Dear Dr. Slevc,

I am very grateful to you and the reviewers for your constructive and thoughtful comments on this manuscript. I found the criticism to be insightful and I have done my best to address each comment thoroughly. Below you will find a copy of the editor and reviewer comments in green, followed by my responses in black, with newly added text *italicized*, where relevant.

Sincerely,

Catherine Laing

Dear Dr Laing,

Thank you very much for submitting your manuscript "Phonological Networks and Systematicity in Early Lexical Acquisition" for review and consideration for publication in Journal of Experimental Psychology: Learning, Memory, and Cognition. I sincerely appreciate the opportunity to review the manuscript. I have now received three expert reviews of your manuscript (appended below) and have read the paper myself. We all agree that this is an interesting and useful contribution, but the reviewers note a number of concerns with the manuscript in its current form. Many of these involve issues of clarification or some additional issues to consider/discuss, which I think can all be feasibly addressed. I thus invite you to revise your manuscript taking these comments into account.

Thank you for your kind remarks about the paper being interesting and useful; I'm glad for the opportunity to strengthen the manuscript further.

Please note that I have made a couple of general changes to the paper (based on reviewer comments) that I would like to flag up here:

- PAT/PAQ values are now termed INT/EXT, respectively, to more clearly mark the distinction between the two;
- I have slightly changed the model structure in both logistic regression models and GAMMs. Following reviewer comments I added a new variable (age of acquisition from comprehensive vocabulary norms) and amended an existing one (using a more general input frequency measure), and to avoid multicollinearity in the data, this led me to remove two of the fixed effects that were included in the initial model (n tokens and vocabulary size).

If you decide to revise the work, please include a cover letter that details your response to each point raised by the Editor and the reviewers in this letter.

To submit a revision, go to https://www.editorialmanager.com/xlm/ and log in as an Author. You will see a menu item call Submission Needing Revision. You will find your submission record there.

Sincerely,
L. Robert Slevc, PhD
Associate Editor
Journal of Experimental Psychology: Learning, Memory, and Cognition

Reviewers' comments:

Reviewer #1: * Review of XLM-2023-2867

Phonological Networks and Systematicity in Early Lexical Acquisition

* Summary and overall assessment

This manuscript adopted the network science framework to study systematicity in the early productive lexicons of very young children. It represents a much needed extension of the comparison between two key network growth models of preferential attachment and preferential acquisition to naturalistic data of infants' actual productions - that has not been done in previous work that used vocabulary norms, as the author rightly points out. The author finds support for the preferential attachment model, in line with research on early production showing that children rely on the word forms that they can produce to guide future learning (what is rich in the lexicon, gets richer), and found stronger effects of PAT over time in contrast to previous literature on this topic. Overall, I have a really positive view of this work - the method and analytic approach are clearly described and presented, the approach of working with naturalistic data is a positive, and the results are important for research in early language acquisition. It is a neat showcase of how network analysis can provide a powerful, quantitative framework to ask theoretical questions about the way that young children build their lexicons. Below are some questions and comments for the author to consider.

Thank you for this positive feedback – I'm really glad R1 found the work clear and saw value in this approach! As a general first comment, please note that, following another reviewer's request, I have re-labelled PAT/PAQ values as INT/EXT values, respectively.

* Specific comments and questions

p. 6 - I know that the Lure of the Associates model is not considered here, but it has been mentioned a couple of times in the Introduction that it may warrant a brief explanation either in-text or in the footnote. Although it is true that previous studies did not find strong evidence for this model, there is some other evidence for it by Holly Storkel and colleagues. They found that nonwords that had more "lures" to the words in the existing lexicon are better acquired - which is conceptually what is predicted by the Lures model.

Storkel, H. L., Armbrüster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. Journal of Speech, Language, and Hearing Research.

I agree that it makes sense to add a brief definition of this model, which I have added to the footnote on p.5:

"*A third model - Lure of the Associates - predicts that new words will be learned that are similar to the highest number of already-known words in the network. This model has been considered in some studies (Hills et al., 2009; Siew & Vitevitch, 2020) but will not be considered here as there is no conclusive evidence for this model in the development literature, though note that there is evidence for this model in adult word learning (e.g. Stamer & Vitevitch, 2012; Storkel, Armbrüster, & Hogan, 2006).*"

p. 6 - Below are a couple of more recent references that also looked at the preferential attachment/acquisition network growth models in language acquisition - those would be

relevant in the literature review of previous work.

Luef, E. M. (2022). Growth algorithms in the phonological networks of second language learners: A replication of Siew and Vitevitch (2020a). Journal of Experimental Psychology: General, 151(12), e26-e44. https://doi.org/10.1037/xge0001248

Ciaglia, F., Stella, M., & Kennington, C. (2023). Investigating preferential acquisition and attachment in early word learning through cognitive, visual and latent multiplex lexical networks. Physica A: Statistical Mechanics and Its Applications, 612, 128468. https://doi.org/10.1016/j.physa.2023.128468

Thank you for these suggestions; these have been added to the literature review on pp.6-7:

"*Ciaglia, Stella and Kennington (2023) analysed complex multiplex networks (including phonological, semantic, sensorimotor and visual associations) to find evidence for both EXT and INT in word learning, though evidence was stronger for EXT.*."

"*A replication of this study using data from adult second-language learners of English found consistent results (Luef, 2022).*"

p. 10 - How many words were excluded on the basis of not being in the CDIs?

I have now added this data to the Methods section on p.12:

"*Altogether, 5483 words were excluded from the data due to not appearing on the French or American English CDIs (2224 in French and 3259 in English).*"

Note that, in addressing this point, I found that ~200 tokens that were CDI words had not been correctly coded in the data, and thus were excluded as being non-CDI words. I have now re-coded these and so the new data sample is slightly larger than that of the initial submission (3096 word types instead of 3013 word types).

p. 11 - Could a concrete example of how the phonological feature approach was used to compute distance? It might also help to explain how this computation was done /without/ the vowels ("in the present analysis only consonants were included"), which was something I did not quite understand. For instance, does this mean that it is not possible to compute the distance between two words that differ by a vowel - like "cat" and "cot"? Some additional explanation for this section would be very helpful.

First to clarify, in this analysis, words that differ only by a vowel, as in "cat" and "cot" would have a distance of 0. To my mind, this is not an issue for infant production, since if the infant was producing both words as, say, /kɛt/, accuracy would be high in both cases (in such cases it would be interesting to test whether different vowels were used to discriminate minimally different words, but that is not within the scope of the current study!).

