Dear Dr. Sloman,

Thank you for the thoughtful comments on our manuscript, “The length of words reflects their conceptual complexity.” Please accept our resubmission. We have addressed your comments and the comments of the reviewers and action editor and we believe that the manuscript is substantially improved. Please find below a point-by-point response to the comments.

Please do not hesitate to contact us if you have any questions or concerns. We look forward to your consideration of this revision.

Sincerely,

Molly Lewis and Michael C. Frank

Reviewer # 1:

> *page 4: Saussure (1916, 1960)'s -> Saussure's (1916, 1960)*

Done.

> *page 6: The ability to link this line of research to operationalized semantic primitives may be an ultimate test of the claims that the authors are making. I do realize that this is a difficult step to make, and do not propose that this is done within the scope of the present review. What would be helpful for the reader however is a stronger contact with the existing literature on semantic primitives, including seminal work by Wierzbicka.*

Thank you for pointing us to this body of work. We have added a paragraph in the Introduction that describes how our work is related to this prior work on primitives by Wierzbicka and colleagues. This work is indeed very relevant to our study, and we feel that our main proposal and findings are now more clear by distinguishing them from this prior work.

> *Similarly, little is said about the decades of linguistic work on principles of iconicity, even though its quantity principle seems to pre-date (with an identical formulation) Complexity Hypothesis 2 in the present paper. To quote from Haspelmath's "Frequency vs iconicity in explaining grammatical asymmetries" "Greater quantities in meaning are expressed by greater quantities in form". I believe prior massive work on iconicity by Sapir, Kay, Jakobson, Moravcsik, Haiman and Haspelmath deserve a much more thorough treatment in the paper, even if because it appears to have obtained much linguistic support both within and across languages for the arguments that the authors make. Equally, the authors may want to explain to their linguistically oriented readers how their claims expand on this prior work.*

Thank you for highlighting this body of work, as well. We have expanded our discussion of prior work on markedness and iconicity in the Introduction. We have also highlighted how our complexity hypothesis is distinguished from this prior work, with reference to Haspelmath's claims about frequency.

> *caption to Figure 2: non-paramedic -> non-parametric*

Done.

> *page 33: "The length-complexity correlation holds even more strongly for words below the median in concreteness, those words whose part structure is presumably much less obvious". I don't see an immediate link between how concrete a word is and the morphological structure of the word. The English lexicon (to take one) is full of perfectly parseable compounds and derived words, whose meaning is anything but concrete (e.g. hogwash).*

Thank you for this helpful point. We agree, it is not obvious that less concrete meanings are less likely to be compounds. The point we were hoping to make was that an iconic link between the parts of a word and parts of a referent is less clear when the referent is abstract. We have

replaced this analysis with a comparison of object labels vs. non-object labels, as requested by Reviewer #3.

Reviewer #2

> *First, if this paper were written before the Piantadosi 2011 paper and the Mahowald 2013 papers, I could understand why the authors would feel that by controlling for frequency in experiment 9, they could get away from the Zipfian assumption that the main predictor of word length is frequency. By now the common assumption (e.g. Seyfarth 2014; Cognition) is that predictability, rather than frequency, should be the main determiner of length. For the first two series of experiments the authors may have felt that they could evade that issue by constructing novel objects, or objects for which no prior predictability of frequency accounts would apply, but that's a very narrow view of predictability-based accounts. For instance, in the artificial object experiment series, if number of parts predicts complexity, it is not far-fetched to assume that subjects do have some predictability assessment for the novel objects, based on a naïve P(object) = <PI>[P(parts)], which would make objects with more parts less predictable, everything else being equal (and the authors seem to be aware of that on page 5, yet maintain the argument for complexity). That is, to rule out a predictability-based explanation, the authors need to contrast their findings with a model that can assign probabilities to first-seen items. If that's what the authors mean by "complex", then it should be spelled out. This is not only an issue for the artificial objects. For instance, body part names seem not to correspond to how difficult it is to understand their action, but rather to how likely we are to see them: ears, eyes and tongue are arguably more complex than intestine, muscle and cartilage. For existing words in Experiment 9 the authors should control for mean predictability rather than frequency (preferably for both), and then use whatever residual effect there is in Experiment 10, not bare complexity. Otherwise they do not rule out the possibility that frequency and predictability account for the correlation, at least in the lower pearson r languages. Without clearer controls for predictability and frequency, the experiments in this paper only show that subjects expect language to match their experience in matching longer words with less predictable and less frequent items, which is still interesting, but perhaps shouldn't take 10 experiments to prove.*

