Dear Dr. Sanborn,

**Uploaded is a revised version of XLM-2020-1418 “Exemplar-Model Account of Categorization and Recognition When Training Instances Never Repeat”, which I am resubmitting with Mingjia Hu in consideration of publication in *Journal of Experimental Psychology: Learning, Memory, and Cognition*.**

**We thank you and the reviewers for your careful reading of the article and excellent suggestions for improving it. We worked hard to address the reviewer comments and believe that the article is much improved.**

**Using the critical comments in your editorial letter as a guide, we start by outlining the major revisions. We then address each of the specific points raised by the reviewers.**

Editor Action Letter

A key point for all reviewers was understanding why the current results and those of Homa et al. (2019), particularly the learning curve results, are so strikingly different. The paper speculates that the past results were an artifact of reusing the same random stimuli for each individual, and that these particular stimuli were (by random chance) not representative of the full set of stimuli. This is possible but could be made less speculative by asking the authors of the previous paper how the stimuli were randomized and perhaps (if the authors are willing to share) taking a look at the stimuli themselves.

**We contacted Drs. Homa and Blair to request their original stimuli. Unfortunately, the original stimuli are no longer available. Because their study was conducted long ago, the precise methods for how the stimuli were randomized are no longer available either. In any case, as we argue in our article, we think our “conceptual replication” approach -- in which a unique set of stimuli was randomly generated from the population of interest for each individual participant -- offers more generality than the approach used by Homa et al. We think our conceptual-replication method and contrasting findings deserve to be made known to the scientific community.**

Both Reviewers 2 and 3 commented that this study is not a precise replication of Homa et al. (2019), so it is very possible that a non-artifactual difference between the studies is driving the results. Reviewer 2 put forward a number of possibilities along these lines: 1) The training patterns in the current study are more distinctive than those in Homa et al. (2019), 2) The terminal levels of learning in the REP and NREP conditions were roughly equivalent in Homa et al. (2019) but are different in the current study, and 3) there may not have been the brief delay (3-4 mins) between training and transfer  that was in Homa et al. (2019).

**In introducing our new experiments and discussing the results, we are now much more careful to make clear that we are not conducting an exact replication of Homa et al. but rather are conducting a conceptual replication (e.g., pp. 24-25, 49-52). Regarding Reviewer 2’s suggestions for explaining some of the differences in our studies:**

**1) It is certainly possible that the training distortions (and new medium distortions) were more distinctive in our study than in Homa et al.’s. However, we used the classic Posner-Keele algorithm for generating the stimuli. Furthermore, broadly speaking, the overall qualitative pattern of results in the classification and recognition transfer phases are extremely similar across the studies. They are also regular and systematic, showing the exact pattern of classification and recognition transfer effects with distortion level that one would expect. Regarding the recognition transfer data, we conducted new extended analyses to trace the exemplar model predictions with variations in its parameter settings (pp. 44-47 and Figures 8 and 9). The exemplar model always predicts small differences in recognition rates for the old- vs. new-medium distortions in the NREP condition, which was one of the central points that we tried to make in our article. The small difference turned out to be significant in our study but not in Homa et al.’s (although tracking down this small predicted effect was not the main motivation for our experiments). Our new analyses indicate that the magnitude of predicted difference does indeed become minuscule as within-category dissimilarities among the medium distortions decreases (i.e., as the stimuli become less distinctive). So we agree with Reviewer 2 that one possibility is that the restricted set of medium distortions sampled by Homa et al. were less distinctive overall than the ones sampled in our study. (Our analyses also suggest that another possibility is that memory sensitivity, measured by the parameter “c” in the exemplar model, was lower for Homa et al.’s subjects than ours.) We now make these points explicitly on pp. 46-47 of our revision.**

**2) In our view, the finding that the levels of learning were roughly equivalent in Homa et al.’s study but different in ours cannot be taken as a “criticism” of our article. That was the VERY POINT of our study: our failure to replicate their learning curve results.**

