# Responses to Action Letter

Revised manuscript for *JEP:G*
"The specificity of sequential Statistical Learning: Statistical Learning accumulates predictive information from unstructured input but is dissociable from (declarative) memory"

*Dear Dr. Endress,*

*I have received reviews of the manuscript entitled The specificity of sequential Statistical Learning: Statistical Learning accumulates predictive information from unstructured input but is dissociable from (declarative) memory (XGE-2021-3907) that you recently submitted to Journal of Experimental Psychology: General. I was fortunate to receive comments and evaluations from individuals who are very knowledgeable and highly respected experts in the topical area you are investigating.*

*Based on the reviews and my own reading, even though I am enthusiastic about the goals and general strategy of your research, I must reject this particular submission. The reviewer comments are mixed, and I will mention key points that drove my decision.*

> We were very pleased to read that the editor was enthusiastic about the goals and general research strategy, and we are grateful to the editor and the reviewers for the constructive and insightful criticisms and suggestions. We believe that these criticisms strengthened our manuscript considerably.
>
> In response to these criticisms, we substantially revised the manuscript and added further analyses in the process. We hope that you agree that our manuscript is now ready for publication in *JEP:G*.

*Your claim is that continuous speech provides clues to segmentation but not recall, whereas segmented speech provides clues to recall, but not segmentation. There might be something to this, but I cannot tell if the dissociation is real or manufactured because you used different materials for segmentation versus recall. In the materials you used for segmentation, in Experiment 1, each bisyllabic fragment that participants are going to try to identify has been heard in multiple contexts (e.g., faIzO:gu, faIzO:vOI, and faIzO:nA). In continuous speech, the bisyllable would perceptually pop out as a result. In segmented speech, as a reviewer mentions, each segmented item might just be learned as a trisyllabic word, overshadowing the effects of repetitions among trisyllables.*

> We believe that we are in full agreement with the editor (and the Reviewer), and some of us have argued that the insertion of silences among words create Gestalt-like groupings (Endress et al., 2009, *TiCS*; Endress & Mehler, 2009, *QJEP*; Endress & Bonatti, 2016, *Wiley Interdis Revi Cognit Sci*).
>
> We agree that one would expect the bisyllabic items to pop out in the continuous version of the word segmentation experiment (i.e., the former Experiment 1, now Experiment 2). If so, the trisyllabic items should also pop out in the recall

experiment (i.e., the former Experiment 2, now Experiment 1), and participants should be able to recall them, which is clearly not the case. As a result, Statistical Learning does not lead repeating elements to pop out.

Conversely, we agree that the insertion of silences creates chunks within which Statistical Learning is more difficult because the chunks are memorized as entire chunks. However, this also implies that Statistical Learning is separate from the (declarative) mechanisms involved in memorizing chunks. This is now mentioned on pp. 42-43:

"*A possible alternative interpretation is that, in the continuous streams of Experiment 2b, repeated bisyllabic items pop out (and are thus remembered), while, in Experiment 2a, chunking cues (in the form of silences) prevent sub-chunks from popping out. However, if repeated bisyllabic items pop out in Experiment 2b, repeated trisyllabic items (i.e., words) should pop out in Experiment 1 as well, and participants should be able to recall them as a result. As this prediction is falsified, a reasonable conclusion is that Statistical Learning does not make repeating elements pop out. Conversely, the availability of chunks might make Statistical Learning of within-chunk regularities more difficult, especially if chunks are memorized as whole units. This possibility would also confirm that Statistical Learning is separable from the (declarative) mechanisms involved in memorizing chunks.*

*Further, while our trisyllabic items are relatively short, so are utterances in infant-directed speech. For example, infant-directed utterances have a typical duration of about 1 s (with some cross-language variability; see e.g., Fernald et al., 1989; Grieser & Kuhl, 1988), with a mean utterance length of about 4 (e.g., Snow, 1977; Smolak & Weinraub, 1983; see also A. Martin, Igarashi, Jincho, & Mazuka, 2016). As a result, if Statistical Learning is difficult in shorter utterances, the utility of Statistical Learning for language acquisition might be reduced.*"

*In contrast, in Experiment 2, to examine recall, there was no such variation. For example, for one language the entire vocabulary was pAbiku, tibudO, dArOpi, gOLAtu. Without variation of the context in which bisyllables occur, a continuous presentation doesn't reveal where the bisyllabic divisions are that would be used in the test. Without such variation, the only recourse is to learn the trisyllables directly. It seems quite possible that if you used the materials from Experiment 1 to examine recall, you would find more frequent recall of trisyllables with segmented speech and bisyllables with continuous speech.*

Here, we are forced to disagree with the editor. In the former Experiment 2 (now Experiment 1), the trisyllabic words occur in variable contexts, namely that of other words. As a result, by the editor's account, the trisyllabic units should pop out just as the bisyllabic units in the former Experiment 1 (now Experiment 2). This is now discussed in Footnote 4:

"*A further possible alternative interpretation of the difference between Experiments 1 and 2 is that the bisyllabic elements in Experiment 2 occurred in different contexts of other syllables. However, the words in Experiment 1 also occurred in different contexts, namely*

*that of other words. As a result, if the availability of variable contexts were sufficient for the formation of declarative memories from continuous speech, such memories should be obtained in both Experiment 1 and 2."*

*If I missed something important please feel free to let me know; the methods section was a bit sparse, and this is my understanding of what you did and how I can interpret it.*

We agree that the method section was incomplete. We now integrate a complete method section for all experiments.