To clarify my approach, I have added two tables to the SI (Tables 1 & 2) that shows the distances between 3 target words in the dataset, each with a different phonological structure (*baby* compared with *balloon* and *sky*). The tables presented here are adapted from Monaghan et al. (2010), as I found their approach very clear and indeed used it to generate my own phonological distance measures in this paper.

I have also clarified this approach in the main text on p.11:

"*This means that two words that differ only in their vowel segments are coded as the same in the current analysis. Words were aligned by syllable nucleus: onset consonants were*

*compared with other onset consonants, and codas were compared with codas. Full criteria for establishing distance, alongside tabulated examples, are included in the Supplementary Information, S1."*

And added the following explanation in the Supplementary Materials, along with examples in two tables, to illustrate:

"*Phonological distance was established following Monaghan, Christiansen, Farmer and Fitneva's (2010) approach, with some adaptations. Note that in their study, only monosyllabic words are included, and so their approach is adapted here to include multisyllabic words. Following their method, each word was first divided into a series of 'slots', according to its phonological structure. For example, the word baby was separated into five slots: /b-e-i-b-i/. Because vowels were not accounted for, the nucleus of each syllable - both monophthongs and diphthongs - was then replaced by a generic V slot, i.e. /b-V-b-V/. Words were then aligned by nucleus to generate a phonological distance measure between each possible word pair. All consonants at word onset and final syllables were aligned, regardless of syllable number, such that the final /d/ of bed would be aligned with the final /n/ of balloon. This is because infants may have a tendency to produce only certain consonants word-finally, and so this approach would capture such systematicity. For the English data, the maximal word structure considered in the analysis is C-C-C-V-C-C-|C-C-C-V-|C-C-C-V-|C-C-C, where syllable boundaries are marked with a |. This accounts for complex clusters at word onset (e.g. splash /splæʃ/, C-C-C-V-C), coda (plant /plænt/, C-C-V-C-C), and across syllable boundaries (pumpkin /pʌmpkɪn/, C-V-C-C-|C-V-C). In French the maximal structure was C-C-C-V-C-C-|C-C-C-V-|C-C-C-V-C|C-C-C-V-C|C-C-C-V-|C-C-C-C. This accounted for multisyllabic target words such as hélicoptère ("helicopter" /elɪkɔptɛʁ/, V-|C-V-|C-V-C-|C-V-C) and appareil photo ("camera" /apaʁɛjfɔto/, V-|C-V-|C-V-C-|C-V-|C-V), and complex codas as in arbre ("tree" /aʁbʁ/, V-C-|C-C). For vowel-initial words, the C1 slot in word-initial position is empty, but all other alignments remain the same. This maximal structure is required in an analysis of infant word production, to account for unexpected complexities such as, production of French mettre "to put" as [mɛʁstʁ] and étoile "star" as [ɛstwal]. In the infant data, it was not always easy to determine exactly where a syllable boundary should occur in complex productions, in part because this was not predictable based on the target form due to the variability in production, so would have to be done on a word-by-word basis, and in part because the syllable boundary of some productions could not be clearly established from its phonetic transcription. Instead, consonants were always assigned to the syllable-initial cluster, rather than assigning part of the cluster to the coda of the previous syllable (e.g. for the examples above, the infant production of mettre was coded as C-V-|C-C-C-C and étoile as V-|C-C-C-V-C).*"

p. 12 - Because I don't know how phonological distance was calculated specifically it makes it hard to understand what 0.25 means - presumably these are standardized in some way before deciding on the cut off. Some clarification and confirmation that the edges in the network are indeed unweighted, undirected edges would be good. A related question I had was how are the results robust if different thresholds were used to decide which words to connect up? Would the 111 hermit words (on p. 13) have phonological edges if a different threshold were used?

I have addressed this question in full in S2, which I have not copied here in order to save space. To clarify in brief here, the distance values were standardised within-subject, and so 0.25 reflects the bottom quartile of connectivity across each child's dataset. I have explored the validity of using this threshold in two different ways: first, I have produced figures to show

the AoP~degree correlation across 1000 different thresholds, from 0.01 to 0.99 (this approach is adapted from that of Amatuni & Bergelson, 2017). Figure 1 in the SI shows spearman's correlations across these different threshold values, and how a threshold of 0.25 compares in relation to all other options. I have also added density plots to show the distribution of different phonological distances between all words in the dataset (SI Figure 2). Second, I re-ran the GLMER models to test whether results would differ across 7 different thresholds (mean connectivity in the network, the lowest and highest viable thresholds according to the correlation plots generated above, and then the lowest quartile of each of the French and English Target and Actual data). The effect of INT on the data remained consistent across all thresholds, while EXT was found to be significant at the lowest viable threshold of 0.15, but not for any of the six higher thresholds. This further supports the robustness of INT as a predictor of network growth for this dataset.

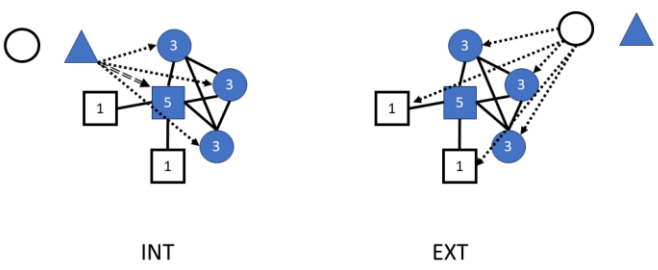
To address the final query in this point, the 111 words that did not form any edges would indeed have edges under different thresholds. This is now shown in Tables 3-4 (S2). However, in re-generating this data I realised that my original point about the 111 unconnected words was not entirely transparent, as this total was generated from Actual and Target data combined. I now show these separately across the two tables and report it more clearly in the main paper on p.14:

"*as both INT and EXT values are established through connectivity in the network (i.e. only words that form an edge with another word are represented), the words included in each network differs slightly; 54 words did not connect to any other word at a threshold of 0.25 in the Actual data, and 63 words in the Target data.*"

Tables 3 and 4 (S2) clarify how the networks would differ under different connectivity thresholds, alongside Spearman's correlation coefficients, and model coefficients with 95% confidence intervals and p values from the GLMERs.

For the methods section, I would suggest including a couple of figures that a provide a visualization of the preferential attachment and preferential acquisition models, as well as to explain the "flow" of network construction and analysis from the "raw" data to the network representation. I believe this would make the study design more accessible to readers who are unfamiliar with network analysis.

