Dear Professor Kim,

We are pleased to resubmit our manuscript for review. We thank our reviewers for their many helpful comments and have revised the paper to address them. Among the more substantial changes, we have revised the General Discussion to highlight both the similarities between our work and previous work, as well as the ways in which our work makes a substantial contribution beyond what was previously known. Additionally, we have moved some of the previously included analyses (specifically, the region-by-region self-paced reading analyses) into an appendix, in order to better highlight the crucial analyses that remain in the main body of the paper.

Below we address the reviewer comments in detail.

***Reviewer #1:***

*This is a in large parts well written and (mostly) very clear paper addressing a long-standing issue in a reasonably convincing way. The methodological approach has novel aspects and appeals also through its relative simplicity (while still being thorough), which makes it more likely that it others will apply it to other phenomena. The authors anchor the paper in the relevant literature with a few exceptions that are easy to fix.*

Thank you for the positive comments! *I therefore am cautiously hopeful that this paper will eventually get published in Cognition. The theoretical framing will be a considerable contribution. However, there are a number of critical issues, some of which might change the results. I will send a separate PDF with minor points I noticed to the editor and authors (none of the points in there are critical).  
  
SIGNED  
Florian Jaeger  
  
Regarding the main criteria reviewers are asked to evaluate:  
  
1. The relation between the value and length of the paper  
Good, with one exception, where I'm recommending to move part of Experiment 2 into an appendix and to use the space to clarify the few places where the ms currently lacks clarity.  
  
2. The quality of language  
Mostly very clear, with a few exceptions.  
  
3. The logical soundness of the paper's arguments  
All sound.  
  
4. The adequacy of the descriptions of methods and analyses  
Mostly adequate, with only minor issues where I'm asking for some clarification and clearer admission of caveats to the employed method (and the extent to which they might affect the results). The are some potentially critical issues, which I summarize below.  
  
5. The abstract (Too short? Too long? Does it serve as an adequate summary?)  
Good  
  
6. The number of references (Too many? Complete?)  
Mostly complete (all exceptions involve very recent work). The additions I'm suggesting mostly serve the further integration with related literature.  
  
7. Tables and Figures (Adequate? Too many? Quality?)  
Figures and Tables were very informative, and I have only minor suggestions for improvement.  
  
  
  
CRITICAL ISSUES  
  
1) On p6 the authors address the issue that contextual effects could confound their results regarding attested and unattested binomials. Their dismiss this possibility " While this assumption would not always hold in a more naturalistic setting, our experimental materials (described in Section 3) will as much as possible avoid local contexts that would influence expression order …."  
However, Section 3 did in no way explain \*how\* effects of local context were avoided. In fact, without further evidence, I'm not convinced that they were. The authors could have opted to put attested and unattested binomials into the same context. This solution is also not ideal but a first step. Alternative, it should at least be assess/normed to what extent the preceding context (Experiment 2) and both the preceding and following context (Experiment 1) 'prime' either of the two constituents.  
  
Without further scrutiny, it is not clear whether the materials were biased in this way and, crucially, whether this bias would have differed between the attested and unattested binomials. I think the paper is publishable without such additional tests, but only if this caveat is clearly and prominently stated, rather than being dismissed.*

Thank you for raising this point. It is one we have thought carefully about, and we have taken this opportunity to address it more explicitly in the paper.

In Section 1.2.1, we have clarified what sort of local context we are primarily worried about (namely, one element being previously mentioned), and provided an estimate (based on other data collected since our initial submission) for how often such effects occur in naturally occurring text. In Section 3, we have clarified that we avoided exactly this sort of local context when constructing items.

More importantly, in Section 4.3 (Discussion of Experiment 1), we have added a discussion of why, even if local context biases slipped in, they would only add noise and would not confound our results.

Following up on that discussion, and recalling that outcomes are always coded in terms of alphabetical/non-alphabetical ordering preferences, we would like to point out that putting attested and novel binomials in the same context would not be a useful control because there is no reason to expect that the bias introduced by a given context for a given attested binomial (specifically, the bias towards alphabetical vs. non-alphabetical order) would be the same as the alphabetical/non-alphabetical order bias introduced by the same context for a given novel binomial (which necessarily contains different words).