*Considering my perception of the experiments and the changes that are recommended by the reviewers, I think the necessary changes are too extensive for me to consider a revision of this submission. I think, though, that if you continue work along these lines and address the points that the reviewers and I raised, you could possibly have a viable new submission, either to our journal (though the paper would go to a new editorial team, as my term has ended) or to a more specialized journal.*

We are grateful for this encouragement as well as the helpful and insightful criticisms and suggestions, and thus decided to submit a (substantial) revision of our manuscript.

*As you probably know, we can accept only small fraction of the papers that are submitted each year. Accordingly, we must make decisions based not only on the scientific merit of the work but also with an eye to the potential level of impact for the findings for our broad and diverse readership. If you decide to pursue publication in another journal at some point (which I hope you will consider), I hope that the suggestions and comments offered in these reviews will be helpful.*

*Thank you for submitting your work to the Journal. I wish you the best in your continued research, and please try us again in the future if you think you have a manuscript that is a good fit for Journal of Experimental Psychology: General.*

*Sincerely,*

*Nelson Cowan*
*Editor*
*Journal of Experimental Psychology: General*

# Responses to Reviewer 1

Revised manuscript for *JEP:G*
"The specificity of sequential Statistical Learning: Statistical Learning accumulates predictive information from unstructured input but is dissociable from (declarative) memory"

*This manuscript describes results from two studies that examined adults' segmentation of syllable sequences from statistical information. Participants heard sequences that were either continuous or had brief pauses every three syllables. The question was whether participants would use statistical information when trying to recognize or recall statistically-defined units in the continuous inputs. In Experiment 1, participants segmented a continuous stream from statistical information better than chance when tested for recognition of units, in line with past studies. In Experiment 2, participants were tested for recall of units in the syllable sequences, and items recalled did not exceed chance levels of performance vs. actual words from the sequence (if I am understanding correctly).*

*I think the paper tackles an important issue and Experiment 2 used an interesting design. The finding that participants did not use statistical information for word learning when tested for recall is important and bears implications for theories of learning, as stated in the manuscript.*

*In my opinion, however, the paper needs an extensive revision. It is unnecessarily confusing at present because of its organization, necessitating a lot of back and forth through the ms and the supplement, and general lack of clarity.*

> We are very pleased to read that the Reviewer thought the paper tackles an important issue and that the design of Experiment 2 was interesting. We are grateful for the insightful and constructive criticisms and suggestions, which, we believe, strengthened our manuscript considerably.
>
> In response to the comments of this and the other reviewers, we completely rewrote and reorganized the manuscript. The introduction is now focused on the relationship between statistical learning and declarative memory. We changed the order of the experiments to create a more logical flow of the argument and added self-contained method sections.

*Here are some comments and suggestions:*

*1. In the Introduction, it should be stressed that a principal innovation of the studies is the testing of recall of learned materials from statistical information. At present this message is muddled—Experiment 2 is described as testing the "function" (p. 8) or "computational function" (p. 13) of statistical learning but this is opaque; how does testing for recall tell us about "function?" What does that mean?*

> As mentioned above, we now completely rewrote the introduction, focusing on the relationship between declarative memory and statistical learning. We also reversed the experiments to create a more logical flow of the arguments.

*2. It's not clear how many of the reported results are important and they are not forecast very consistently (e.g., discussion of forward vs. backward TPs; the many comparisons of conditions). The ms should be streamlined to stress the principal results with a limited number of figures showing the main findings, and it should be reorganized tor maximum clarity. Specific problems:*

> While we streamlined the results section, we still need to report a substantial number of results. For example, to show that participants do not remember words, we need to show that they did not simply fail to track transitional probabilities, that they produce a sufficient number of items and so on.

*(a) There are descriptions of methods in the Results section.*

> We now include self-contained Methods sections and removed the methods from the Results sections.

*(b) Figures 2 and 3 are a mixture of central and subsidiary findings, and they took me quite a while to disentagle.*

> We now split these overcomplicated figures into several more focused ones.

*(c) It's not clear to me that the reader's time is well spent by going through all the comparisons in Table 2.*

> While we removed some information from the Table, we decided to keep it, as we believe that including the numerical information from the Table into the main text would disrupt the text flow.

*(d) The section on Experiment 2 results (pp. 13-18) is particularly difficult to follow. (e) Analysis in Experiment 2 is limited to a subset of participants (p. 14, third paragraph); this took me by surprise and needs justification.*

> We now include self-contained method sections. We excluded participants if their recognition performance did not exceed 50% to make sure that a failure to recall items was not due to a failure to learn statistical information to begin with. We also excluded participants with no analyzable responses (e.g., participants who produced only unattested items or only single items). This is now clarified in the method section on p. 12.

*3. Principal results belong in the paper and subsidiary results & methods (e.g. Tables S4 & S5) are better placed in the supplement. Currently the supplement contains material that belongs in the main paper, such as information about participants, apparatus, and design. It should not be required that the reader go back and forth between the ms and supplement to figure out and evaluate the findings.*

As mentioned above, we now include self-contained method sections.

*4. The paper should be written with a general Introduction and Discussion and include three experiments. Supplementary materials should be minimal. The two recall experiments (in-lab and on-line) are currently combined in Experiment 2 but they should be described separately (viz., Experiments 2 and 3 in a revision). Each of the three experiments should have its own Methods and Results sections. Organizing the paper in this more traditional fashion might help the reader keep track.*

We followed this advice, and now present the paper in a more traditional format. The two recall experiments are now Experiments 1a and 1b; we decided to present them jointly, given that the results are essentially identical. The former Experiment 1 is now combined with some results that were originally reported in the Supplementary Online material, and split into Experiment 2a (with stimuli synthesized using a British-English voice) and Experiment 2b (with stimuli synthesized using an American-English voice).