**3) Regarding the delay issue, our method also had a brief delay between training and transfer (in which subjects took a brief break and also read instructions for the next phase). However, to address the possibility that the above-chance recognition discrimination for old vs. new-medium distortions in the NREP condition of our study may simply have reflected recency effects, we conducted the analysis suggested by Reviewer 2: a plot of hit rates for the old distortions as a function of the block of training in which they appeared (p. 31 and Figure 6). The plot is flat and there is no evidence of recency effects, so this factor does not appear to the main one driving the quantitative differences in recognition-discrimination results across Homa et al.’s study and ours.**

I think each of these explanations deserves careful consideration, and Reviewer 3 made a comment along these lines as well. Reviewer 3 also suggested including an exact replication in order to more carefully assess the differences, or rewriting the paper to reach more definitive conclusions about which is the more appropriate way to conduct the experiment. An exact replication that addressed any randomness issue would be welcome, but the paper at least needs to have a more comprehensive discussion about what the crucial difference is between the two sets of experiments which includes the possibility that it is not an artifact.

**As indicated in our previous paragraph, we gave each of the above issues careful consideration. Also as explained above, we are unable to conduct an exact replication because the original stimulus materials and precise methods are no longer available. However, as already indicated, we did our best to provide a more comprehensive discussion of the differences between the experiments; and to advance our argument that our way of conducting the study was, at the least, an extremely reasonable one, with certain clear advantages compared to Homa et al.’s method. To reiterate, the advantage is that we are pursuing effects that are general across the population of patterns generated from the classic Posner-Keele algorithm, rather than limiting the inquiry to a restricted sample of patterns with potentially highly idiosyncratic properties. (This is not to say that there may not also be advantages to conducting studies with small sets of sampled materials in this type of paradigm; e.g., such designs allows MDS analyses of the dot patterns and prediction of performance at the level of individual items rather than only types of items – see Shin & Nosofsky, 1992). We now include in our General Discussion a new paragraph that urges future research on the issues, with the possibility of someday reconstructing the types of stimulus conditions that produce Homa et al.s’ pattern of results (p. 52).**

Quite a few other comments concerned the modelling in this paper. Reviewers 2 and 3 both thought that the learning data should be modelled with the GCM. I understand the hesitancy to do so expressed in the footnote, but it could be done as a complementary analysis to show what the effect of including the training data would be.

**We followed this request of conducting a complementary analysis of the learning data (pp. 42-44 and Figure 7). In a nutshell, the analysis shows that even a rudimentary learning version of the exemplar model can capture our learning data while also being consistent with the transfer data.**

Reviewer 3 commented that it is important to fit the Mixed model from Homa et al. (2019) to the current data. While the formal Mixed model may be incomparable to the exemplar model fit here, I think the larger question is very important: could the difference in the learning curves observed in this paper by explained by using an exemplar model in the REP condition and prototype model in the NREP condition? Answering this question will help frame what the current results mean for the Mixed model as well for other models (e.g., SUSTAIN and the Rational Model of Categorization) that can behave either like a prototype or exemplar model.

**We added a paragraph to our General Discussion (p. 54) to acknowledge the possibility of mixed models. We are certainly not arguing AGAINST mixed models in this article, and we acknowledge the plausibility of such models in our revised General Discussion, including the possibility that alternative forms of category representation operate across the REP and NREP conditions. Instead, the central point of our article is that the qualitative patterns of results do not severely CHALLENGE the single-system exemplar model account, as Homa et al. had claimed in their article. For reasons also explained extensively in our paper, we do not think this is a good domain to carry out fine-grained quantitative model-fitting comparisons (with the likely need for complicated model-complexity-adjusted fit values), because we have no idea what are the psychological dimensions that define these stimuli.**

Relatedly, Reviewer 1 noted that the introduction stressed qualitative predictions, but that the results showed fitted predictions. This reviewer wanted to know how robust the predictions were across a range of parameters. Finally, this reviewer commented on the writing, suggesting that the “Brief Review of the Debate” section should be shortened to focus on the critical findings for this study.