Regarding an assessment of to what extent context primes either constituent, we in fact predict that preferred-first elements of binomials would score higher in most obvious tests of priming (e.g. cloze responses or plausibility ratings), even absent effects of the sentence context, because properties that contribute to elements being preferred first (e.g. frequency, cultural priority, concreteness) would likewise contribute to higher scores on any test of priming. This makes it difficult to disentangle effects of specific sentence contexts from other factors contributing to binomial ordering preferences. Fortunately, as noted above, any effects of local context that do exist within our materials will only add noise, not confound our results. *2) One critical issue regards the clarity of the exposition of Experiment 2. A good deal more work is required to assure comparability to previous work (beyond binomials) and to conduct and present analyses in a more standard way.  Below I list specific suggestions. One point that I consider absolutely essential for publication is that a region-by-region multivariate analysis should be conducted. Currently, the paper only presents region-by-region analyses that deviate from the analysis approach taken in the remainder of the paper. The critical analysis, however, is based on summing up the residuals over all critical regions. This is not standard. If the results of the multivariate analysis are not uniform across the sentence regions, this should be clearly stated.*

We thank you for this and other comments regarding the analyses of Experiment 2. A discussion of this issue in particular has been added to the Discussion of the region-by-region analysis (now in Appendix C). As we explain there, we do not expect the results of a multivariate analysis to be uniform across all regions. *3) There are a few places where caveats to the chosen analysis should be more clearly stated (the second author will recognize some of my points from previous reviews of other papers he was involved in).*

We will respond to the point by point suggestions below. *4) It seems that Experiment 1 and 2 come to somewhat different conclusions regarding the role of abstract knowledge for attested (reasonably frequent) binomials. In Experiment 1 (2AFC) abstract knowledge is found to affect attested and unattested binomials equally strongly (no interaction). However, in Experiment 2 (SPR) a significant (simple) effect is only found for unattested binomials. Did I miss something? I didn't find that this conflict was discussed in much detail. Perhaps what I propose under 2) above will resolve this issue, but that remains to be seen. In any case, I would expect more of a discussion of what is shared and what differs between the experiments, possible explanations of the differences, and what type of limits this puts on what can be concluded from both experiments taken together.*

This is a good point. Our primary claim in this paper rests upon a shared finding between the two experiments: that processing of attested expressions relies more heavily upon direct experience than abstract knowledge. However we cannot with certainty say whether abstract knowledge is differentially active between novel and attested expressions, due to the conflict pointed out here. We have added a paragraph to this effect in the discussion of Experiment 2 (Section 5.3).  
 *PRESENTATION  
As I said above, the paper is overall well written. The introduction is beautiful. The attached PDF contains some minor editing suggestions. Here I mention some more that stood out to me.*

Thank you again for your kind words! *1) p8, footnote 6 is hard to follow where it is currently. I recommend moving the footnote into section 2.1.*

This footnote has been moved as suggested. *2) p10, Experiment 1 analysis: this is one of the places where a few more sentences will make it a lot clearer how you're analyzing your data E.g., for " We nest the abstract knowledge predictor within type, allowing us to consider the effects of abstract knowledge on novel and attested binomials independently." — Elaborate. Looking at the result table it seems to me that you mean that you're using a simple effect analysis. (as it's known in the ANOVA world).*

Thank you for this suggestion. The descriptions of the Type and Abstract Knowledge predictors have been expanded as suggested (pages 11-12). *3) p12, towards the end of Experiment 1: " This pattern of results supports a theory wherein both abstract knowledge and direct experience play a role in processing." The experiment also finds that abstract knowledge affects judgments equally strong, regardless of whether binomials were attested or unattested. This should be stated again, as it further constraints the specific type of model that can account for these findings.*

Thank you for raising this point. We do not think it is justified to draw strong conclusions on the basis of this result, given that it is not replicated in Experiment 2. We have taken the opportunity to add a paragraph explaining our thoughts on this in the discussion of Experiment 1 (Section 4.3). *4) A few specific suggestions for the revision of Experiment 2:  
  
a) Please say more clearly (and earlier) what types of analyses will be conducted. For example, anticipate the multivariate analysis and say early on why the other analysis is even presented. Alternatively, I think the comparison to  Siyanova-Chanturia et al.'s (2011) could be moved into an appendix without loss of information. You could state in the main text that their results replicate and that you for the first time show the effect for novel items.*

This is a great suggestion: we have moved the region-by-region analysis to an appendix, which we think highlights the main results (of the multivariate analysis) as well as helps to clarify a number of the further presentational issues discussed below. *b) Please be clearer how \*exactly\* your results compare to Siyanova-Chanturia et al. For example, did they find effects on the same words? Such region-specific comparison is rather standard for SPR research.*

Unfortunately, Siyanova-Chanturia only report aggregate reading times, not word-by-word reading times, so we cannot make this comparison directly. We have added a footnote to this effect on page 29. (In fact, even if we had Siyanova-Chanturia’s data, they used eyetracking while we used self-paced reading, which might also lead to differences in where exactly effects appear.)