*Minor comments:*

*The first complete sentence on p. 7 seems to contradict the first complete sentence in paragraph 3 on p. 5.*

We now make clear that the view that the speech signal is continuous is not universal, and warn readers that we will present an opposing view. This is now clarified on p. 5:

"*Speech is often thought to be a continuous signal (and often perceived as such in unknown languages, but see below), and before learners can commit any words to memory, they need to learn where words start and where they end.*"

*p. 8, Participants: The numbers of participants for each condition in Experiments 1 and 2 do not consistently match the numbers provided in the Supplementary Materials.*

We now present our final sample in Table 1.

*p. 14, second paragraph, line 4: "participant" —> "participants"*

Fixed!

*Scott P. Johnson, UCLA (I sign all reviews)*

# Responses to Reviewer 2

Revised manuscript for *JEP:G*
"The specificity of sequential Statistical Learning: Statistical Learning accumulates predictive information from unstructured input but is dissociable from (declarative) memory"

*The authors ask whether tracking transitional probabilities between syllables (i.e., statistical learning) can account for the acquisition of declarative knowledge of word forms. In Experiment 1, they show that listeners distinguish between statistical words and partwords only when exposed to a continuous stream of speech (i.e., they do not track statistical structure of a speech stream when the speech stream is pre-segmented). In Experiment 2, they show that listeners only show declarative memory for statistical words when exposed to pre-segmented streams of words (i.e., they cannot recall words after exposure to continuous speech stream). The authors conclude that there is a disconnect between what listeners learn from continuous and pre-segmented streams of speech and argue that statistical learning is not used to form declarative memories for lexical items.*

*I really enjoyed reading this manuscript. For decades, it has been an uphill battle to publish any findings that question the role of transitional probability tracking for early language development. Indeed, publishing work in this area is so frustrating that many researchers have just abandoned attempts to pursue this type of work. For this reason, although I am still slightly skeptical about some aspects of the authors' argument, I really applaud them for this daring work. The authors' unique stance on statistical learning for word learning is a breath of fresh air for the field! The authors have identified a really important issue that will generate new questions for language researchers and have serious ramifications for currently dominant models of speech processing and word learning. I also enjoyed how the authors embedded this work in a broad literature, discussing learning and evolution in other species in relation to language acquisition in humans. I would love to see this work published in some format. But sections of this manuscript are pretty rough and difficult to follow. And some design and procedure decisions need to be better explained. I list my main concerns below.*

> We are very pleased that this reviewer found that our unique stance on statistical learning for word learning is a breath of fresh air for the field, and we are grateful for the constructive and insightful criticisms and suggestions, which, we believe, strengthened our manuscript considerably.
>
> We agree that the previous version of the manuscript was less clear than it could have been, and completely rewrote it as a result.

*- In the second line of the abstract, do you mean 'might' or 'has been argued to'? And do you mean 'learning' or 'memorizing'? Use of the word memorizing makes it sound a bit too intentional to me. Relatedly, for a general audience journal, the authors need to be careful in their use of some terms that may have different slightly meanings in different areas of psychology - like the terms statistical learning and declarative memory.*

> We now edited the abstract in line with these suggestions. Specifically, we now write:

"*Learning statistical regularities from the environment is ubiquitous across domains and species. It has been argued to support the earliest stages of language acquisition, including identifying and learning words from fluent speech (word-segmentation). We ask how the Statistical Learning mechanisms involved in word-segmentation interact with the memory mechanisms needed to remember words, if they are tuned to specific learning situations. We show that, when completing a memory recall task after exposure to continuous, statistically structured speech sequences, participants track the statistical structure of the speech stream, but hardly remember any items at all and initiate their productions with random syllables (rather than word-onsets) despite being sensitive to probable syllable transitions. Only discrete familiarization sequences with isolated words produce memories of actual items. Conversely, Statistical Learning predominantly operates in continuous speech sequences like those used in earlier experiments, but not in discrete chunk sequences likely encountered during language acquisition. Statistical Learning might thus be specialized to accumulate distributional information, but dissociable from the (declarative) memory mechanisms needed to acquire words.*"

*- Although the Introduction is well-written and interesting, I found the methodology and results sections incredibly confusing. I had to read this section of the manuscript many times to understand it. I think part of the problem was that the layout did not pattern as I expected. I think the manuscript would be much easier to read if the authors followed a more traditional format where the three experiments are reported in succession one after another rather than in an interleaved fashion (each with their own participant and stimulus section, etc.).*

We now present the paper in a more traditional format, including dedicated methods sections.