**Following Reviewer 1, we have included an extended new section that addresses the extent to which the qualitative predictions are robust across a range of parameters (pp. 44-47 and Figures 8 and 9). (The short story is that the predictions are indeed reasonably robust. A pet peeve of mine is that, even if the model did require specialized parameter settings to reproduce the qualitative pattern of results, in my view that would STILL be an important contribution. When researchers publish provocative findings, it often takes place in a context that involved the search for specialized procedures or stimulus conditions that allowed the provocative phenomena to be observed. It is then not surprising that even a correct model would require specialized parameter settings to reproduce those provocative results.)**

**Second, we followed Reviewer 1’s suggestion of significantly shortening the introductory “Brief Review of the Debate” section (pp. 3-5).**

Reviewer #1: XLM-2020-1418  
  
- The "Brief Review of the Debate" could probably be shortened a bit (probably by half), focusing more on the Homa, Blair findings (getting to them more quickly).

**We significantly shortened the previous “Brief Review of the Debate” section (pp. 3-5).**

- It seemed odd to refer to "randomly sampling z-scores". Why not "a random sample from a normal distribution with mean zero and standard deviation one?" Then there are two terms that multiplicatively scale the standard deviation : "low", "medium", and "high" are fixed based on Homa, Blair et al. and "within" is a free parameter.

**We rewrote using the reviewer’s suggested terminology (pp. 14-15).**

- Footnote 2 might also note the absence of a response bias parameter.

**We now note the absence of a response bias parameter in the relevant footnote of our revision (Footnote 4).**

- Can Figure 2 reproduce Homa's error bars? And the titles might explicitly say "from Homa et al." (to not confuse "Experiment 1" and "Experiment 2" from Homa vs. "Experiment 1" and "Experiment 2" from the present article).

**We are unable to reproduce the error bars (and are unsure if the original individual-subject data files are still even available). Our main goal, however, is simply to reproduce the major qualitative pattern of results. We think it is going down the wrong path to attempt rigorous quantitative model comparison in this domain. Following your recommendation, we have modified the titles in Figure 2.**

- The article opened by stressing qualitative predictions but showed fitted predictions. Is there a sense of how robust the predictions are (across a range of parameters) as has been demonstrated for other exemplar model predictions?

**We have added an extended new section to the article that confirms that the qualitative pattern of predictions is reasonably robust across a range of parameters (pp. 44-47 and Figures 8 and 9). In showing that the model CAN reproduce the qualitative pattern of results with at least some sets of parameters (a significant goal in and of itself), it seems reasonable to us to illustrate with parameters that also give good quantitative fit.**  
  
- I would say "removed" or "eliminated" subjects from analysis (not "deleted").

**We now say “removed” instead of “deleted”.**

- Maybe I missed it, but was there a note about how many subjects (power) were in the original Homa, Blair, et al. paper (compared to this article)?

**We preferred not to get into the details of statistical power analyses in this article. It would be pretty complicated. Homa et al. ran a bunch of different conditions across multiple experiments (some two-category conditions, some three-category conditions, some conditions involving use of foils as new items, some conditions using prototypes as new items, some with 15-block learning phases and others with 20-block learning phases). Any given condition did not have nearly as many subjects as in our Experiment 1, but meta-analyses that aggregated data across their experiments may have had comparable numbers of subjects. Calculating power would also require some knowledge of the predicted effect sizes, but we really don’t exactly know those either.**

- It might be useful to have a bit more information about the modeling results reported by Homa, Blair et al. and how they differ from the modeling done in this article. From what I understand, Homa, Blair, et al. reported modeling results that differed from those in this article - but I wasn't quite sure how it differed (and if it differed, why it mattered). Was it merely a matter of presentation (as described in the G.D.) or was there something more substantively different? We you able to replicate their modeling results or was there something fundamentally wrong with what they did.  
  