*c) It would help to make the analyses of the judgment data and the SPR as parallel as possible. Currently, the are rather different (e.g., separate analyses of attested and unattested data for the SPR data, but a joint analysis for the judgment data).*

We agree: we have tried from the beginning to make the multivariate analyses for the two datasets as analogous as possible. We believe that, having moved the region-by-region SPR analysis to an appendix so that it no longer intervenes between the two multivariate analyses, the parallels between the two multivariate analyses will now be much more apparent.  *d) Remove the random by-item slope for alphabetical order from the first analysis (you can add it later). It's rather confusing at this point and its purpose is not at all apparent.*

Because what was previously the “first analysis” is now in an appendix, the random by-item slope for alphabetical order will now be familiar to readers when they encounter it in this context. *e) State why trial was included in the analyses. You might want to refer to other works that have done so (e.g., Hofmeister et al., 2011-LCP; Fine et al 2013-PLOS One; Farmer et al 2013-CogSci) and the functional motivations these paper provide for this step.*

These reasons and citations have been added (Section 5.2.2). *f) Why was only the main effect but no interactional of trial included (again, compare to previous work) and justify or change the analyses.*

No interaction of trial with other variables was included because we made no predictions about the existence of such interactions. But for completeness’ sake, we have now run an analysis including such interactions, reported in footnote 14. *g) Motivate exclusion criteria with reference to earlier work.*

These citations have been added (Section 5.2.2). *h) Motivate why 400 (!) subjects were run in this study, compared to far fewer in Experiment 1. Previous web-based SPR studies have done with far fewer subjects. Was this decision made at the start of the experiment or after looking at the data at earlier stages. If the latter, pls adjust your family-wise Type I error rate (cf. Simmons et al, 2011).*

Some motivation has been added in footnote 12: in particular, the self-paced reading data is noisier than the forced-choice data, and each subject in Experiment 2 saw approximately half of the experimental items (compared to all of the items in Experiment 1).

Simmons et al. include correcting alpha levels in their list of “nonsolutions”, noting that it is generally ambiguous what an appropriate correction factor would be. We think this is certainly true in our case: Each subject in Experiment 2 saw approximately half of the experimental items due to a programming error that was not discovered until many subjects’ data was already collected, leading us to run more subjects than originally intended. Later, after having finished data collection, discussions with other researchers led us to change our method of analysis, so the final analyses were only performed on the full set of data. We do not believe there is any reasonable way to come up with an alpha correction factor that corresponds to these circumstances. *i) Cite some previous web-based SPR studies, especially some that replicate previously found effects (and ideally some that used webspr). We have a bunch of those, but I think Hal and Ted Gibson probably are the people to cite.*

According to Ted Gibson (personal communication; July 2015), there is no published work that explicitly replicates previously found effects using flexspr. Please point us to anything we’re missing! However, we have included citations to novel work using flexspr, as well as replications of known effects using other web-based SPR software (Section 5.1.2). *j) You included Trial in the analyses even for the first set of analyses (according to p15). This stands in direct contradiction to what you claim on p.16. If you used it, please state its effects clearly. As an aside: this further increased the impression that many different analyses were explored, raising questions about how you decided which ones to report  (cf. Simmons et al, 2011).*

Trial was included in all SPR analysis, but not in the forced choice analysis. (It would not be a sensible predictor in the forced choice analysis model, as is now explained in Footnote 10.) This is consistent with what we claimed in the previous version of this manuscript: the statement on page 16 was comparing the multivariate SPR analysis with the multivariate forced choice analysis, not with the region-by-region analysis. However, we recognize that the organization of the analyses in the original submission made this confusing. We believe that having moved the region-by-region SPR analysis to an appendix significantly clarifies this point. *k) Please motivate why the second set of SPR analyses (section 5.2.3) is using a slightly different residualization and outlier remove method that the first set of analyses. All of this just adds to the burden of the reader and creates questions about why you did what you did.  
  