*The authors also need to give better motivation for all design and procedure decisions. For example, why were the silences placed after every word instead of after every few words? Given how the authors describe the task of the listener (e.g., deciding whether hi baby should be parsed into hiba by or hi baby), wouldn't it make more sense to do the latter?*

We now discuss this issue on p. 15:

"*As the role of the silences in the pre-segmented stream was to create clearly identifiable chunks, the silence duration was chosen to result in clearly perceptible syllable groups (according to the experimenters' perception). Other investigations with pre-segmented material used shorter silences (e.g., Peña et al., 2002), longer ones (e.g., Sohail & Johnson, 2016; Endress & Mehler, 2009a) or natural prosodic phrasing (Shukla et al., 2007; Seidl & Johnson, 2008). Relatedly, other experiments mimicking the prosodic organization of speech used natural prosodic phrasing (Shukla et al., 2007; Seidl & Johnson, 2008) or grouped several "words" together using silences (Sohail & Johnson, 2016). In the light of Experiment 2, where we ask if Statistical Learning can be used to break up small prosodic groups such*

*as "thebaby" into their underlying words (i.e., "the+baby"), we follow Peña et al. (2002) and present silences after each word instead of inducing longer groupings.*"

*And choosing a sample size to be convenient for a third-year psychology student project doesn't seem like good justification for a sample size? Moreover, I'm not sure if 'third year psychology student project' refers to a project run on third year students, or a project run by a third-year student? Are you sure you didn't choose the project that had a needed sample size that was appropriate for an undergraduate project/class?*

We now make clear that we refer to a project *run* by a third-year student. While this choice of the sample size might result in underpowered experiments, we replicated our critical findings in independent samples. For example, we now write on p. 12 and p. 33:

"*This number was chosen because it is realistic in the time-frame available for a third-year honors project.*"

*- I was a bit confused regarding the authors' explanation for the seemingly bimodal patterning of participants' responses in the continuous condition in Figure 3a.*

We now write on p. 27

*"However, inspection of Figure 3a shows that the distribution in the continuous condition is bimodal, with some participants producing only words, and others producing only part-words. Such a behavior can arise if participants pick a syllable as their starting-point, and segment the rest of the stream accordingly. If they happen to pick a word-initial syllable a, they will produce only words; if they pick the second or the third syllable of a word, all subsequent items will be part-words.*"

*- I had a hard time working out what the words in the Experiment 1 language were, based on the information provided.*

We now added a traditional methods section; the words are given in Table 5.

*- The authors report that participants reported to be native English speakers. In my experience, most participants don't know what it means to be a native English speaker (and indeed, many language researchers don't even agree on what this means). Did you give your participants a definition for what it means to be a native English speaker?*

We did not provide participants with a definition of native proficiency, and did not evaluate their proficiency. However, at least in the lab-based experiments, it can be assumed that the overwhelming majority were exposed to English from childhood, given that the experiment was run in London, UK and we did not notice any clear

non-native accents when welcoming the participants to the lab. This is now mentioned on p. 12 (Experiment 1) and p. 34 (Experiment 2):

"*Participants reported to be native speakers of English, but we did not assess their English proficiency. However, participants were most likely exposed to English from childhood, as the experiment took place in London, UK, and the experimenters did not notice any clear non-native accents.*"

*- Past studies have shown that 3-1-2 versus 2-3-1 words are learned differentially (i.e., participants often mistake partwords containing the second and third syllable of a word as a whole word). Do the authors think their results were affected by what type of partwords they were tested on? This might be a particularly interesting factor to consider in the production data?*

We assume that this comment refers to the former Experiment 1 (now Experiment 2), where half of the participants had to split a triplet like ABC into a bisyllabic item and an extra syllable (i.e., AB+C), while the other half had to split the items into an extra syllable followed by a bisyllabic item (A+BC). Given that, when participants are presented with trisyllabic items of consistent length, they recognize 2-3-1 part-words somewhat better than 3-1-2 part-words, one might wonder whether a similar difference emerges in the former Experiment 1. However, as shown in Table 6, we generally did not detect such a difference.

This might be because the difference between 3-1-2 and 2-3-1 items might emerge from basic Hebbian learning (Endress & Johnson, 2021, *Cognition*), presumably because the strong association in the 2-3 transition leads to greater overall activity in the associative network when it is first exposed to the test item compared to a weaker 3-1. As a result, such effects might result from the traditional design of statistical learning tasks rather than being a core property of statistical learning.

*- How did the authors choose a silence of 222 or 520 ms? Why did it differ across studies? And didn't the stimuli sound odd without a ramping in and out of intensity and F0?*

As mentioned above, we now made clearer that the silences resulted in the perception of clearly defined chunks. The ramping in and out of individual chunks was taken care of by the speech synthesizer.

Unfortunately, we could not recover the reasoning for the different silence duration across the experiments. While the syllable durations in the former Experiment 2 (now Experiment 1) were based on Saffran et al.'s (1996) experiments, the segment durations in the former Experiment 1 (now Experiment 2) were based on measurements. Given that the role of the silences is to induce clearly defined chunks, we would assume that the actual silence duration will not affect the results as long as the silences are clearly perceptible, but we could not recover the reason for choosing different silence durations.

*- For the most part, the authors do an excellent job reviewing the relevant literature. However, they really should cite and discuss Sohail and Johnson, another paper where listeners' ability to use statistical cues to locate word boundaries was compared when prosodic bracketing was present or absent in an artificial speech stream.*

*Sohail, J., & Johnson, E.K. (2016). How transitional probabilities and the edge effect contribute to listeners' phonological bootstrapping success. Language Learning and Development, 12(2), 105-115.*

This paper is now discussed on pp. 8, 15, 32, and notably on pp. 43-44 :

"*This is not to say that Statistical Learning can never occur in pre-segmented units. While the available statistical information does not always improve performance when chunking information is available (e.g., Sohail & Johnson, 2016), Shukla et al. (2007) showed that, when adults learners are exposed to 10 syllables chunks (defined by intonational contours), they have some sensitivity to statistical information within the chunks, though they might also use declarative memory mechanisms to remember sub-chunks (see also Endress & Bonatti, 2007; Endress & Mehler, 2009a for additional results suggesting that Statistical Learning is possible within chunks). However, Shukla et al. (2007) also found that participants predominantly retain information at chunk edges rather than at chunk medial positions. At minimum, it is thus an empirical question to what extent Statistical Learning is useful for word segmentation in the short utterances infants are faced with.*"