We appreciate this important point about the relationship between our work and previous work on predictability. We have two different responses, one for the part of our work that concerns inferences about novel words and one for the part that concerns known words.

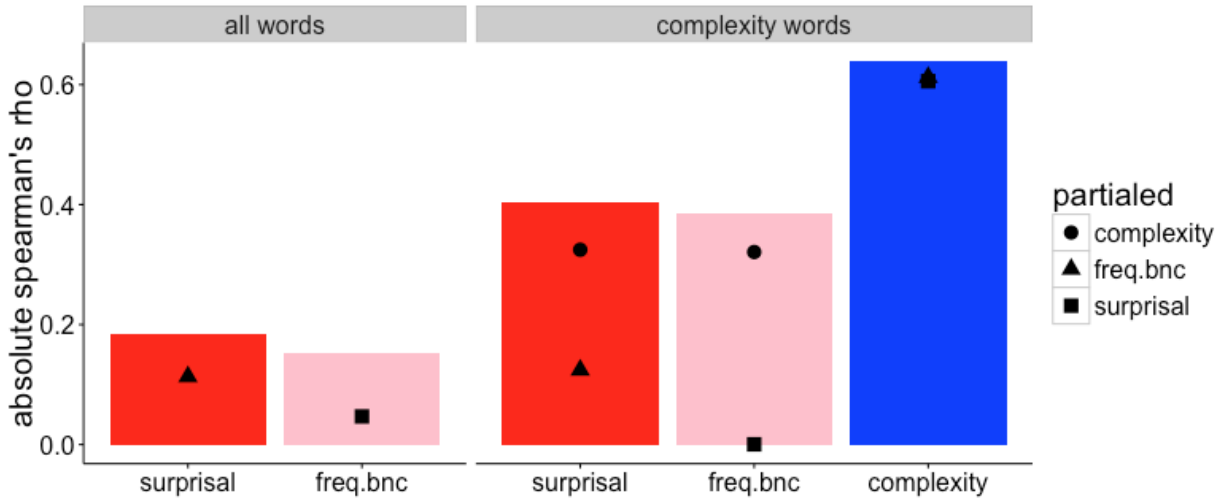
Novel words. In relating our work to previous on predictability literature, we believe that it is critical to distinguish between two types of predictability: word versus referent predictability. The work cited above concerns the predictability of a word in its linguistic context, not referent predictability. Because our words are novel in our experiments, participants cannot rely on linguistic predictability to guide their judgements. We agree, however, that predictability based on the number of object parts may underlie our effect. As we note in the Introduction, this account is entirely consistent with our findings. If true, this finding would be novel since no prior research has demonstrated an effect of referent predictability on word length.

An alternative hypothesis about referent predictability is that the relevant dimension is the predictability of the entire object over time, rather than the likelihood of the co-occurrence of

parts that make up an object. This account would explain why a complex but frequent object like the “eye” has such a short label. However, we think this possibility is unlikely given Experiments 11 and 12 (described in the Supplementary Information). In these experiments, we manipulated the raw frequency of objects and found no evidence that less frequent objects tended to be mapped to longer labels.