**Part of the problem with directly addressing your question is that we found Homa et al.’s modeling section extremely difficult to follow and understand, and unfortunately some of their equations are not sensible. (E.g. their Equations 11b and 12b state that the probability of responding “old” to a new item is equal to one minus the probability of responding “old” to an old item.) In other cases, separate parameters are estimated for fitting the REP vs. NREP conditions with no explanation as to why (e.g., see Table 3 of their paper). Rather than launching into a detailed critique of their modeling section, we decided it would be better to take the high road by describing the gist of their modeling approach and pointing to conceptual limitations. (E.g., we noted that inserting their “average-similarity” parameters into their equations would not allow the model to be sensitive to individual-stimulus differences across different tokens of the main types of patterns nor to the differential “density” effects that we argue exist across the REP and NREP conditions.) However, to address your point, we have now commented a bit more about the nature of Homa et al.’s modeling in our expanded General Discussion by arguing that their proposed mixed model provides a post hoc account of the null effect of the REP vs. NREP manipulation on the learning-curve data (pp. 53-54). In general, an advantage of our modeling account is that the stimulus parameters (between, within, c, beta) are held fixed across REP and NREP, so the differential predictions arise from the processing machinery of the model itself rather than simply fitting different parameters to the different conditions. (The response criteria k are EXPECTED to vary in the manner that they do across REP and NREP, for reasons we explained in our article.)**

Reviewer #2:

However, there are numerous, serious problems with this study - both empirical and theoretical - that preclude my recommendation of acceptance of their manuscript. The problem is fundamental - empirically, Hu and Nosofsky replicate none of the critical results obtained by Homa et al. (2019).

**We conducted conceptual METHODOLOGICAL replications of Homa et al.’s studies. The fact that we do not replicate some of their critical RESULTS cannot be taken as a criticism of our study: That is in fact one of the central points of our new contribution. It is antithetical to the principles of science to argue against publication on grounds of failure to replicate empirical results, especially results that have not yet been well established.**

In the current manuscript, learning was significantly better in their repeat-condition, classification was virtually identical in the two conditions with no advantage for the new-medium transfer patterns following non-repeat learning, and recognition for training patterns was statistically significant between old and new patterns in the non-repeat condition. To repeat, none of these results were obtained by Homa et al. (2019). The question is why, and raises the issue of why this study by Hu and Nosofsky can be considered a replication.

**Our study can clearly be considered a conceptual METHODOLICAL replication of their studies. One of the major points of our study is that we are NOT replicating some of the claimed RESULTS. Failure to replicate results is not uncommon in our field, and the last thing one wants to do in advancing science is to suppress failures to replicate results.**

**For the record, the reviewer is also exaggerating some of the differences in the results across the studies: Homa et al. themselves did not observe a significant main effect of REP vs. NREP on classification transfer, nor did we. The reviewer’s focus here on the new medium-distortion result is post hoc: it is not the case that the mixed model predicted an effect for the medium patterns and not the others. Likewise, the central point emphasized in our article is that the exemplar model PREDICTS only a very small difference in recognition rates for the old vs. new medium distortions in the NREP condition. Homa et al observed zero effect, we observed a small-size one.**

Statistical power doesn't seem to be a critical distinction - Hu and Nosofsky ran 287 subjects in their two experiments; Homa et al conducted 4 experiments, and ran about 250 subjects overall.

**Our point was that if a model is predicting an extremely small-size effect, then it is weak evidence to suggest that the model is wrong if a significant difference is not observed and the sample sizes are small. Within Homa et al.’s individual conditions, sample size did tend to be small, and no meta-analysis that combined results across all the conditions was ever reported. In any case, the statistical-power issue is not a central point in our current submission.**

One obvious difference is that terminal levels of learning in the current study is considerably worse than found by Homa et al., at least for the non-repeat condition. As a consequence, the concern is how to interpret this 'replication' because, in short, nothing is replicated, and the exemplar-model fits are for results not obtained previously. Some potential explanations are suggested later.

**The reviewer is just repeating that we did not replicate the learning-curve RESULTS. Regarding the overall performance levels across the studies, extreme caution is needed in interpretation. As we argued throughout our article, Homa et al. limited consideration to a very restricted set of patterns (2 sets of 3 prototypes each generated from the Posner-Keele algorithm.) Perhaps the prototypes they generated in their restricted sample were easier to discriminate from one another than occurs on average when prototypes are generated randomly anew for each individual participant (our procedure). There is a huge number of other possibilities similar to the possibility just described.**