# Responses to Reviewer 3

Revised manuscript for *JEP:G*
"The specificity of sequential Statistical Learning: Statistical Learning accumulates predictive information from unstructured input but is dissociable from (declarative) memory"

*Through two experiments, the authors aim to show that statistical learning does not play a role in word segmentation insofar as it is claimed not to encode word candidates in declarative memory. In Experiment 1, participants were exposed to repeated trisyllabic sequences, consisting of a monosyllabic and a bisyllabic nonsense word--the latter defined by having transitional probability (TP) of 1 between the two syllables whereas TPs between words were 1/3. During familiarization, half of the participants were presented with syllables in which the transitional probability (TP) and chunk frequency was higher between the first and second syllable as compared to the second and third syllable (AB + C), and the other half were presented with syllables in which the TP and chunk frequency was higher between the second and the third syllable as compared to the first and the second (A + BC). The familiarization stream was either a continuous stream of syllables or it was segmented into the trisyllabic sequences, comprising the monosyllabic and bisyllabic words. Afterwards, the participants were given a two-alternative-forced choice task (2AFC), where they were asked to pick the (bisyllabic) word they had heard during familiarization (AB vs. BC). The results indicated that only participants exposed to the continuous stream tended to choose the bisyllabic words they had heard over the foil. In Experiment 2, participants were exposed to four trisyllabic nonsense words (TP = 1 within words and 1/3 between words), either concatenated into a continuous stream or presented one after another with pauses between them. After familiarization, participants were then asked to free recall the words they could remember from the speech stream. Not surprisingly, given the pauses between words, the participants that were thus exposed to pre-segmented individual word forms performed considerably better than participants exposed to the continuous stream. The latter group seems to have had problems picking up on even word beginnings/endings. The participants were also administered a very short recognition task, using a 2AFC task with 4 stimulus pairs. From the pattern of results the authors conclude that statistical learning does not play a role in speech segmentation because participants are not able to explicitly remember words from the continuous speech stream in Experiment 2 and there was no sensitivity to bigram regularities in short pre-segmented trisyllabic sequences in Experiment 1.*

*This is a generally well written paper, though perhaps too many methodological details have been placed in the SOM (making it all but impossible for a reader to evaluate the experiments from the article alone). The authors are right that it seems unlikely that statistical learning to be the only source of information useful for speech segmentation, yet it seems premature to conclude that sensitivity to statistical regularities is not important or even needed for learning words. As detailed further below, the two experiments do not provide compelling evidence for this conclusion either, though they might provide the basis for more definitive research. The study also has several methodological shortcomings that limit the theoretical implications, making it better suited for a more specialized journal.*

We were very pleased to read that the reviewer thought that the paper was well written, and we are grateful for the insightful and constructive criticisms, which, we believe, strengthened our paper considerably.

We substantially rewrote the manuscript in response to these and other comments, including a traditional method section.

*EXPERIMENT 1*

*The ideas behind this experiment are interesting in principle but the implementation leaves something to be desired. In particular, given the very short nature of the pre-segmented trisyllabic sequences involving just 3 bisyllabic words and 3 monosyllabic words, it seems likely that participants in this condition simply treated them as a single trisyllabic word. Indeed, given work in the perception literature related to Gestalt perceptual-grouping principles (e.g., Bregman, 1990), it seems likely that because of their short duration -- 360 ms compared to 540 ms for the pauses -- the relative shortness of the trisyllabic sequences interspersed with considerably longer pauses are likely to have induced such Gestalt-like groupings. Indeed, prior work on this has demonstrated different kinds of Gestalt effects in both auditory (Morgan et al., 2019) and visual (e.g., Glicksohn & Cohen, 2011) statistical learning, making this Gestalt account a likely explanation for the results of Experiment 1. Such groupings would make it harder for participants to notice any internal regularities. And given the explicit reasoning-based nature of the 2AFC task used in Experiment 1, this makes the results less compelling and definitive, given that they can be explained by other factors.*

We believe that we are in full agreement with the reviewer about the mechanistic interpretation of the results of the former Experiment 1 (now Experiment 2), though not necessarily about the implications of the results. In fact, some of us have argued that the insertion of silences among words creates Gestalt-like groupings that make strings more compatible with the memory representations used for speech items (Endress et al., 2009, *TiCS*; Endress & Mehler, 2009, *QJEP*; Endress & Bonatti, 2016, *Wiley Interdis Revi Cognit Sci*). This is now discussed on p. 9:

"*There is some evidence that learners might process continuous speech sequences differently from discrete ones (e.g., Endress & Bonatti, 2016; Marchetto & Bonatti, 2015; Peña, Bonatti, Nespor, & Mehler, 2002). For example, Peña et al. (2002) familiarized participants with continuous speech streams as well as with discrete, "pre-segmented" speech streams, where each word was followed by a brief silence. The brief silences triggered additional processes such as rule-like generalizations that were not available after continuous familiarizations. Critically, the rule-like generalizations observed after pre-segmented familiarizations might reflect memory processes. Endress and Mehler (2009a) suggested that the role of the silences was to act as Gestalt-like grouping cues that provided learners with the location of the word edges (i.e., onsets and offsets), and thus enabled generalizations based on those word-edges (see also Glicksohn & Cohen, 2011; Morgan, Fogel, Nair, & Patel, 2019 for other perceptual grouping effects in Statistical Learning). Given that the grouping cues resulted in a sequence of discrete chunks, the grouping cues might also support declarative memory processing.*"