I have now added this to the manuscript on p.39, and have copied the figure here for convenience.




*"Figure 1. Visualisation of INT and EXT models of network growth. Shapes represent nodes*

*in the network and filled lines represent edges between nodes. The two images demonstrate the likelihood of two new nodes - a filled triangle or an open circle - being added to the network under conditions of INT- and EXT-like network growth. In each case, the node that would be acquired is added to the network, and new edges are shown with dashed arrows. The double-dashed arrow in the INT model shows the new edge formed with the most highly-connected node in the existing network."*

On a final note, I agree with the authors' argument that vocabulary norms represent the "average" out of many infants, which may explain why analyses of such norms are biased toward a PAQ growth model - I am curious to know if there is evidence from the current data analyses to conclude if there are individual differences in the growth patterns, that is, do children show greater or lesser extents of the preferential attachment pattern in their lexical acquisition?

This is a very interesting question and I have been grappling how best to address it. To some extent, the clear variability in the correlations shown in S3 can address this as the data clearly shows that connectivity varies across the nine infants' networks, to the extent that two infants' networks do not change predictably as vocabulary size grows (i.e. there is no correlation between age of production and degree; for one infant this is marginal, but for Anais the rho value is close to 0). This suggests that, at least for CDI words, acquisition is driven by factors other than INT or EXT (since both assume that connectivity and age should correlate negatively) for these infants.

To inspect this further, I have re-generated Figure 4 with a by-speaker grid, shown in S6, which shows the trajectory of each infant's INT values over time. Here we can see differences in the extent to which INT-like network growth drives learning across the infants, accompanied by the following text:

"*Between-child differences are well established in the phonological development literature (Maekawa & Storkel, 2006; Vihman, Ferguson, & Elbert, 1986). There is thus reason to expect that the nine infants in the current dataset may have differential paths to word production, and so results may be variable when considered across speakers. Figure 4 below shows a by-infant breakdown of the INT value trajectories visualised in Figure 4 of the main paper. Descriptively, in Anais, Nathan and Alex's data, INT values increase exponentially over time, suggesting that INT is an increasingly influential factor in their early production. On the other hand, INT values plateau relatively early in the data for William, Naima and Lily, suggesting that other factors may have a stronger influence in these infants' word production and acquisition after a certain time-point.*"

Reviewer #2: The author considers whether infant and early child spoken word production can best be predicted by high similarity to preexisting produced forms (if you know pat, bat is highly likely to be learned) as opposed to high connectivity to a variety of words in the language. The former is called preferential attachment (PAT) in the nomenclature of network science, the latter preferential acquisition (PAQ).

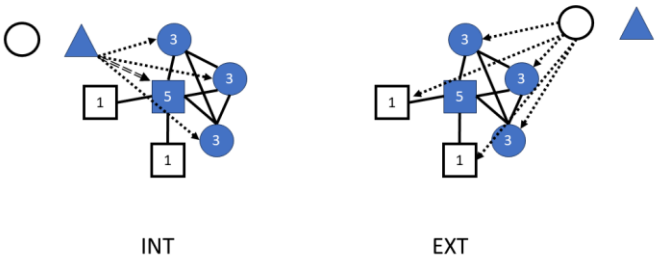
The author finds strong evidence for PAT, that is, producing more words that sound like already-produced words. This is true both of network relationships between target word forms (child is trying to produce pat, patch, panda) and network relationships between actual productions (child says "pa" when trying to produce pat, patch, panda). There is little evidence for PAQ. They argue that previous findings showing strong PAQ evidence may reflect general trends in children's acquisition (they reflect norms of data, not individual children) which would tend to obscure phonological patterning.

I am convinced by most of the material presented here. I'm especially convinced by arguments that normative data (such as 10% of kids know "bat" at age 13 months) obscures the course of individual children's production patterns. I'm still confused by exactly what PAQ is in the phonological context: I think it means something like contextual diversity (similarity to a lot of other words in the language in general, but not similarity to stuff in the child's productions), but it's not quite clear. Am I correct that there are *two* differences between PAT and PAQ, one being whether connections are within the child-specific network (PAT) or general network (PAQ), and the other whether there are connections to "popular" nodes (PAT) vs. lots of different nodes (PAQ)? If both of these are the case, how do we know which one is more important? Some  network illustrations and/or concrete examples might help readers understand this better.

I am glad to know that R2 found the work convincing and relevant in the context of the related literature! As a general first comment, please note that, following R2's recommendation below, I have re-labelled PAT/PAQ values as INT/EXT values, respectively.

To return to this first comment, R2's understanding of the two differences between EXT vs. INT is correct, though note that the "general network" referred to here is limited to the data being analysed – that is, the global network only includes words we know are eventually produced in the data, so ultimately all words included in the analysis of EXT will eventually be considered in the INT models, and vice versa. However, this difference is indeed crucial in the models: in EXT, the connectivity of the eventually-acquired network should predict learning (i.e. nodes with the highest number of edges in the end-state network are acquired earliest), whereas in INT, the state of the network at the given timepoint should predict learning (i.e. the most densely-connected nodes drive learning of new nodes that will connect to them).

To clarify the difference in predictions made by the two models, and following a request from R1 as well, I have added an explanatory figure to p.39 of the manuscript, which I have copied here for convenience.




*"Figure 1. Visualisation of INT and EXT models of network growth. Shapes represent nodes in the network and filled lines represent edges between nodes. The two images demonstrate the likelihood of two new nodes - a filled triangle or an open circle - being added to the network under conditions of INT- and EXT-like network growth. In each case, the node that would be acquired is added to the network, and new edges are shown with dashed arrows. The double-dashed arrow in the INT model shows the new edge formed with the most highly-connected node in the existing network."*

I have also expanded on the explanation in the paper with some examples, on p.5:

"*EXT-like growth assumes that forms that connect to (i.e. share properties with) a higher number of different nodes in the target network will be acquired first. EXT models of network growth thus assume that external factors in the learning environment influence acquisition – that is, forms that are most well-connected within the target language will be acquired earlier. In phonological terms, this would mean that early productions would constitute the distribution of segments and structures that co-occur most frequently in the input, thus leading early forms to resemble the statistical properties of the ambient language more closely, rather than a 'pattern force' driven by dominant features of the existing lexicon. For example, given an existing lexicon that included the forms pat and bat, an INT model would predict that a highly phonologically-similar form such as pit or bit might be acquired next, whereas EXT would predict that more variable forms would be acquired, such as /p/-initial or /t/-final words, which have high phonotactic probability in English and thus connect to a wider range of different forms.*"

In terms of the importance of the two differences outlined in this comment, I don't think there is necessarily a need to weight one above the other in general terms, though the difference in what each model predicts is of course relevant to this paper. The difference between child-specific and general network construction is a difference that is inherent in what the models each expect: one predicts that learning will be bottom-up, driven by what the child already knows; the other predicts top-down learning, driven by the linguistic environment.