Steps like these will help to reduce the impression that of excessive researchers' degrees of freedom. This is particularly critical since you do some of the things that Simmons et al (2011-PsychScience) explicitly warn against (such as including covariates without further motivation).*

Thank you for raising this issue: it made apparent to us that there was a small error in our description of our analysis methods in the original submission, which we have correct in this resubmission. In the region-by-region analyses, we mistakenly reported that outlier removal was done before residualization (in contrast to the multivariate analyses, in which we reported that residualization was done first). In reality, residualization was done first in both cases. Once again, thank you for bringing this to our attention!

The only other differences between how residualization and outlier removal are done in the two analyses is that outlier removal is done for the summed reading time in the multivariate analysis and for each region in the region-by-region analysis (i.e. outlier removal is done with respect to the unit of analysis—this is standard), and that outlier removal is done without regard for type in the multivariate analysis, but separately for each type in the word-by-word analysis. This second difference follows directly from the fact that we report one analysis (including both types) in the multivariate analysis, but separate analyses for each type in the word-by-word analyses. We have added a sentence clarifying the reason for this difference in the region-by-region analysis section (now in Appendix C). *5) p21, discussion of the gradient trade-off idea between specific exemplar storage and reliance on abstract knowledge: At the end of this review, I've included a number of references and perspectives that are directly relevant to this and other ideas in this part of the discussion. Additionally, the idea of this trade-off is discussed in quite a bit of detail for speech perception in Kleinschmidt and Jaeger (under review, p. 70-73, [http://goo.gl/tDo1kr](http://goo.gl/tDo1kr" \t "_blank))*We appreciate the suggested references. In addition to incorporating some of these suggestions into the Introduction, we have revised and expanded the General Discussion (particularly Sections 6.1 and 6.2) to include more detailed comparison of some of these perspective with our current work. *ANALYSIS  
Overall the analysis is presented very clearly, allowing a broad audience to follow. I have only a few requests:  
  
1) Please state that backward model selection is known to be anti-conservative (Harrell 2001), but an acceptable evil given your problem of having way too few cases for the number of constraints that are being considered.*

We appreciate this suggestion and have added this caveat, but with an alternate explanation: “Our goal is to develop the best possible model of binomial expression preferences that is nonetheless reasonably parsimonious (in particular, does not include those constraints that are clearly poor predictors), but it is not our goal to conclusively demonstrate that particular constraints are significant predictors of preferences: rather, our goal is to develop an effective predictive model that can be used to investigate the link between abstract knowledge of binomial ordering preferences and behavioral responses in offline and online processing tasks. We thus adopt relatively lenient criteria for inclusion of constraints in our final model.” (Section 2, page 8)

*2) Please state a) how many of the 379 types favored one outcome over the other, as this determines the suggested limits on how many predictors can reasonably included in the model (cf. references in Jaeger, 2011-Chapter) and b) acknowledge these recommendations (e.g., 15 times number of predictors < less common outcome). The starting model of 18 predictors clearly is above this limit (18 \* 15 > 379 / 2), likely exacerbating the anti-conservativity of the backward approach.*

We have considered this suggestion, but given that, as discussed above, the goal of this particular paper is not to conclusively determine the statistical significance of individual predictors of binomial ordering preferences, we do not think these are pertinent details.

*3) At least in an appendix, I'd like to see the constraints that were not selected by the backward model procedure (with as much of an explanation as provided for the constraints that were selected).*

Given that the data we use was originally collected and coded by Benor & Levy (2006), we believe that readers should refer to that paper for further details about the not-selected constraints, rather than including those in the current paper. (Even our explanations of the selected constraints are merely summaries of the constraint definitions provided by Benor & Levy.) *This isn't arguing for a change in method as I think it's the best that can be done with the available data (though more could have been collected!), but readers should know about these potential caveats.  
  
4) Experiment 1: a) pls provide a measure of the remaining collinearity in the model that is reported in Table 2.*

We have added a summary of the VIF for the model to the caption, and done the same for the model in Table 4.

*b) Please also state whether an interaction of relative and abstract-predicted frequencies would be significant if added to the model.*

It would not: this is now reported in footnote 10.

*Finally, c) I would love to see a footnote that states the the result of abstract-predicted and relative frequency (for attested binomials only) are significant if attested and unattested binomials are analyzed separately.*

This follows directly from how we parameterized the model shown in Table 2. Because we have nested the effect of abstract knowledge within binomial type, the effects as estimated in our model are equivalent to what one would find for each type if analyzed separately. Additionally, because relative frequency is zero for all novel binomials, the effect in our current model applies only to attested binomials (and would be identical if attested binomials were analyzed separately).