Known words. In the case of existing words, we agree that it is important to know how our measure of referent complexity is related to word predictability (surprisal), and we appreciate this suggestion. We have added an additional analysis to Experiment 9 in the paper that controls for the surprisal. We used the bigram measure of surprisal reported in Piantadosi et al. (2011) from the BNC corpus.

We replicate the Piantadosi finding, and find that the relationship we reported between length and complexity remains reliable after partialing out the effect of surprisal. Below we present the correlation coefficient between predictability measures (log word frequency in the BNC corpus, surprisal, and complexity) and word length for both the full set of words (left facet) and our subset of 499 words (right facet). As in Piantadosi et al. (2011), we report Spearman correlations. The symbols on the bars show the correlation between predictability measures and length in terms of number of characters, partialing out one of the other factors (where the partialed factor is indicated by the shape of the symbol).



For all words, the correlation between surprisal and length is greater than the correlation between frequency and length (z = 10.35, p<.001). For our set of 499 words, surprisal and complexity are weakly correlated with each other (ρ = .25, p<.001). As in previous work, we find that surprisal is correlated with length (ρ = .40, p<.001). Critically, we find that the relationship between length and complexity remains reliable partialing out surprisal (ρ = .59, p<.001). This analysis suggests that our measure of complexity explains different variance than that accounted for by surprisal.

> *Second, the authors seem to change their definition of "complex" throughout the paper. For the first series of experiments, complex objects are objects that have more parts, as*

the authors construct artificial objects from object primitives, making the assumption plausible in that context. An identical paradigm is replicated in norming real-world objects starting in Experiment 4, in which the relationship between number of parts and complexity is at most implicit. However, in Experiment 9, the authors explicitly define "complex" as "difficult to understand and has many parts." The first series of experiments did not rely on number of parts explicitly. What made the authors choose to change their definition of complexity? I could accept "difficult to understand" as a proxy for "complex", but outside the artificial objects experiments, I see no reason why number of parts should predict complexity. A valve is complex and has only a few parts, while a necklace is simple and has many parts.

Thank you for this comment. We appeal to a broad, intuitive definition of complexity throughout the paper that is motivated by the notion of semantic primitives. We see these primitives as conceptual rather than concrete, and so they may be applied to meanings both abstract and concrete. In Experiment 9, we view our explicit definition of complexity, "difficult to understand and has many parts," as a strength rather than a weakness because it allows us to operationalize complexity for more abstract meanings in the same way it was manipulated visually in the object experiments. This similarity makes it more likely that the same conceptual construct underlies the bias in both Experiments 1-8 and Experiments 9 and 10.

However, as you suggest, a limitation of our work is that we are not able to provide a more precise definition of the nature of these parts. We see this work as an important first step in testing the general hypothesis that complexity---operationalized as the number of conceptual parts---is related to word length. Future work will be needed to understand what exactly the nature of these primitives is. In our revisions to the General Discussion, we highlight this limitation more fully.

page 5: such a unit exists -> such units exist

Done.

page 7: saying that "some" is ambiguous between "one or more, possibly all" and "one or more, but not all" ignores decades of research in semantics and more recently in psycholinguistics. This entire section seems to conflate a word having multiple meanings, multiple referents (for indexicals), and multiples uses.

Thank you for this point. We have removed this discussion from the paper.

page 11: truth functional -> truth conditional?

Done.

page 14: I don't see why the markedness account cannot be extended beyond the morpheme level, and it certainly does taken to apply at the phonemic level.

Thanks for this comment. The markedness account could indeed be extended to the morpheme level, but previous work has not done this. We see our contribution as providing a broader and more quantitative account of the relationship between length and meaning. It is difficult to

directly distinguish our proposal from work on markedness, because this work does not make predictions about differences in markedness beyond particular semantic domains. For example, under a theory of markedness, why would the word “generation” be more marked than the word “bed”? In contrast, we are able to predict the difference in length of these two words on the basis of our measurement of conceptual complexity.

figure 2: why aren't 1/1 grouped with 2/2, 3/3, 4/4 and 5/5?

