) had a 7-seconds presentation phase and a 10-seconds retention interval. It is true that they wrote that the had created a sound recording that lasted approximately 20 seconds but the fact that the sound recording lasted 20 seconds does not imply that it was played for 20 seconds. Jones and Macken (1995) are clear in the description of the procedure that the presentation phase was 7 seconds and the retention phase was 10 seconds, resulting in a total exposure of 17 seconds. I don't think that this is a critical issue because, of course, the authors of the Stage-1 manuscript can argue that exposure is maximized by presenting the sounds for the full 20 seconds, thereby increasing the chances of finding effects if they existed, but I wanted to mention it because it is a deviation from the original study that does not seem to be strictly necessary.

*Thank you for bringing up this important point about stimulus and trial durations, which is somewhat ambiguously described in Jones and Macken (1995b). We'll stick to the 7 s presentation + 10 s retention interval, totaling 17 s for the irrelevant sound and we have specified the description of the stimuli and procedure accordingly.*

According to APA, quotation marks should not be used to highlight key phrases, therefore I suggest to remove the quotation marks highlighting "stereo", "mono", "irrelevant speech effect", "changing-state effect", "unitary attentional account", "spatial planning", "changing-state hypothesis", "streams", , "steady state", "changing state", "stream out", "interference-by-process", "spatialize", "in the head", "babble speech", "standard babble", "streamed babble", and "babble".

*Thank you for that suggestion, and for checking with APA style: We removed what felt like 100 quotation marks, except for the first mention of "mono" and "stereo", since that usage by Jones & Macken is not quite what is meant by it in audio reproduction. After we define those as "monotic" and "diotic", we do not use the quotation marks any longer.*

Page 10: "While the critical stereo condition affording spatial streaming into three steady-state sources should completely abolish the changing-state effect (due the eliminated interference with serial-order processing), the memory disruption observed in all three experiments is still substantial and – showing mean performance between the silent control and the most disruptive monophonic changing-state condition (see Figure 2)." – I found this sentence hard to follow – please revise.

*Thanks for pointing that out: We reworded, simplified, and split it into 2 sentences.*

Page 11: "differences differences" –-> differences?

*Corrected.*

Page 12: Please note in the figure caption what the error bars stand for and please add an explanation as to why the error bars are not displayed for the steady-state conditions.

*We added a sentence to the figure caption explaining why no error bars are available for 2 control conditions in Experiment 1C, where only a single value was read off the graph.*

Page 12: "low statistical power and unequivocal results" —> "low statistical power and ambiguous results"?

*We agree, "ambiguous" makes more sense here.*

Page 16: “test feedback” —> feedback?

*It should be “text feedback” and has been corrected.*

**Review by Massimo Grassi, 20 Jun 2023 11:46**

I read the stage 1 RR by Hassanzadeh et al. In my opinion the work is good, but there is some work to be done on two sides:
1. make more explicit the direct comparison between original study and current replication
2. improve the computational replicability of the protocol submitted by the authors

*We thank the reviewer for the positive evaluation and we attempted to carefully address both issues in the revision.*

1. In the paper, authors stress a lot the theory and the importance of the original study. This is certainly correct. But when it comes to replication, they should stress more similarities and (more importantly) differences between original and replication. I guess authors are familiar with the distinction between direct and indirect replication (Zwaan et al. 2018). This looks like a direct replication to me. When it comes to a direct replication (a hypothetical ideal dream in psychology) authors should stress explicitly the differences between the original experiment and the replication and ask themselves whether the differences are theoretically relevant or not and they may impact the results (see references) because in the case the results of the replication differ from the original, we have two possible scenarios: if the replication is a direct replication we can criticize the original result. If it is not, there is not much that can be done. If differences are theoretically relevant authors should report them. In particular, in the description of method and analysis I would report here and there the differences. We tried to do something similar in our paper replicating the attentional blink (see references). Note that in many cases we do not know whether differences may be theoretically relevant, but if authors acknowledge them, at least they will be evident for the future readers. Differences I noticed: different recorded voice (is it relevant?).
[By the way, I didn't understand why authors criticize the stereophonical simplicity of the original experiment but then replicated exactly the same type of stereophonic stimuli (page 9). Note that I do prefer the direct replication of the original stimuli!].

*Yes, this is meant to be a direct replication! Based on your criticism and that by Reviewer 1, we have reduced the differences between our planned replication and the original to a minimum: (1) We are now using the original distractor sounds with English pronunciation of the letters V, J, X. (2) We have removed the feedback we were going to implement for motivation. (3) We have adjusted the timing of presentation and retention intervals to what most likely was what Jones and Macken (1995b) used. The only remaining difference is that we ask subjects to click on a numerical array to enter the digits in the order they recall, rather than having them write down numbers (see our response to Reviewer 1, Main point 1, ii).*

*As to mentioning that Jones and Macken (1995b)’s “spatialization” of the sound is of the simplest, and most unnatural kind conceivable, following your suggestion (and that of Reviewer 1), we now just leave it at that, and abstain from even naming the spatial-*

*audio alternatives suggesting themselves - having realized that this will only divert readers' attention from the main idea of this replication project.*

2. I would make the materials (softwares) reproducible. May I ask the authors to translate the power calculation into R-language? g*power is rather opaque and in many cases it is not clear the power calculation that is performed. The same applies for the analysis scripts that will be used for the statistical analysis. Note also that it was unclear how the power analysis was exactly calculated.

*We replaced the G*Power analysis with a simulation-based power analysis conducted in R. The code has been made available in the OSF repository. In the revised manuscript, we are providing a detailed description of how the simulations were conducted to determine the statistical power to obtain the crucial effects with our sample size.*

3. Suggestion. Authors claim they want to conduct an extremely powerful replication (in statistical terms). If this is true and the outcome of the experiment will be a super clean result (super replicable) an independent replication of the replication maybe less powerful (see our paper) should return the same result. This would strengthen the result of the study.

*Thanks for the references you give at the end of your review, highlighting the differences between different strategies one may adopt in replicating empirical studies. That was interesting to study! We decided to explicitly point out on p. 14 that we'll attempt a "direct replication", citing Zwaan et al. (2018).*

Minors
- I do not understand exactly the "errors". If I type one letter in the incorrect position it is certainly an error (I guess). What about the question mark? Is this still an error? Note that typing something sets a floor for correct guesses.

*Since this is serial recall, the question mark serves as a placeholder for a digit the participant did not remember (the identity of), but it enables them to enter the next couple of digits in the correct serial position. Of course, entering a question mark scores as incorrect/not-recalled at the respective serial position.*
- is there any filter for outlier performances? I would use the same of the original experiment

*Since we saw no hint at data cleaning in the report of the original study, we will use all data, even if they show ceiling or floor performance.*
- page 13. "statistically indistinguishable". Please operationalize this criterion.

*We replaced "statistically indistinguishable" by "not significantly different".*