We believe that the implications of how we have parameterized the model are more clear now that we have expanded the description of the predictors used in the model (as suggested above), but we are happy to also include such a footnote in a future revision if the reviewer still thinks it necessary.

*5) OPTIONAL BUT RECOMMENDED: In Experiments 1 did repeated exposure to unexpected binomials have any effect on judgments (cf. Hofmeister et al., 2011-LCP)? Also, in Experiment 1 and 2, did the effects of other variables change of the course of the Experiment —e.g., as a function of repeated exposure to non-preferred binomial orders (cf. Fine et al., 2013)?*

Our experiments were not designed to optimally address these questions. However, we can look for such effects post-hoc by looking for interactions of trial order with other predictors (e.g. a positive interaction of trial order with abstract knowledge would indicate that abstract knowledge was becoming a stronger predictor of preferences over the course of the experiment, while a negative interaction would indicate the opposite.) Following the reviewer’s suggestion, we now report models including interactions of trial order with all other predictors in footnotes 10 and 14. Trial order did not interact significantly with any other predictors in either experiment. Therefore, we have no evidence from the work presented in this paper that repeated exposure to unexpected binomials affects their processing. However, as noted above, our experiment was not designed to optimally test for such effects, so we cannot conclusively state that they do not exist. *6) Why does the second set of analyses in Experiment 2 not proceed region by region? Did unexpected results emerge when this is done? As I've mentioned above, I very much recommend to simplify the presentation of Experiment 2: make it parallel to Experiment 1, but region-by-region and include the first set of analyses in an appendix (extended to include the necessary details that are so far missing). This will lead to a much more accessible paper with the same message that at the same will allow you to be sufficiently detailed where it is required.*

As discussed above, we have taken the reviewer’s excellent advice here and moved the region-by-region analyses to an appendix, addressing the question of multivariate region-by-region analyses in the Discussion of Appendix C. *7) It is not clear to me that in the terminology of Barr et al (2013), random slopes for trial are justified by the design. Admittedly, the Barr et al terminology isn't particularly clear on this issue, but it's not clear from the description of Experiment 2 that trial order was systematically manipulated within both items and subjects (random slopes for both are included).*

The reviewer is correct: Trial order in our experiment would fall under the category that Barr et al call “control predictors”. These predictors are not required to be included under their description of maximal random effect structures, nor are they excluded. In our design, trial order varies both within subjects and within items (but was not systematically manipulated); therefore, a model with random slopes for trial order for both subjects and items is identifiable, and so we chose to include it. Following the reviewer’s suggestion, we have modified our description of the random effects structures to clarify that the random slope for trial order is not a required part of the maximal structure as defined by Barr et al (Section 5.2.2 and Appendix C).

*INTEGRATION OF RELEVANT LITERATURE  
Regarding other literature, most of my comments (incl. specific references) are added to the attached PDF, but I would like highlight a few points here.  
  
1) at the end of Section 1 (and in some places before as well as in the discussion), there should be more of a mention of work that has asked the same question about the trade-off between abstract knowledge and lexical co-occurrence. There are a number of recent approaches in this vein, but a few stand out. This includes work by Baayen on colleagues on NDA; work by O'Donnell and Post & Gildea on morphological and syntactic parsing using tree-substitution grammars and similar approaches; and work by Bod and colleagues within an exemplar-based approach. Many of these recent works were included in a special issue on \*parsimony and redundancy\* in "Language and Speech" in 2012, edited by Wiechman et al. (see also their introduction).*

Thank you for the suggested literature, particularly the special issue of Language and Speech. We have added some of these references when first introducing exemplar-based models (p. 2, “Similar claims are made by exemplar-based computational models…”). Additionally, we have clarified the difference between these lines of previous work and our current work on page 3 (“Another outstanding question is how to empirically measure…”). *2) the \*absence\* of animacy (and certain other conceptual accessibility effects, such as 'proto-typicality') in binomial structures, compared to essentially all other constituent alternations has played a considerable role in research on production research, where it influenced theory development. This work starts with Bock and Warren 1985 and motivates recent papers by Branigan et al 2009 and Tanaka et al 2011 (though neither of these is on binomials). For a summary of this literature and further references, Jaeger and Norcliffe (2009) might be a good starting point for the authors. I don't want to de-rail the paper, but this should be mentioned somewhere, especially since one of the few psycholinguistic papers that \*does\* find an animacy effect on binomial order preferences \*is\* cited (McDonald et al).*Thank you for referring us to this line of work. We have added a reference to Bock and Warren, as well as a couple of related references motivated by Branigan and Jaeger & Norcliffe (Section 1.1.1).