The objects in each condition were randomly sampled from the first quintile of objects based on the complexity rating norms. The fact that the 1/1 group is greater than 1 may reflect either variability in this sampling process, such that objects on the left on average had slightly higher complexity ratings than objects on the right, or simply measurement error.

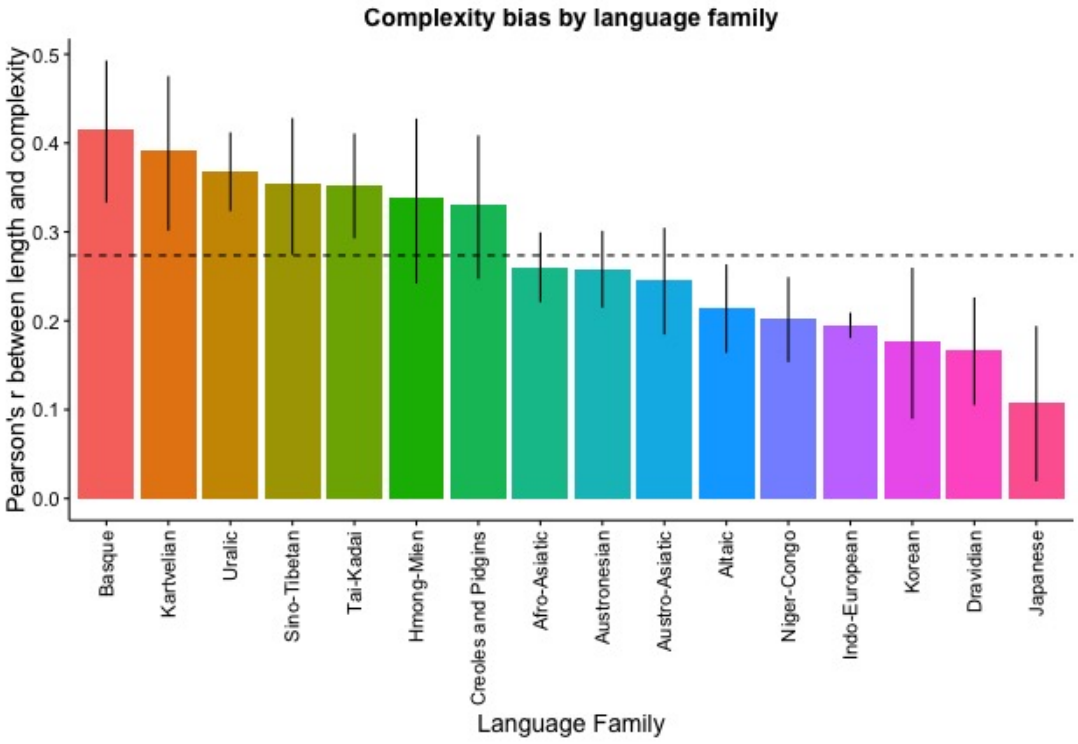
page 29: which previous work?

We were referring to all previous work that has used a rating scale approach to measure semantic variables. As an example of this body of work, we refer to the MRC corpus (Wilson, 1998). We have moved this citation earlier in the paragraph.

Study 10: English and other Germanic languages use compounding (cell+phone, micro+wave) as a way to create new words. This is not a strategy all other languages take. Of the first 5 languages, 4 are Germanic.

Thank you for this observation. We agree that part of the effect may be related to the extent to which a language uses concatenation as a morphosyntactic strategy. However, the robustness of this effect across all the languages we analyzed suggests that this cannot fully account for the bias. Furthermore, we find a relationship between length and complexity in the subsample of only monomorphemic words in our dataset for English words (Study 9; n = 387; $r_{phonemes}$ = .53, $r_{syllables}$ = .47, all ps < .0001). We also find a reliable correlation between length and complexity across languages subsetting to only monomorphemic words in English (Study 10; r= .23).