Other than that, I have a few remaining questions, and some suggestions to connect the work better to language processing and language development literatures.

Remaining questions—the first two are the most important.

1. Is it possible that some of the changes in connectivity with vocabulary size are inevitable? What about some 'random word' models—if you select random sets of tokens from the full dataset and grow a vocabulary network from that, does connectivity increase/decrease the more words you have? The current data would be more convincing if random networks showed very different properties than the ones based on real data.

Thank you for this interesting suggestion, which I hadn't considered in the initial submission but I agree would strengthen the argument in the paper. I have added a short section to the first part of the Results, copied below from p.16. I have opted for what is a simpler but I think equally representative solution, by randomising AoPs in the original dataset, and then generating degrees from this. If connectivity increases with vocabulary size, as proposed in this comment, then we would expect the same outcome as is assumed by both EXT/INT models: that later-acquired words are less well-connected (i.e. negative AoP~degree correlation). However, this was not the case for the randomised data, which was in fact found to have an AoP~Degree correlation of 0.01.

"*To ascertain that this negative relationship between AoP and connectivity is not simply a given in vocabulary-based networks that increase in size over time, this analysis was re-run on an identical dataset that was randomized by AoP, such that new words were added to the French and English networks at random ages, and then the degree of each word in this random network was calculated. Across the data, there was no correlation between AoP and degree (r=0.01; p=.678); evidence for INT/EXT-like growth in the real data is thus not an inevitable outcome of vocabulary growth.*"

2. How does this finding relate to comprehension? This seems important for fully characterizing language development. Are children limited (or swayed) in their comprehension to forms they can more easily produce? Or are they simply refusing to say words they easily recognize because they are hard to produce?

One way to address this in the analysis – albeit an imperfect measure – is to include age of acquisition norms for comprehensive vocabulary in the models, which is introduced on pp.16-17 as follows:

"*As well as INT and EXT growth values, each model also included target word length in phonemes, reported age of acquisition for each item in the comprehensive vocabulary according to vocabulary checklists (CDIs; see below), input frequency in child-directed speech, word category (based on CDI word categories), and corpus (English vs. French) as fixed effects. Infant was specified as a random effect with a by-infant random slope for the effect of age. Input frequency for each word was derived from Braginsky et al.'s (2019) frequency estimates, which includes a unigram count for every word produced in adult speech in all CHILDES corpora for the respective language. Normed comprehensive vocabulary data for English and French CDI words was taken from WordBank (Frank et al., 2017); again following Braginsky and colleagues (2019), age of acquisition (AoA) was taken as the month in which >50% of children were reported to understand a given word. As comprehensive vocabulary norms are only available up to ages 16/18 months for French/US English data, respectively, 1470 total tokens did not include this measure (603 word types actoss all infants), either due to the word being acquired after the cut-off for the CDI checklist (i.e. it was included on the checklist but fewer than 50% of infants understood the word by 16/18 months), or due to it not being included on the checklist in the first place (i.e. it is included on the productive vocabulary checklist only).*"

This measure significantly predicted learning in both the Target and Actual data (b=-.22, p<.001); words that are typically acquired earlier in comprehensive vocabulary norms were acquired earlier in this data. To address this comment more directly, I don't think there is any reason to expect that infants' comprehension would be determined by their production, given the clear evidence to suggest that infants understand words at 6 months, long before they can produce them (e.g., Bergleson & Swingley, 2012). However, I hadn't considered the second point and I agree that this would be a really valuable addition to the findings presented here. I have reflected on this in the Discussion on p.27:

"*Finally, it would have been valuable to have data on these infants' comprehensive vocabularies over the course of the analysis. While comprehensive vocabulary norm data was included in the models, this is a wide step away from the expectation posited throughout this paper that individual trajectories shape learning. Comprehensive vocabulary data would allow an analysis of the extent to which known (but not yet produced) words "fit" existing segments and structures in the child's productive repertoire; in this way, models could be devised that predict which words in the comprehensive vocabulary are most likely to appear next in the productive vocabulary.*"

3. Is it possible that characterization of children's language is a little inaccurate in the sense that they may produce a form at (say) age 1 year that they later stop producing? Thus, carrying forward a previously-produced word assumes some structure that is not there?

This is a good point and is definitely a possibility in this dataset. The fact that the average word type has 24 different tokens produced across the dataset suggests that this is unlikely to be the case for a substantial proportion of words; however, it was easy to test for this.

I've now added some text to the methods that considers the number of different word tokens produced across the data, by-word, and then added a column to Table 1 that shows total token production by child. The new text is on p.12 and copied below:

"*On average, there were 32 tokens of each word type (SD = 144); 3 words occurred only once in the data, and on average each word type was produced across 6 different months (SD = 8), which supports the (admittedly imperfect) assumption made here that the first production of a word in the dataset indicates its acquisition in the infant's lexicon.*"

4. Euclidean vs. edit distance: I think the distinction the author is making is more between features and phonemes. You could calculate edit distance over features, which would be a city-block metric but still better than phoneme-based edit distance. Still, I agree that feature-based seems like a more fine-grained characterization of similarity than phoneme-based.

Thank you for this point, which I had not considered. I have briefly clarified this in the manuscript on p.11 by referring to "*Levenshtein distance in phonemes*" and "*edit distance (in phonemes)*".

Suggestion for discussion
It would be fruitful to connect this to the neighborhood structure vs. phonotactic probability literature, both in child learners and adults (children: Dollaghan, Coady and Aslin, Luce and Charles-Luce, Storkel; adults: Luce, Magnuson among others). The findings here seem to suggest that high-neighborhood words are easier to acquire (despite phonological competition?), though in practice it's hard to pull apart these influences. I was surprised not to see that any of that work referenced here. It would certainly increase the relevance of the manuscript.