We would also agree with the reviewer that the fact of remembering entire units makes participants less sensitive to the internal statistical structure of the items. However, we (presumably) disagree about the implications of this finding. While our trisyllabic items are relatively short, so are utterances in infant-directed speech. As a result, if learners find it difficult to use Statistical Learning in relatively short

utterances, the utility of Statistical Learning for language acquisition would be substantially reduced. This is now discussed on pp. 42-43:

"*A possible alternative interpretation is that, in the continuous streams of Experiment 2b, repeated bisyllabic items pop out (and are thus remembered), while, in Experiment 2a, chunking cues (in the form of silences) prevent sub-chunks from popping out. However, if repeated bisyllabic items pop out in Experiment 2b, repeated trisyllabic items (i.e., words) should pop out in Experiment 1 as well, and participants should be able to recall them as a result. As this prediction is falsified, a reasonable conclusion is that Statistical Learning does not make repeating elements pop out. Conversely, the availability of chunks might make Statistical Learning of within-chunk regularities more difficult, especially if chunks are memorized as whole units. This possibility would also confirm that Statistical Learning is separable from the (declarative) mechanisms involved in memorizing chunks.*

*Further, while our trisyllabic items are relatively short, so are utterances in infant-directed speech. For example, infant-directed utterances have a typical duration of about 1 s (with some cross-language variability; see e.g., Fernald et al., 1989; Grieser & Kuhl, 1988), with a mean utterance length of about 4 (e.g., Snow, 1977; Smolak & Weinraub, 1983; see also A. Martin, Igarashi, Jincho, & Mazuka, 2016). As a result, if Statistical Learning is difficult in shorter utterances, the utility of Statistical Learning for language acquisition might be reduced.*"

*A more informative experiment might involve pre-segmented combinations of 2 trisyllabic nonsense words (e.g., taken from a six word version of the language from Experiment 2), such that stimuli might include a pre-segmented input along the lines of: ABCDEF GHIJKL MNOPQR GHIABC DEFPQR ... and performance then compared with a continuous version. Although Gestalt grouping might also be possible across 6 syllable strings, the fact that more combinations would be possible, the longer strings relative to the pauses would make for a more compelling experiment. Thus, a pre-registered version of such an experiment would likely be more definitive.*

We believe that there is also some evidence addressing this question. Shukla et al. (2007, *Cognit Psychol*) showed that, when adult learners are exposed to 10 syllable chunks (defined by intonational phrases), they either have some sensitivity to statistical information within the chunks or remember sub-chunks using declarative memory mechanisms. However, Shukla et al. (2007) also showed that participants predominantly retain information at chunk edges rather than at chunk medial positions. As a result, within chunk learning is possible to some extent, but we would argue that the difficulty to learn from structured input is problematic for a role of Statistical Learning in word segmentation. This is now discussed on pp. 43-44:

"*This is not to say that Statistical Learning can never occur in pre-segmented units. While the availability statistical information does not always improve performance when chunking information is available (e.g., Sohail & Johnson, 2016), Shukla et al. (2007) showed that,*

*when adults learners are exposed to 10 syllables chunks (defined by intonational contours), they have some sensitivity to statistical information within the chunks, though they might also use declarative memory mechanisms to remember sub-chunks. However, Shukla et al. (2007) also found that participants predominantly retain information at chunk edges rather than at chunk medial positions. At minimum, it is thus an empirical question to what extent Statistical Learning is useful for word segmentation in the short utterances infants are faced with.*"

*Additionally, why was only the continuous condition replicated? This seems like a post hoc addition to the experiment.*

We replicate only the continuous condition due to the unexpected results with the *en1* voice when the speech stream was continuous. In contrast, the pre-segmented condition was successfully replicated with the *en1* voice; we also report several conceptual replications with a different population in the pilot experiment (SOM5). In other words, while we obtained several independent replications of the failure to use statistical information with pre-segmented streams, we wanted to be sure that, with a continuous familiarization, participants can really track statistical information in our design. This is now mentioned on pp. 7-8 and p. 8:

pp. 7-8

"*As shown in Table S3, performance was much better for Language 1 than for Language 2, presumably due to some click-like sounds the synthesizer produced for some stops and fricatives (notably /f/ and /g/). These sounds might have prevented participants from using statistical learning. We thus decided to replicate the results with a different, American English voice.*"

p.8

"*Given the unexpected results with the en1 voice above, we replicated the successful tracking of statistical information using a new sample of participants.*"

*The pre-segmented condition appears to have one extreme outlier who got 5-6% correct (which would be significantly below chance, perhaps indicating that they had learned something but was applying the "wrong rule"). It seems worthwhile to exclude this outlier. Given this issue, it might seem pertinent also to replicate this result given that more than half of the participant in this condition appear to have scored above 50% (though not all necessarily significantly so).*

In SOM4, we now report an analysis where we exclude all outliers who deviate by more than 2.5 SD from each condition mean. The results are essentially unchanged.