We additionally thank Reviewer #1 for the more minor comments provided on a pdf. We have implemented the majority of the changes suggested therein. We respond here only to two comments that deserve further discussion.

*pdf page 5*

On the topic of whether our predictors for relative frequency and abstract knowledge are logged/in logits: Our predictors are in probability space (i.e. [0,1]), as stated in the paper (“a real number between 0 and 1”). The reason for this decision is that our predictors sometimes take on the values 0 and 1 (for relative frequency, in cases where one order is unattested), or values arbitrarily close to those (for abstract knowledge, in cases where the Iconic sequencing constraint is active, as its regression coefficient is effectively infinity), which correspond to values of positive or negative infinity in logit space. Using predictors in probability space allows us to avoid having to make arbitrary choices about how to tweak these infinite values in order to make computations tractable. However, if we do make these arbitrary choices in order to run our crucial analyses in logit space, the results are qualitatively similar.

*pdf page 13: Please say more about what the answers [to comprehension questions] DID depend on. E.g. how many of the questions focused on the NP that the binomial formed or was part of?*

Because our experimental materials encompassed a wide variety of sentence structures, there is no convenient way to summarize the focus of the comprehension questions. Instead, we have now included the comprehension questions along with the experimental materials in Appendix A so that readers can see for themselves.***Reviewer #2:***

*Review of "Abstract knowledge versus direct experience in the processing of binomial expressions."  
  
Summary:  This paper reports the outcome of a preference judgment task and a self-paced reading task for sentences containing conjoined NPs (e.g., the bishop and the seamstress/the seamstress and the bishop).  The main research questions related to the effects of prior exposure/frequency and abstract ordering constraints, phonologically-based preferences, chiefly.  The preference study showed that abstract ordering constraints have a greater effect on ordering preferences for un-attested items than for attested items.  Similar results were obtained in the self-paced reading experiment.  The paper concludes that both abstract knowledge and exposure affect ordering preferences.  
  
Evaluation:  There is probably some value in determining how abstract phonological properties influence choices of production in conjoined NPs.  There does not seem to be as much value in documenting another kind of frequency/familiarity effect.  In the limit, would anyone be surprised if "Jill and Jack went up the hill" were preferred less and read slower than "Jack and Jill?"  In the general case, many types of frequency effects, including structural frequency effects have been documented.  Given that prior published work has already addressed how phonological properties affect ordering preferences in conjoined NPs, it is not clear whether this paper adds enough new information to warrant publication in Cognition.  On its face, the paper represents an incremental contribution, rather than a potentially transformative one.*

We thank Reviewer #2 for their comments, but we disagree with this summary of our work and this evaluation of its potential impact in two important regards:

First, the abstract ordering constraints we consider are not primarily phonologically-based. We refer to these constraints as “semantic, phonological, and lexical” in both the abstract and the body of the paper. In fact, only two out of our seven constraints (No Final Stress and Length; Section 2) are phonologically based, and Table 1 demonstrates that these two in fact have some of the lowest weights in our model of binomial ordering preferences.

Second, and more importantly, our primary finding is not merely a frequency effect. Indeed, were that the case, our paper would simply be a replication of Siyanova-Chanturia et al. (2011), who, as we discuss in Section 1.1.1, have already demonstrated frequency effects for binomials. Rather, our paper presents a dissociation between when processing relies upon abstract constraints and when it relies upon frequency of one’s direct experience. These two knowledge sources provide competing explanations for previous findings (including those of Siyanova-Chanturia et al.), and the previous literature has not resolved which of these competing explanations is correct in which situations. We have highlighted this point in our revised General Discussion, which explicitly compares our approach to previous related work. We have also highlighted the competition between these explanations—and the fact that our work resolves it—in other places in the paper, including the title (“Abstract knowledge versus direct experience”) and the last paragraph of the Introduction (Section 1.0): “we will quantify the extent to which people’s processing of attested expressions is influenced by their frequency of direct experience with those specific expressions versus by the abstract linguistic knowledge that allows them to generate such expressions compositionally.”

As we discuss in Section 6.3, the work presented here also has substantial potential for future impact, as the novel findings in this paper lead directly to further predictions about language structure and change, some of which we have already seen borne out in our more recent work.