To also address this concern, we added an additional analysis to Study 10 that controls for the non-independence across languages due to genetic relationships and language contact. Following Jaeger et al. (2010), we fit mixed effect models to control for these variables, using data from the WALS dataset (Haspelmath, et al., 2005). The relationship between length and complexity remains reliable even after controlling for these variables. Below we present the correlation between length (characters) and complexity for each of the 16 language families present in our dataset. Across all 16 families, there is a reliable correlation between length and complexity.



Reviewer #3:

[statistical comments]

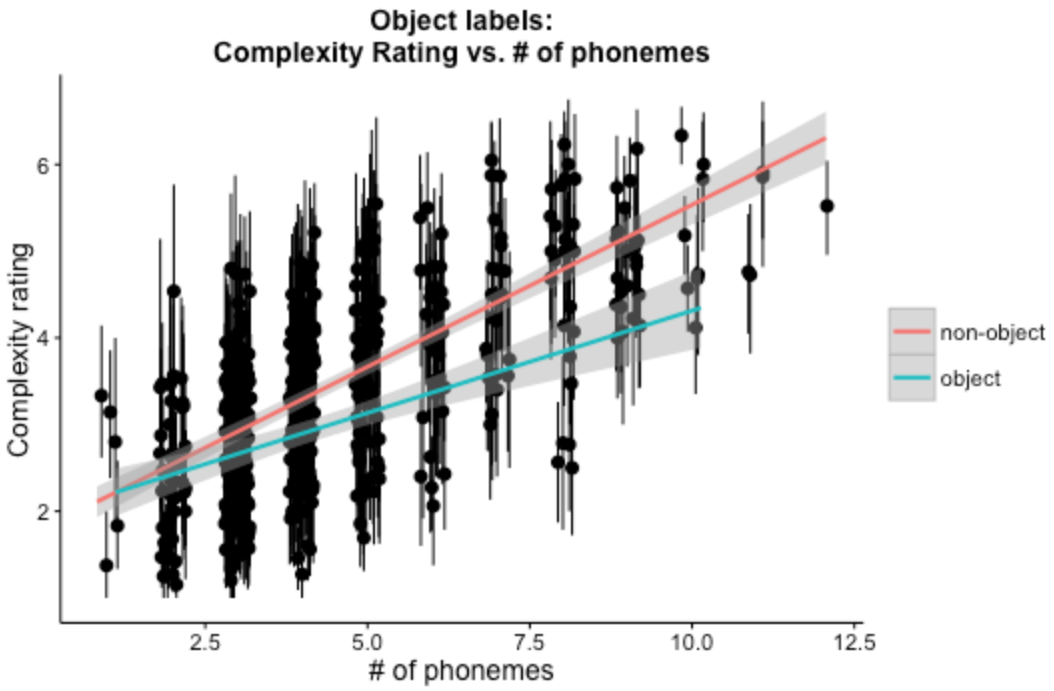
*1. The details of the statistical analyses are rather opaque - the authors make extensive use of correlations, but (presumably in the interests of brevity) it's often not clear what exactly the correlations are being calculated for. I detail these points of uncertainty in a few places in the annotated manuscript, but in general I think rather more information is needed on how the statistical tests were conducted. Given that there are essentially 10 experiments in the paper there's a danger this could make the paper rather unwieldy, so I think the authors' approach to keep the analyses as simple as possible is sensible; furthermore, the results *look* pretty striking, so I don't think the authors are concealing questionable decisions in the analyses by glossing over these details. Nonetheless I think it's important that the details of the analysis are available for evaluation by readers (particularly sceptical readers).*

pg. 19: Is this 'just' a correlation? how you handle the repeated-measures aspect of the data? Bit more detail needed on analysis.
pg. 20: Details of analysis in terms of random slopes etc? If this slope is negative, does that mean this is log-odds of selecting the simpler object, not the more complex object as the framing implies?
pg. 25: so is this a paired samples t-test where you calculate mean length for the two quantiles per participant? Why do that rather than e.g. a regression analysis which wouldn't require averaging over items?