*EXPERIMENT 2*

*The production results from this experiment seem trivial: Participants in the pre-segmented condition are given the word initial and word final edges of words. Although the pauses (222 ms) in this experiment were shorter relative to the trisyllabic words (648 ms), they're still likely to be short enough to potentially induce a Gestalt grouping across the three syllables in the 4 nonsense words. Thus, from this perspective it is not surprising that participants in the pre-segmented condition were able to recall most of the nonwords. Importantly, this kind of free recall task is known to be a test of explicit knowledge (see e.g., Meulemans & Van der Linden, 1997), whereas statistical learning is often considered to be related to implicit learning and therefore the low performance in the continuous condition is unsurprising. It is therefore not the case, as the authors suggest on page 18, that "TPs do not allow learners to reliably detect onsets and offsets of words." There were no implicit measures assessing whether participants' processing might have been reliably affected by statistical learning. Thus, a fundamental problem with this experiment is that it assumes that sensitivity to statistical regularities translates into explicit knowledge thereof (being able to recall words) as supposed to affecting their processing implicitly (which is not assessed).*

Here, we are forced to disagree with the reviewer. We agree that a free recall task taps into explicit processes, while Statistical Learning might be implicit (though it is certainly sensitive to attentional manipulations; see e.g., Toro et al., 2005, *Cognition*; Turk-Browne et al, 2005, *JEP:G*). However, learners eventually need to produce words, just as in our free recall tasks. If Statistical Learning does not lead to the kinds of representations that allow learners to produce items even when they are demonstrably sensitive to statistical information (as they are in our experiments), it is unclear how Statistical Learning might support word learning. This is now discussed on p. 32 and in our reply to a comment below:

"*Experiment 1 suggests that participants do not form declarative memory traces of words when the only available cues are statistical in nature. In contrast, they readily form declarative memories when items are pre-segmented. These results do not imply that Statistical Learning might not play a critical role in word segmentation. As mentioned above, speech is prosodically organized (Cutler et al., 1997; Nespor & Vogel, 1986; Shattuck-Hufnagel & Turk, 1996), and a learner's segmentation task is not so much to integrate distributional information over long stretches of continuous speech, but rather to decide whether the correct grouping in prosodic groups such as "thebaby" is "theba + by" or "the + baby". In principle, Statistical Learning might be well suited to this task. In line with the two-step explanation of Graf-Estes et al.'s (2007), Hay et al.'s (2011), Isbilen et al.'s (2020) experiments above, implicit knowledge of statistical regularities might help learners acquire words more effectively once (prosodic) segmentation cues are given (but see e.g. Ngon et al., 2013; Sohail & Johnson, 2016).*"

*A second issue is why only 4 pairs of items were used in the recognition test in Experiment 2, when more were possible. Without a pre-registration this seems a bit odd given that the results of this recognition test are used to exclude participants. With only 4 pairs, passing performance requires 75% correct to be included and the impact of strategic guess may additionally play an outsized role here.*

We used the words from Saffran et al.'s (1996, *Science*) classic Experiment 2. Given that even 8-month-old infants can track the statistical information in these speech streams, we believe that the learning situation is sufficiently simple for adults to demonstrate their knowledge of the underlying words. However, our results reveal that participants have no (explicit) knowledge of these words after dozens of exposures when tested in a free recall task.

While we created a version of the experiments using more words from Saffran et al.'s (1996) adult experiments, we opted for the simpler version as informal pilot tests suggested that participants did not remember the underlying words.

*Apart from the methodological limitations with Experiments 1 and 2, another key issue with the paper is that it does not consider the possible role of explicit vs. implicit measures as they relate to the theoretical issues that it aims to address (though there is some reference towards the end of the paper). The authors seem to assume that word segmentation requires explicit learning (what they refer to as declarative learning), and suggest that statistical learning is akin to procedural learning (which they presumably take to be implicit). But there would seem to be at least some aspects of word learning that involve implicit knowledge (e.g., related to phonotactics, which can determine which possible word forms form legal combinations of sounds). The primary tasks used in this study -- 2AFC in Experiment 1; free recall in Experiment 2 -- are all geared toward explicit knowledge. However, if statistical learning (as at least assumed by some) is more implicit in nature then the results are less meaningful (certainly if one doesn't adopt the theoretical declarative/procedural distinction as it applies to lexicon/grammar).*

We believe that the distinction between explicit and implicit processes is a rather tricky one. The end state of word learning is presumably declarative and explicit, given that competent learners can recognize and produce any word they know. However, once lexical entries have been established, lexical access can be facilitated by implicit cues, such as subliminal priming and steam completion, suggesting that declarative forms of memory are not immune to implicit manipulations. Conversely, Statistical Learning might be implicit (e.g., Meulemans & Van der Linden, 1997; Perruchet & Pacton, 2006; Christiansen, 2018), but it is also sensitive to attentional manipulations (see e.g., Toro et al., 2005; Turk-Browne et al, 2005). Likewise, phonotactic regularities can help segmentation (Friederici & Wessels, 1993; Mattys et al., 1999; McQueen, 1998), and such regularities can be learned from mere exposure (Onishi et al., 2002, Chambers et al., 2003, Chambers et al., 2011). However, they can also be learned by keeping track of information at utterance boundaries (Monaghan & Christiansen, 2010). Given that utterance boundaries are just the kind of information that led to explicit processing in our studies, it is unclear to what extent knowledge of phonotactic constraints is truly implicit, or whether the phonotactic wellformedness of a word depends on the number of similar words in the mental lexicon (i.e. if phonotactic constraints are based on implicit rules or rather on exemplars).