*There are also problems with the framing of the issues and basing strong theoretical conclusions on null results.  In terms of framing, the paper does not present a detailed processing account that rules out the kind of frequency effects that they purport to document.  Words and rules is claimed to rule out such effects, but that does not seem quite right.  Even in a system that generates regular past tense by rule and stores irregulars separately, there is no obvious reason why frequency effects would not be observed.  The paper itself does not present a compelling argument for the blanket claim that no version of words-and-rules processing could produce the observed pattern.*

We agree that some version of a words-and-rules theory could produce multi-word expression frequency effects (See Footnote 1); however, it would depend upon details of the parser that are not generally specified by those who propose this theory. A standard version of this theory does not predict multi-word frequency effects. In fact, the lack of frequency effects for regular past tense forms has been one of the primary arguments used by its proponents in favor of rule-based generation of regular forms in morphology. For example, Pinker (1991), which we cite, says:

If regular past tense forms can be computed on-line by concatenation of symbols for the stem and affix, they do not require prior storage of a past tense entry and thus need not be harder or stranger for low-frequency verbs than higher ones. Judgments by native English speakers of the naturalness of word forms bear this prediction out.

*In addition, the paper makes little or no attempt to describe other existing lexical processing frameworks that could produce frequency by (phonological) regularity effects in conjoined NPs.  Such effects are compatible with a wide variety of accounts.  Trace, for example, would straightforwardly  
predict the kinds of effects claimed to be compatible only with a Bayesian mechanism.*

We are confused about this discussion of “lexical processing frameworks” because our claim is not about lexical processing: it is about **sentence** processing. It is not clear to us how Trace, which is a model of single word recognition, would predict the effects we show here. We would welcome any clarification from the reviewer on this matter.

However, we agree that other connectionist models make predictions consistent with our experimental findings. Our revised General Discussion discusses a number of different frameworks that make convergent predictions, and hopefully makes clear that we do not believe our findings are only compatible with a Bayesian mechanism. *Other theoretical claims also go unsupported with references.  For example, storage efficiency is presented as a potential constraint that prevents conjoined NPs from being stored in long-term memory.  No references are provided in which the claim is advanced that LTM is too small to store a large number of attested conjoined NPs.  So this feels like a straw man.  MINERVA, which is not referenced, assumes more or less unlimited storage of individual memory traces.  Unsupported assertions of this type should be avoided.*

Following the reviewer’s suggestion, we have included another reference to Pinker (2000) to clarify that this assertion about storage efficiency is not a claim we are making ourselves, but rather a motivating concern for Pinker and others in developing the words and rules theory (Section 1.0, page 3). *The way the results are presented creates other problems.  The graphs do rather more to obscure than reveal.  The figure on page 11 is indecipherable.*

We have expanded the caption for Figure 2 (now on page 12), which we hope makes it easier to understand. This figure displays standard kernel density plot, which we believe are the most straightforward way to present this information. However, we would welcome constructive suggestions for alternate ways to visualize these data and analyses.

*Figures 3 and 8 would convey the information more clearly in a standard scatter-plot format.*

Figures 3 and 8 (now Figures 4 and 6) are, in fact, scatter plots, with additional dimensions of information overlaid as colors. We hope that the inclusion of our new Figure 1 (the standard scatter plot without color overlaid) makes the later figures easier to interpret.

*The tables are also not particularly illuminating (e.g., Table 4).*

The table in question (now Table C.7) presents information about the magnitude and statistical significance of our key regression predictor in each region for each subset of data. This table seems to us to be the most compact format for including this critical information. Once again, we would welcome constructive suggestions for alternate ways to present these results. *Aspects of the methods and data analysis are also troubling.  Experiment 2 was apparently run differently than it was intended.  Why not run the design that was intended?*

We are not entirely sure what the reviewer is referring to here. Experiment 2 was run as intended, except that subjects only saw 80 out of 168 items in each list of experimental items. As noted in Section 5.1.2, this was due to a programming error. Aside from requiring us to run more subjects, it does not affect our ability to draw inferences from the data.