Thank you for this suggestion. I have now introduced the link between phonological networks and phonological neighbourhood density in the literature review on p.5 ("*In phonological development terms, this model implies that the lexicon will constitute clusters of similar-sounding words (i.e. denser phonological neighbourhoods)*"), and have added a section to the Discussion exploring the link between the findings in this paper and those of the existing work on neighbourhood density and phonotactic probability (pp.25-26). It makes sense to me that higher PND and phonotactic probability in early development (and its facilitative effect in word learning as seen in experimental studies, e.g. Zamuner, 2009; Zamuner et al., 2014) may result from an early bias towards the acquisition of words that a child can produce (word selection). I have therefore proposed that this may be the case in the Discussion, copied below, but I welcome R2's further input on this.

"*These results align with and expand on previous work observing phonological neighbourhood density (PND) and phonotactic probability in early word learning. Both have been found to positively influence new word acquisition earlier on in development (Coady & Aslin, 2003; Dollaghan, 1994; Storkel, 2004), though for older children (Charles-Luce & Luce, 1990) and adults (Gordon & Kurczek, 2014; Vitevitch & Luce, 1999), low neighbourhood density appears to be more beneficial in learning and remembering novel words. The present findings suggest that, at least in early development, high PND (i.e. phonologically more similar words in the lexicon) may in part be derived from systematicity in production. That is, if infants are selecting new words that match their output capacity in*

*early development, then we would expect a higher number of phonological neighbours in the Target and Actual forms, as observed here, and consistent with the PND literature. On the other hand, the fact that INT predicts acquisition in both Target and Actual forms may be due to the increased learnability of words that belong to denser neighbourhoods, leading infants to produce these earlier on – the fact that they are also phonotactically similar (due to PND and phonotactic probability being correlated, Vitevitch & Luce, 1999) would no doubt support their early production as infants need to draw on fewer resources to produce a number of new words. Results in this study lend preference to the first explanation (i.e. that higher PND is motivated by production, rather than the other way around): we see a continuous increase in Actual INT values over time as new words are adapted to fit existing well-rehearsed segments and structures (i.e. existing dense neighbourhoods attract phonologically similar words for acquisition), which is significantly higher than Target INT values over the same period (see Figure 5). If higher PND was motivated by learning, we would expect to see no difference in acquisition of Actual and Target forms, since infants would be learning words that clustered together just as densely in the Target network as in the Actual network, i.e. they wouldn't be systematically adapting words to fit the dominant patterns and structures in their existing lexicon."*

Clarifications and line edits

PAT and PAQ are rather abstract. And phonologically similar, so it's hard to remember which is which (even though I can produce both ɘ ). Possible to rename them something like "highly connected" vs. "contextual diversity" or something like that?

I agree with this suggestion, as I have had issues with the phonological similarity (!) of these two terms myself when presenting this work at conferences. Following recent similar work in this area (e.g. Kalinowski et al., submitted) I have opted to refer to the two networks as INT (PAT; an internally-driven network) and EXT (PAQ; an externally-driven network). I have clarified this terminology in the paper on pp.4-5:

"*Network growth models analyse changes within a system over time, and two key models1 of development have been proposed for lexical acquisition: preferential attachment (hereafter INT due to the assumption that network growth is internally-driven in this model; note that some studies refer to this model as PATT) and preferential acquisition (hereafter EXT, due to the assumption that network growth is externally-driven in this model, note that this model is otherwise known as PAQ Hills et al., 2009; see also Steyvers & Tenenbaum, 2005).*"

And I have otherwise INT/EXT have replaced all instances of PAT/PAQ.

p. 6: "Mak and Twitchell's (2020) work with paired-association learning in adults shows that participants were better at remembering word pairs when items had been paired with highly-connected cue words in semantic space"
highly connected to the target word, or highly connected to a lot of other words? If the latter, doesn't that support PAQ more than PAT?

Thank you, I agree this is confusing! I have double-checked the paper to ensure I have interpreted the findings correctly, and re-phrased this sentence as follows (p.7):

"*Mak and Twitchell's (2020) work with paired-association learning in adults shows that*

*participants were better at remembering word pairs when items had been paired with cue words in semantic space that had a higher degree (i.e. were connected to a larger number of semantically-similar words).”*

Top of p. 7: This is a really good argument: vocabulary norming data abstracts away from the individual differences expected in early phonological development … To better understand the role of systematicity in early word production, it is essential to consider infants' actual productions of their early word forms...

Thank you for the positive feedback; I’m glad R2 finds this perspective convincing!

p. 12: "multiple tokens of a given word type in that session were 'averaged out' to one unique value for each word"
Does that mean there was a different connectivity for each token, and connectivity for the word type was the average of token values? I thought only a single form was used for each word type. So then how do you get different connectivity values?

It was not plausible to derive a connectivity for each token in the dataset, as the number of tokens was so big (n=159,043 tokens). The distinctive features for each token were averaged out (so, all sonorant values, all nasal values, etc) to create one “average” word type, and then connectivity between this word type and all other word types was established. I admit that this is not a perfect measure – I initially included only the first token of a given word in the dataset, but I prefer this option as it better represents variability in the data; results are consistent with both approaches, though. I have now clarified this on p.13 as follows:

“*multiple tokens of a given word type in that session were ‘averaged out’ to one unique set of distinctive feature values for each word, from which connectivity with all other word types was then derived.*”

p. 14, Age of production ~ connectivity correlations. Is there any other way this could have turned out? That is, is it possible that, given an increasing vocabulary, later items will necessarily be less well connected?
I included this component of the analysis because this is typically the first step taken in the analyses of the previous work I drew on here (Hills et al., 2009; Fourtassi et al., 2020; Amatuni & Bergelson, 2017), and to my mind it is a sensible first-step validity check of the data. However I agree that it seems almost a given that a negative correlation would be identified in vocabulary acquisition data. To check this, I randomized my dataset by AoP (age of production), re-generated connectivity values across the data, and then re-ran the correlations on this random sample. The correlation was non-significant, and this time the direction was positive. This reassures me that the overall network structure is indeed in line with assumptions made by INT/EXT, and not just an artefact of the data.