We agree that the nature of these correlations were confusing and apologize for this issue. We have added a footnote in Experiment 1 that clarifies our analytical methods. In the majority of cases we conduct simple Pearson correlations across means of items, except where noted. Also, in response to this and other feedback, we have added mixed-effect models in experiments 3, 6, and 7 to control for the repeated-measure design.

> *2. I found the norming study slightly odd when set against the first 8 experiments. Experiments 1-8 are focussed squarely on object complexity, and the instructions for the norming study (Experiment 9) also focus on this type of complexity (referring specifically to number of parts). However, the norming and corpus studies consider a far wider range of word types - for instance, participants are asked to rate the complexity of "a" in these terms. The results come out as predicted, even after controlling for frequency, but I do wonder what the participants are doing here, and how they reconcile their instructions to the types of words they are asked to rate. The authors also note that the correlation between complexity and length is *higher* for less concrete words; presumably this means that if the norming study was restricted to object labels (i.e. made more similar to to Experiments 1-8) the effect would be weaker? Can you run the analyses on just the object labels in this corpus and see if that's the case, to make the artificial and natural cases more directly comparable? In general, I think a little more work could be done considering how this norming study for English relates to the preceding experimental work.*

Thank you very much for this informative suggestion. We ran this analysis for the subset of words that were object labels (n = 163). The correlation was in fact lower for object labels ($r_{phonemes}$ = .44, p<.001) relative to non-object labels ($r_{phonemes}$ = .73, p<.001). We also constructed a linear model where complexity, object label status and their interaction were included as predictors of number of phonemes. In this model, object label was a reliable predictor of complexity (B =-0.29, t = -10.61, p<.001; plot below). These effects remained reliable when controlling for frequency. We have added this analysis to the paper, along with a comment about how it relates to the experimental work.



*3. I like the fact that the authors attempt to see if this effect generalises to other languages, but I do wonder exactly what this analysis (Study 10) shows. I *think* what it shows is that word lengths are correlated across languages - so modulo a language-specific scaling factor, words that are long in English will tend to be translated as long words in these other languages. Then, because length correlates with rated complexity in English, it also correlates in these translations. I think it's intriguing that length correlates across languages, even when controlling for frequency, and the complexity effect is one possible explanation for this, but I think the authors should acknowledge the somewhat more indirect nature of the correlations they show.//pg. 37: Thinking aloud: what this result actually shows is that words which are translations of each other tend to be of similar length (modulo some language-specific scaling factor)? and this is true even when controlling for frequency. and *then* since complexity is correlated with length in english, it is in the translations too. So the link between length and complexity in other languages is somewhat indirect.*

Thank you for this comment. Indeed, we do find reliable correlations with length across languages. However, given that all of this data is correlational, it is difficult to infer causality. Since we find that the complexity bias remains reliable after controlling for genetic relationships, this suggests that the systematicity in word length across languages is likely due to some latent variable rather than non-independence of word forms. Nevertheless, we cannot rule out this possibility.

4. More prosaically: the authors mention two additional experiments in Supporting Information, but I don't have access to that document. Presumably the SI should be reviewed too prior to acceptance.

We apologize. The Supplemental Information can be viewed here: https://mllewis.github.io/projects/RC/RCSI.html.

> *pg. 4: This feels like a quite different kind of non-arbitrariness to that discussed in the opening paragraphs - this is non-arbitrariness given a shared convention of how to refer, the first few paragraphs deal with the arbitrariness of what that convention actually is.*

Thanks for this comment. We agree that these two cases have not been traditionally discussed in the same way in the literature. However, we see them as related to each other: the suggestion is that pragmatic pressures constrain how we refer at the utterance level, and it is these same pragmatic pressures that shape word meaning mapping at the word level (via a complexity-length relationship), though over a much longer timescale.