*It is not standard (p. 15) to take a 2 x 2 design and analyze each half separately, without first reporting the interaction.*

Once again, we thank Reviewer #1 for his suggestion of moving the region-by-region SPR analyses into an appendix. The main body of the paper now presents only analyses in which the two binomial types are analyzed jointly. *Assuming lexical matching across the attested and un-attested conditions, there is no compelling reason to analyze residual reading times.  This technique carries assumptions and leaps of inference that are not articulated in the paper, but that could affect the interpretation of the results.  Minimally, the raw RT data (means, some index of variability within conditions) should be reported.  If the raw pattern does not map onto the residual data, that needs to be explained.*

There is no lexical matching across binomial types (see experimental materials in Appendix A), so we believe that residualizing reading times is appropriate. Following the reviewer’s suggestion, we have clarified that the purpose of residualizing is to account for influences of word length, and cited Ferreira & Clifton (1986), who popularized this method of residualizing (Section 5.2.2). Additionally, at the reviewer’s request, we have included the results of analyses run on the raw reading times in Appendix D. These results are qualitatively similar except that the abstract knowledge x binomial type interaction only reaches marginal significance, which we attribute to the presence of extra noise (due to variable word lengths) in the non-residualized data. *Finally, some of the main theoretical conclusions are based on null results.  This is a weak basis for inference.  For example, on page (12) "...the effect of abstract knowledge does not differ significantly between novel and attested expressions" could be restated as "...the data did not show an interaction between frequency and new vs. old expressions."  This absence of an effect in the data set could reflect the true absence of an effect in the world, but need not.  Similar logic applies to the information presented on page 17.*

We believe this comment stems from a misunderstanding. We do not draw any theoretical conclusions on the basis of the null result—specifically, the lack of a significant interaction between abstract knowledge and binomial type—in Experiment 1 (on what was in the previous draft page 12). We have added a paragraph clarifying this in the Discussion (Section 4.3): “Although the effect of abstract knowledge does not differ significantly across binomial types, we do not think it is justified to draw strong theoretical conclusions from this null result…”

In contrast, the conclusions drawn from Experiment 2 (previously on page 17) are **not** based on a null result. We find a significant interaction of abstract knowledge with binomial type: “In a likelihood ratio test comparing this model to a model with only an additive (non-nested) effect of abstract knowledge, we find a significant difference (χ2(1) = 4.24,p < 0.04); in other words, the effect of abstract knowledge differs significantly between novel and attested expressions, playing a significant role in online processing for novel expressions only.” (Section 5.2.2)

As Reviewer #1 also pointed out, the fact that we find a significant interaction in Experiment 1 but not Experiment 2 is worthy of future study, and we have added a paragraph discussing this in the Discussion of Experiment 2 (Section 5.3).

Most importantly, the crucial result that preferences for attested binomial expressions are more strongly determined by relative frequency than by abstract knowledge is consistent across both experiments. *Minor point:  Submitting the paper in type-set format rather than draft format makes there viewers' job harder than it needs to be.*

***Al Kim***

*R2's review:  I believe R2's comments are important.  R2's main point is that your manuscript, in its original form, did not present a major advance over the large body of work dealing with tradeoffs between memory of items vs. composition in language processing.  Empirically, there are many prior demonstrations of influences from combinatory frequency, which are similar in spirit to those you present here, even if they do not address binomial expressions specifically.  Theoretically, the idea of tradeoff between exemplar-specific memory and composition, which you emphasize in your conclusions, seems to have numerous precursors in the literature (both R1 and R2 mention this, in different ways) and this reduces the theoretical novelty of your manuscript.  You mention Rumelhart & McClelland (1986) late in the paper but seem not to appreciate how much this type of work already makes the points you make. You need to be much more clear about how your work provides a major  
conceptual advance.*

*R1 feels that the methodological advance is potentially valuable to the field, and I suspect that your paper will be strengthened by highlighting and clarifying exactly what the methodological advance is and how it could be applied to other phenomena.*

Thank you for this summary of the key issues. We hope that the revised General Discussion provides a clearer picture of why we see our work as a major conceptual advance. In particular, while we recognize and appreciate the contributions of the two literatures (empirical and theoretical) mentioned above, we see a major gap at the intersection of the two: namely, **an empirical demonstration of the trade-off between exemplar-specific memory and composition.**The previous empirical literature on multi-word expression frequency has not accounted for possible influences of abstract knowledge, while the previous literature on the memory-versus-composition trade-off has relied on computational models to demonstrate this trade-off. The work we present here combines behavioral experiments with use of modern corpora and multivariate statistics to fill this critical gap. *Minor comment: I agree with R2 that several of your figures are hard to read.  I think there has to be room for improvement either in the visual presentation or in the text you use to describe what's in the figures.  Please don't get hung up on whether the term "scatterplot"  is appropriate or not.*

We have expanded our figure captions and added a new figure (Figure 1), which we hope will make our figures easier to understand. We continue to welcome concrete suggestions for how to present these data more effectively.

Yours truly,

Emily Morgan (corresponding author)

Roger Levy