I have re-worded some of this paragraph, and added the additional results, on p.15:

“*First, to assess the broader assumption that connectivity in the network will change systematically over time, regardless of whether that is through INT- or EXT-like changes, the relationship between age of production (AoP) and connectivity (degree) in the static network was considered. Both INT and EXT models of network growth assume that later-acquired words will be less well-connected in the network. Across all infants, there was a mean AoP~degree correlation of r=-0.21 (Spearman’s, SD=0.09; English: r=-0.26, SD =0.04; French: r=-0.15, SD=0.11); overall, later-learned words were less well-connected in the*

*networks. Negative correlations were found in all children's data; these were all significant at p<.05 except Anais (French corpus). See Table S3 and Figure S3. This is consistent with previous similar work showing that earlier-learned words are more highly-connected in the network (Fourtassi et al., 2020; Hills et al., 2009), and replicates these findings with a naturalistic sample of infant production data. To ascertain that this negative relationship between AoP and connectivity is not simply a given in vocabulary-based networks that increase in size over time, this analysis was re-run on an identical dataset that was randomized by AoP, such that new words were added to the French and English networks at random ages, and then the degree of each word in this random network was calculated. Across the data, there was no correlation between AoP and degree (r=0.01; p=.678); evidence for INT/EXT-like growth in the real data is thus not an inevitable outcome of vocabulary growth."*

p. 15 (not really a question for this page as much as a general one): were to-be-learned words just selected out of the total word types in the sample, or some larger dictionary of English or French word types?

These were selected out of all words in the sample. I have now clarified this on p.13:

"*an as-yet-unknown word (i.e. all the words in the global network - that is, all words produced in a given child's data up to and including age 2;6 - that had not yet been produced)"*

p. 16: "less frequent words were more likely to be learned, as were words with a higher token count"
Higher token count makes sense, but less frequent words doesn't. Is there some sort of interaction/suppression going on between token count and word frequency? Where did the word frequency information come from? Oh, I see—token count in the CHILD's speech, not the mother's speech. It would be a lot clearer to keep this bit of info present, for example, n tokens (child), n tokens (mother) or something like that. It still doesn't make sense to me why the mother's speech effects would go in the directions they do, though. Or is frequency calculated across all mothers rather than being specific to each child's mother?

To avoid confusion, I have re-labelled 'Word frequency' as 'Input frequency' throughout the paper. I have also removed the token count variable from the models, as, due to the addition of comprehensive AoA into the model, both n tokens and vocabulary size were removed to avoid multicollinearity.

Multiple reviewers queried the direction of the input frequency result, and so I opted to use a more robust measure of word frequency across the dataset. The initial measure drew on token counts of each child word produced in their own mothers' speech – while I think there is real value in using such a child-specific measure, because the dataset was filtered to include CDI words only, there were ~125 word types in the data that did not appear in mothers' speech and were coded as 0. There was also high variability in the number of words produced across corpora and infants.

Instead, I now include an input frequency measured developed and used by Braginsky and colleagues (2019), which incorporates token counts from all adult speech in the given language on CHILDES. This measure has been applied in a number of previous studies, including Fourtassi et al. (2020), which I draw on heavily in this paper. This measure includes an input frequency count for every word in my dataset.

As shown in Table 3 on p.37, this measure of frequency significantly predicts learning in the expected direction: more frequent words were acquired earlier (Actual: b=.17, p=.001, CI: [0.07,0.27]; Target: b=.19 p<.001, CI: [0.09,0.29]).

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, 3, 52-67.

Reviewer #3: The manuscript is well-written and the project makes a strong contribution to an existing literature modeling word acquisition using network growth models. In particular, it shows that when individual children's actual productions are analyzed (as opposed to analyzing target form productions of aggregated parent-reported vocabulary), systematicity between new words and words already in the child's productive repertoire is clearly apparent.

I am glad that R3 found the paper well-written, and sees the value in its contribution! As a general first comment, please note that, following another reviewer's request, I have re-labelled PAT/PAQ values as INT/EXT values, respectively.

There are several places where I think the paper could be clearer and I have a few questions about how particular modeling decisions might have affected the results:

Abstract: The statement "systematicity is most apparent in early acquisition" doesn't seem to me to match the results. Rather, systematicity increases over the age range included in this study.

I have now changed this statement in the Abstract to "*systematicity becomes increasingly apparent over the course of acquisition*".

Intro., p. 5: I wasn't following how PAQ necessarily implies dominance of external factors. In general, couldn't the motivation to maximize diversity of connections to a new node potentially be entirely internally driven? I can see the logic of assuming PAQ represents a less internal-motor-constraints-driven mechanism, in the current project, so my suggestion is just to be a little clearer about the logic behind that assumption here in the intro.

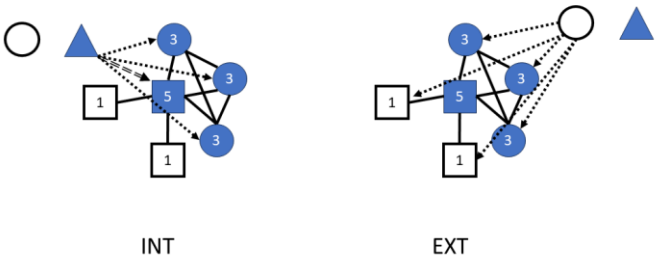
My understanding is that EXT/PAQ models of network growth are defined as being driven by external factors in the learning environment (as set out in Hills et al., 2009; Fourtassi et al., 2020). I see R3's point about the potential for EXT-like growth to be internally-driven, but as I understand it this would have different theoretical implications to those of EXT/PAQ.

That being said, all three reviewers have asked for a clearer explanation of EXT models in particular. I have therefore expanded on the explanation in the paper with some examples, on p.5:

"*EXT-like growth assumes that forms that connect to (i.e. share properties with) a higher number of different nodes in the target network will be acquired first. EXT models of network growth thus assume that external factors in the learning environment influence acquisition – that is, forms that are most well-connected within the target language will be acquired earlier. In phonological terms, this would mean that early productions would constitute the distribution of segments and structures that co-occur most frequently in the input, thus leading early forms to resemble the statistical properties of the ambient language more closely, rather than a 'pattern force' driven by dominant features of the existing lexicon. For example, given an existing lexicon that included the forms pat and bat, an INT model would predict that a highly phonologically-similar form such as pit or bit might be acquired next,*

*whereas EXT would predict that more variable forms would be acquired, such as /p/-initial or /t/-final words, which have high phonotactic probability in English and thus connect to a wider range of different forms.*"

Intro./Methods: A schematic illustration to help the reader understand PAQ and PAT network growth models as applied to the current domain might enhance the paper.
All three reviewers have requested an illustration of the two models. I have added an explanatory figure to p.39 of the manuscript, which I have copied here for convenience.