As a result, we are open to the possibility that implicit, statistical processes might contribute to the establishment of lexical representations. However, given that one of the best performing segmentation models relies on information at utterance boundaries rather than statistical information (Monaghan & Christiansen, 2010), that the memory format resulting from Statistical Learning might not match the format of memory representations of linguistic items (e.g., Endress & Langus, 2017; Fischer-Baum et al., 2011; Miozzo et al., 2016), and that Statistical Learning is clearly not sufficient to allow for the production of words, it is an empirical question to what extent Statistical Learning contributes to lexical acquisition.

This is now discussed on p. 45:

"*This is no to say that Statistical Learning might play no implicit role in word learning even when it is not sufficient to produce memories that can be recalled. For example, and as mentioned above, associations among syllables might facilitate the establishment of declarative memories once suitable (and explicit) segmentation cues become available (Endress & Langus, 2017), and, once words are acquired, word processing is not immune to unconscious stimuli such as masked primes (e.g., Forster, 1998; Kouider & Dupoux, 2005). Statistical Learning might also facilitate word learning indirectly, for example through the acquisition of phonotactic constraint that might affect word learning in turn (e.g., Friederici & Wessels, 1993; Mattys, Jusczyk, Luce, & Morgan, 1999; McQueen, 1998). However, the extent to which Statistical Learning supports such computations remains to be established. For example, the phonotactic regularities above can be learned by keeping track of material at utterance boundaries (Monaghan & Christiansen, 2010), and thus just using the type of cues we introduced in the pre-segmented conditions. As a result, we believe that it is an important topic for further research to determine the role Statistical Learning plays in word acquisition.*"

*MINOR ISSUES*

*INTRODUCTION: It might be useful for the reader if specific hypotheses could be outlined for the experiments based on contrasting theories. This will help the reader to better situate the experiments.*

We now completely rewrote the introduction and made the hypotheses clearer. (We don't copy a passage here because the entire introduction has been rewritten.)

*Page 14. Estimates are reported in Table 1, but a generalized linear model was used to assess performance. It is unclear why estimates rather than odd ratios are provided?*

We now converted the estimates to odds ratios in Tables 6 and S6.

*Page 19: "It is also consistent with the view that Statistical Learning may be less important for memorizing utterances" — it's not clear who would suggest that SL is for memorization of utterances rather than breaking those utterances into useful subparts?*

We agree, given that we meant to propose that Statistical Learning might be involved in memory for words (rather than utterances), we are not sure if the reviewer objected to this typo, or rather to a role of Statistical Learning for memory in general. We now address both issues by writing on p. 42:

*"[This result] is also consistent with the view that Statistical Learning may be less important for memorizing words (or at least to break up utterances so that the underlying words can be memorized)."*

*Page 20. Having been involved in the review process for Isbilen et al., it seems unlikely that predictive processing can explain the results from that study. They asked people to recall sequences of six syllables (not isolated words) either from the language or randomized and found that the former was better recalled. They also observed better recall of syllable triples from the language. Importantly, they do not attribute the results to declarative memory as such but to long-term changes to processing (similar to what the authors here might be suggesting but with different implications). The Isbilen et al. measure is implicit, of course, whereas the current two experiments focus on explicit measures which are subject to a variety of factors that may limit performance.*

We agree with the reviewer that the question of whether Statistical Learning promotes predictive processing is orthogonal to the Isbilen et al. results. Our point is that these results do not imply that Statistical Learning is used to memorize chunks. Rather, in line with Endress and Langus (2017), we propose a two-step explanation: During the initial exposure phase, participants learn associations among items. During the memory face, where participants are presented with isolated 6-item chunks, knowledge of associations facilitates memory for these chunks, but (declarative) memory requires these chunks to be presented as chunks in the first case.

As we now reversed the order of the experiments, and focused the introduction on the role of Statistical Learning for (declarative) memory, we decided to leave this passage, as it should be clear in the context of the more focused introduction.

*Aside from the fact that SOM1 is not mentioned in the manuscript, it is stated that there are 30, 30, 31 participants (pre-segmented, continuous, replication). However, when looking at the supplementary materials different numbers are listed (i.e., 30, 32, 30). Why do those differ from one another? Moreover, the aim was 30 participants per experiment (15 per language) but there are far more than 30 participants in Experiment 2, as already 157 were tested for the online version. It is unclear why this is the case. Also, what is the distribution of participants for experiment 2 (i.e., how many participants were assigned to the pre-segmented and continuous condition)?*

*We now report full demographic information for all experiments in Table 1. In the first replication of the continuous condition of the former Experiment 1 (now Experiment 2), we recruited 32 rather than 30 participants by mistake, but the results of the replication with 30 participants is undistinguishable. As mentioned on p. 12, we aimed for a greater sample in the online experiments, given how easy it is to recruit online participants, but the results are essentially identical to the lab-based experiments with the smaller sample size.*

*REFERENCES*

*Bregman, A. S. (1990). Auditory scene analysis: The perceptual organization of sound. Cambridge, MA: MIT Press.*

*Glicksohn, A., & Cohen, A. (2011). The role of Gestalt grouping principles in visual statistical learning. Attention, Perception, & Psychophysics, 73(3), 708-713. https://doi.org/10.3758/s13414-010-0084-4*

*Meulemans, T., & Van der Linden, M. (1997). Associative chunk strength in artificial grammar learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23(4), 1007-1028. https://doi.org/10.1037/0278-7393.23.4.1007*

*Morgan, E., Fogel, A., Nair, A., & Patel, A. D. (2019). Statistical learning and Gestalt-like principles predict melodic expectations. Cognition, 189, 23-34. https://doi.org/10.1016/j.cognition.2018.12.015*