> *pg. 7: Isn't it weird to assume that speakers don't care about getting their meaning across (if they don't, and are just trying to minimise production effort, why bother speaking at all?) or that hearers don't care about having to deal with overdescriptions?*

This is a great point, thank you. We assume this analysis reflects only speakers *non*-aligned utility, but that speakers also have shared utility associated with successfully communicating information. We have added a footnote clarifying this point.

> *pg. 9: Surely (according to the logic set out earlier) the hearer wants each word to map to a single meaning, but doesn't care if multiple words map to the same meaning, since they will nonetheless be able to interpret the intended utterance? On these terms it seems like the bias against synonymy would have to come from the speaker?/ Exactly - but synonymy would not the violate the hearer's principle, because the hearer doesn't care. / Similarly, under this account, isn't it the *hearer* bias that eliminates ambiguity?*

Thank you for noting this inconsistency in our logic – indeed it is the hearer's principle, not the speaker's principle, that is primarily responsible for the bias against synonymy. We've changed this paragraph to address this issue.

> *pg. 12: Your description of Horn's idea at the bottom of the previous page makes it sound like it's frequency that arbitrates between these two possibilities - isn't that fairly straightforward?*

Thank you for this point. We have edited this paragraph to more clearly reflect the fact that frequency arbitrates between these two possibilities.

> *pg. 13: This phrase is starting to bother me now, since it implies that you will be looking at change over time. In fact you are simply looking at language at two separate (synchronic) levels: use by individuals, and structure in the lexicon. I think the mapping between these is fascinating, which is what language evolution is about, but I think*

describing either of the synchronic perspectives as being "the language evolution timescale" is a little confusing.

Thanks for this comment.

pg. 16: How many? 10? 100? 1000? It would be useful for people to know, but also might play a role in evaluating whether I'd rather see the results reported for the non-repeated data instead of the data including repeats participants.

Across all experiments, an average of 3% of participants (n = 81 total) participated in a related study.

pg. 17: rated

Thank you, fixed.

pg. 18: Number of syllables should be 2, 4 or 6 in this figure and also in the equivalent panel in one of the later figures.

We appreciate you noting this error. The figure is actually correct, but the number of syllables reported in the text is wrong. The number of syllables was either 1, 3, and 5. This has now been corrected in the text.

pg. 18: What is "complexity rating ratio"? It's somewhat opaque, and why is 1/1 at 1.25?

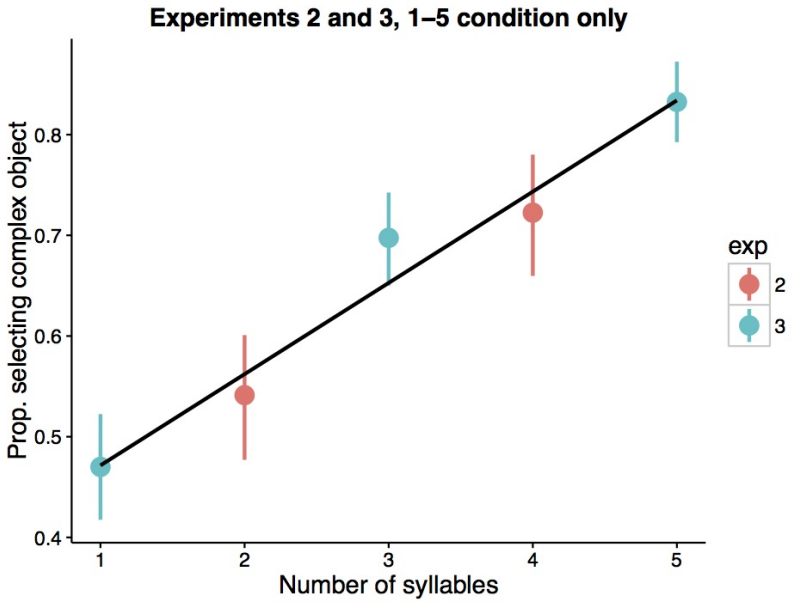
See comment to Review 2.

pg. 20: I was very confused by this until I read the next sentence, so maybe move that info earlier re. the last syllable ending in a consonant.