*"Figure 1. Visualisation of INT and EXT models of network growth. Shapes represent nodes in the network and filled lines represent edges between nodes. The two images demonstrate the likelihood of two new nodes - a filled triangle or an open circle - being added to the network under conditions of INT- and EXT-like network growth. In each case, the node that would be acquired is added to the network, and new edges are shown with dashed arrows. The double-dashed arrow in the INT model shows the new edge formed with the most highly-connected node in the existing network."*

p. 8: "PAQ thus assumes that Actual and Target networks do not differ": PAQ as a general approach seems agnostic to whether Actual and Target networks are the same or different. It's conceivable that PAT would best fit Actual data and PAQ would best fit Target data. So for the theory that Actual and Target data both grow in a PAQ fashion, perhaps some other descriptive term(s) could be used.

I see your point here and agree that the statement quoted above is not fully in line with expectations regarding INT and EXT. I have re-phrased this statement as follows:

"*the question of differences between Actual and Target forms is thus not of central theoretical interest for EXT models in this analysis*" (p.8)

p. 11, top paragraph: How was syllable structure factored into phonological distance? Also, a table of examples of pairs plus their phonetic features and phonological distance scores would be helpful to make the methods clearer.

This query aligns with a similar question from another reviewer, who also requested a table showing how phonological distance between word pairs was calculated. I have thus copied my previous response here:

I have added two tables to the SI that shows the distances between 3 target words in the dataset, each with a different phonological structure (*baby* compared with *balloon* and *sky*).

The table presented here is adapted from Monaghan et al. (2010), on which my own phonological distance measures were based when writing this paper.

I have also clarified this approach in the main text on pp.11-12:

"*This means that two words that differ only in their vowel segments are coded as the same in the current analysis. Words were aligned by syllable nucleus: onset consonants were compared with other onset consonants,, and codas were compared with codas. Full criteria for establishing distance, alongside tabulated examples, are included in the Supplementary Information, S1."*

And added the following explanation in the Supplementary Materials:

*"Phonological distance was established following Monaghan, Christiansen, Farmer and Fitneva's (2010) approach, with some adaptations. Note that in their study, only monosyllabic words are included, and so their approach is adapted here to include multisyllabic words. Following their method, each word was first divided into a series of 'slots', according to its phonological structure. For example, the word baby was separated into five slots: /b-e-i-b-i/. Because vowels were not accounted for, the nucleus of each syllable - both monophthongs and diphthongs - was then replaced by a generic V slot, i.e. /b-V-b-V/. Words were then aligned by nucleus to generate a phonological distance measure between each possible word pair. All consonants at word onset and final syllables were aligned, regardless of syllable number, such that the final /d/ of bed would be aligned with the final /n/ of balloon. This is because infants may have a tendency to produce only certain consonants word-finally, and so this approach would capture such systematicity. For the English data, the maximal word structure considered in the analysis is C-C-C-V-C-C-|C-C-C-V-|C-C-C-V-|C-C-C, where syllable boundaries are marked with a |. This accounts for complex clusters at word onset (e.g. splash /splæʃ/, C-C-C-V-C), coda (plant /plænt/, C-C-V-C-C), and across syllable boundaries (pumpkin /pʌmpkɪn/, C-V-C-C-|C-V-C). In French the maximal structure was C-C-C-V-C-C-|C-C-C-V-|C-C-C-V-C|C-C-C-V-C|C-C-C-V-|C-C-C-C. This accounted for multisyllabic target words such as hélicoptère ("helicopter" /elɪkɔptɛʁ/, V-|C-V-|C-V-C-|C-V-C) and appareil photo ("camera" /apaʁɛjfɔto/, V-|C-V-|C-V-C-|C-V-|C-V), and complex codas as in arbre ("tree" /aʁbʁ/, V-C-|C-C). For vowel-initial words, the C1 slot in word-initial position is empty, but all other alignments remain the same. This maximal structure is required in an analysis of infant word production, to account for unexpected complexities such as, production of French mettre "to put" as [mɛʁstʁ] and étoile "star" as [ɛstwal]. In the infant data, it was not always easy to determine exactly where a syllable boundary should occur in complex productions, in part because this was not predictable based on the target form due to the variability in production, so would have to be done on a word-by-word basis, and in part because the syllable boundary of some productions could not be clearly established from its phonetic transcription. Instead, consonants were always assigned to the syllable-initial cluster, rather than assigning part of the cluster to the coda of the previous syllable (e.g. for the examples above, the infant production of mettre was coded as C-V-|C-C-C-C and étoile as V-|C-C-C-V-C)."*

p. 11, bottom paragraph: An alternative to analyzing Target forms of all words produced lol by 2;6 would be to analyze the adult productions from the same sessions. Some mention of alternative ways of representing external targets and rationale for the current approach rather than alternatives would be nice to see added.

I'm not sure if I've fully understood this comment, since the Target forms essentially are the "adult" forms, though granted not the words that the caregiver necessarily produce. By using the Target forms, the analysis essentially replicates that of previous work that has drawn

exclusively on vocabulary norm data, with actual production age, rather than normed AoA data, as an index of developmental time.

However, I agree that the rationale for making this decision would be valuable in the Discussion, and so I have added the following on p.23:

"*It may also be the case that the representation of the Target network was not sufficiently aligned with the reality of the end-state network that the infants will acquire. Analysing the Target network on a larger scale – for example, including all words produced in the infants' inputs across their recordings, and building a network based on which of these words infants produce in the dataset – might better represent the role of EXT-like growth on early word learning. This is an avenue that could plausibly be considered in future work.*"

p. 12: It seems a limitation that semantic info. is not included in these models. Can that be added to the Discussion as a limitation?

I have added a sentence to the Discussion on p.27 to consider how future studies could build on this one to consider semantic networks:

"*The influence of semantic networks on acquisition has also not been considered here - further studies may want to analyse similar naturalistic data to consider semantic network growth within infants' actual productions, or even combine indices of semantic connectivity with that of phonological networks to observe how/whether the two interact in early development."*

p. 16, top paragraph: Confidence intervals on the estimates could potentially provide evidence that PAQ is significantly less predictive than PAT. Currently, the conclusion rests, technically, on (over-)interpretation of a null result. The chi sq. values are very different, so I would bet the difference is significant. Moreover, confidence intervals could potentially show that even though PAQ and PAT are both predictive for Actual data, PAT is a better predictor that PAQ for Actual data as well.

I have now added confidence intervals to the model summaries in Table 3, as well as to the relevant tables in the SI.

p. 16, middle paragraph: The result that "less frequent words were more likely to be learned" is very counter-intuitive. I imagine this is related to the way the model is set up, but I'm having a hard time reasoning beyond that. Can you provide some possible explanations?

Thank you for flagging this up; R2 also had a similar query. For convenience, I have copied my response to their comment below:

Multiple reviewers queried the direction of the input frequency result, and so I opted to use a more robust measure of word frequency across the dataset. The initial measure drew on token counts of each child word produced in their own mothers' speech – while I think there is real value in using such a child-specific measure, because the dataset was filtered to include CDI words only, there were ~125 word types in the data that did not appear in mothers' speech and were coded as 0. There was also high variability in the number of words produced across corpora and infants.

Instead, I now include an input frequency measured developed and used by Braginsky and colleagues (2019), which incorporates token counts from all adult speech in the given language on CHILDES. This measure has been applied in a number of previous studies,

including Fourtassi et al. (2020), which I draw on heavily in this paper. This measure includes an input frequency count for every word in my dataset.

As shown in Table 3 on p.37, this measure of frequency significantly predicts learning in the expected direction: more frequent words were acquired earlier (Actual: b=.17, p=.001, CI: [0.07,0.27]; Target: b=.19 p<.001, CI: [0.09,0.29]).

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, 3, 52-67.

p. 17, top paragraph, "the direction of this effect is not as expected: in the Actual data, PAT values of newly-learned words are lower earlier on in development, while they are higher in the Target data": What are some potential explanations and implications of this? It would be good to add some elaboration to the Discussion section about this.

Following other reviewer comments, I have changed the model structure slightly to include comprehensive age of acquisition from vocabulary norms, and the new input frequency measure from Braginsky et al (2019), and have removed vocabulary size and token frequency owing to multicollinearity. As a result, only the Age*INT interaction is significant in the Actual data, which again is positive. The Age*EXT interaction remains negative, according to model estimates, but is now non-significant. I have consequently removed this section from the analysis.

p. 17: For the GAMM analysis, was the unit of analysis words learned at a given age point? Or all unlearned words at an age point? Or something else?

I have reworded the text on p.19 as follows:

"*The data was subsetted to include only INT values at the time-point immediately prior to the word's production as the dependent variable in the model (i.e. for a word produced at 17 months, its INT value at 16 months was analysed); higher INT values are expected to predict that a word would be learned in the next month."*

p. 18, top paragraph, "the data was subsetted such that only the PAT values at the time-point immediately prior were analysed": Was this also done for the linear age models presented earlier? If not, why not?

This approach wasn't taken on the GLMER models because the dependent variable was whether or not a word was learned at the next time-point based on the INT/EXT values; all time-points were thus included for each target word. In contrast, the GAMM models are testing how the INT values change over time for learned words, as vocabulary grows. Sub-setting the data in this way for the GLMERs would require a full change of the model and all predictions.

I have clarified this in a short addition to the text on p.19:

"*To explore these results further, GAMMs were run using the mgcv() package in R (Wood, 2011), to observe how INT values changed over time as new words were learned."*

p. 19, middle paragraph, "in both Actual and Target data, earlier-acquired words tended to have lower PAT values, while later-acquired words had higher PAT values": Could it be because of vocabulary size limiting the number of connections between words? What if that

were somehow accounted for in network construction, e.g. by choosing the phonetic similarity threshold so as to always have a consistent proportion connectivity?

To address the first point, and following Hills et al. (2009), I have now scaled the INT and EXT values in the GLMER models not only by speaker but also by age. This accounts for the changing network size over the course of the analysis. I have clarified this on p.17:

"*INT/EXT values were scaled by speaker and age to account for the effect that increased vocabulary size at each month has on INT/EXT values (i.e. when the network is bigger, a newly-acquired word has the opportunity to connect to a higher number of different words by default).*"

The significant positive Age*INT interaction continues to be significant in the Actual but not the target data, even when changes in vocabulary size are controlled for (see Table 3).

p. 22, bottom paragraph: "it appears that infants are selectively acquiring forms that match their own production preferences": I think it should also be noted that parents may be selectively eliciting/highlighting words they think their child is capable of producing.

While I absolutely agree with this point, I don't think it would explain the fact that infants' words are phonologically similar to one another. However, I agree that the fact that the words could well be elicited in some way – whether explicitly or implicitly – is relevant to the claims being made in this paper, so I have added a note to the methods section acknowledging this possibility:

"*Note that while interactions were naturalistic and thus not at all directed by the original researchers, the data was not coded for infant productions that were imitated from or prompted by the caregiver, and so data includes both spontaneous and non-spontaneous infant productions."*

p. 23: It could also be nice in the future to consider babbling forms within the network analysis.

Thank you for this suggestion, which I completely agree with. I am about to start data collection on a longitudinal sample that would incorporate babble, with the intention of doing just that. For now, the kind of data required for this sort of analysis is not easily available.

p. 23, last paragraph, first line: Please change "within a networks account" to "within a phonological networks account".

This has been amended in the text.

p. 34 (Table 4): Why is there PAQ for Actual when PAQ only applied to 2;6 Target forms?

This is the analysis of the Actual network (i.e. the connectivity of all words based on their Actual form) using EXT values that were generated based on connectivity of the Target form. EXT values are generated from Target forms because they represent the final network that the child is moving towards, but given that connectivity differed between Actual and Target networks I analysed both for completeness. I have clarified this on p.14:

"*As EXT-like growth is assumed to represent the connectivity of words in the ambient language, global networks were established with Target forms only, since the way infants produce words in the existing network is not relevant to this model. However, given that connectivity differs across Target and Actual networks (i.e., the known words in the Actual network at month n may be different from the known words in the Target network in the*

*same month), both Actual and Target network structure will be tested in the analysis.*"

p. 35 (Figure 1): Please add a color legend. When printed grayscale, I can't tell which is red and which is blue.

The legend was already in colour so I imagine the issue is with the rendering within the manuscript submission system. I have changed the line types in the figure so that hopefully these are more easily differentiated in greyscale.

p. 36 (Figure 2): I thought PAQ was only done on Target data, not Actual data? I think the text mentioned that both were plotted for exploratory purposes, but I think a little more detailed explanation could help, for understanding this as well as Table 4.

As per the comment about Table 4 above, I have now clarified this in the Methods section.