Done.

pg. 20: How do these results compare to Exp 2, the 2 and 4 syllable words? Is there any evidence, by comparing across these experiments, that the identity of the labels did in fact matter?

We find no evidence for an effect of label-type. Below is a plot showing the 1-5 condition from Experiment 2 (2 and 4 syllables) along with the data from Experiment 3 (1, 3, and 5 syllables). Together, the conditions show a linear trend with proportion complex object selections increasing as a function of the number of syllables. We did not add this analysis to the manuscript because it seemed distracting, but are open to doing so if the reviewer or editor think it would be helpful.



pg. 21: Are these the means of the complexity ratings in the two replicates? Wouldn't it be more informative to report the mean difference in ratings for objects or something?

This is a good suggestion, thank you. We have added the mean difference in ratings across the objects.

pg. 22: Or could it just be that there is more variability in complexity ratings for these real objects?

We actually observe identical variance in the complexity ratings for the artificial objects and the real objects (*SD* = .18), so this possibility seems unlikely.

pg. 25: Why not report mean length, rather than mean log length? In later analyses you report non-logged values, since those values are more interpretable I'd do so throughout.

Thank you, this is a helpful suggestion. We have now reported the means and standard deviations in terms of number of characters.

pg. 26: Why not measure the RT on the old/new decision? Is there a precedent for this study time measure? I can see that it's relevant, but I'd like a little more spelling out of the rationale / any precedents in the literature.

We didn't analyze decision times because we expected there might be baseline differences between old and new decision times. Study time provides a way to uniformly measure processing time across all stimuli. Other than work using search time as a measure of

complexity (Alvarez & Cavanagh, 2004), we do not know of any previous work using this measure.

> *pg. 27: presumably this is mean study time, not mean log study time?*

Thank you for noting this -- the times here were mean log study time, but the units were milliseconds (not seconds). However, following your suggestion for Experiment 7, we have changed these to reflect non-logged values (in seconds) to be more interpretable.

> *pg. 28: This correlation is a lot weaker - more variability in complexity ratings in the real objects?*

This is a great point – thank you! As you suggest, the item-level variance in study time is much larger for the geons, relative to the real objects (.28 vs. .18). We have added a note about this in Study 8b.

> *pg. 28: It's not actually clear to me what this means - why would it be informative? Or is it more of a sanity check?*

Thank you, yes, this is not directly relevant to our hypothesis. We have removed this analysis from Study 8 for clarity.

> *pg. 32: Yuk to these p values - should format them appropriately.*

Thank you, done.

> *pg. 32: So does this imply it holds *less* for concrete objects, like those studied in Exps 1-8? Is that a strength? A weakness? It seems worth a comment at least.*

Yes. See comment above.

> *pg. 33: *English* monomorphemic words right?*

We've added this clarification.

> *pg. 35: In this context it becomes clear that "language evolution timescale" isn't really the right term.*

Thanks for this comment. We understand why this term may be confusing, but we think that it allows us to connect with a body of literature that addresses the broader question of interest -- why this regularity might emerge in natural language. To explain this kind of data, it is necessary to appeal to process that happen on a very different timescale than we are able to observe in our laboratory experiments.

> *pg. 35: Actually you showed an effect on *study* time, which is not how (I think) people will interpret this term.*

Yes, thank you for this note. We agree that “study time” better reflects what we measured, and have changed the text to reflect this.