The modeling is also different in an important way. Hu and Nosofsky admit that their exemplar model cannot explain why learning should be little different between the repeat and non-repeat conditions, and their quantitative fit to the Homa et al. data is confined to the transfer data. In contrast, Homa et al. weighted equally the critical three data sets, learning, transfer-classification, and transfer-recognition. This seems appropriate, since many of the same parameters appear in each phase, e.g., sensitivity, within- and between-category similarity, how these values enter into predicted learning and transfer, etc. There were other differences between the two modeling approaches that are less critical here (e.g., Homa et al. suggested that parameters should reflect changing similarity relationships as learning progressed). Nonetheless, once Hu and Nosofsky ignore their inability to explain the learning data with their exemplar model, they argue that transfer classification and recognition reported by Homa et al. appear to be captured by their model ("…the quantitative fit to the data is outstanding", pg. 19). This approach is unsatisfying because Hu and Nosofsky provide no theoretical rationale for data-fitting using only the data sets consistent with their model. Regardless, Figure 2, which shows their "outstanding' quantitative fits, is so cramped that critical analysis is precluded (and none is provided in their document). For example, the statistically significant advantage of classifying the new-medium transfer patterns found by Homa et al. in the non-repeat condition goes unmentioned.

**We feel that our lines of argument and mode of analysis were sensible. Part 1 of our article was to point out that the general pattern of classification and recognition TRANSFER data reported by Homa et al. were fully in keeping with the exemplar model, a point not acknowledged by Homa et al. It is important to make these facts clear to the field. We clearly acknowledged that the model failed to predict the learning-curve results, and that was the central purpose of our conceptual replications, but we failed to replicate the learning-curve RESULTS. All of our data – both the patterns of learning and transfer -- were captured reasonably well by the exemplar model. (We also fail to understand the reviewer’s final gripe regarding the new-medium distortions in the Homa et al. experiment: examination of our Figure 2A shows that the exemplar model does predict a small quantitative advantage for the new-medium distortions in the NREP vs REP conditions.)**

Hu and Nosofsky then proceed to run two experiments that mirror the manipulations used in the Homa et al. study, which they fit (transfer-recognition & classification) with their model. However, none of the major results - learning, transfer-classification, and transfer-recognition - replicate what was reported by Homa et al. Instead, they found that learning statistically favored the repeat condition, that transfer-classification was equivalent between the two conditions, and recognition responses were greater for the old vs. new transfer patterns in the non-repeat condition. Again, none of these outcomes capture the original results that was the intent of this article, and that's, effectively, where we are.

**The reviewer is simply repeating his/her complaints about our observing different results. We do not feel that these points can be viewed as a “criticism” of our research. A major point of our article is that we are in fact observing different results. It violates the principles of scientific method to suppress the publication of results on grounds of failures to replicate findings.**

The issue that should be raised is the concept of a replication. We have two studies, one published and one under review, that arrive at quite different outcomes. Their modeling fits their obtained results but not the results of Homa et al. Failure to replicate is not rare. For example, the frequently-referenced Shin and Nosofsky (1992) in this manuscript reported substantial and significant forgetting of new and prototype patterns in their replication of other studies (Exp. 2), an outcome never obtained by anyone else. Then, as now, they fit their model to these results.

**This swipe at Shin and Nosofsky seems irrelevant to the current issues and is not helpful.**

Some additional comments are listed below.  
1. Subjects in the non-repeat condition were predicted, according to their model, to learn the training patterns more slowly than those in the repeat condition. In their two experiments, this outcome was obtained, with subjects in Hu and Nosofsky study asymptoting around .80. In contrast, the subjects in the repeat and non-repeat conditions in Homa et al. asymptoted around .95. This poorer learning may then be an insight into why Hu and Nosofsky obtain the (different) transfer results they report.

**To begin, the reviewer is a bit misleading in summarizing the learning outcomes. Homa et al. did observe roughly 95% correct by the end of learning in their Experiment 1, but that was a 20-block learning condition (involving the collection of only classification data at time of transfer). We conducted 15-block learning conditions, as did Homa et al. in their Experiments 2 and 3 (the experiments that examined recognition-transfer data). In their three-category conditions (the same ones we tested), learning finished at around 90% for both REP and NREP. In our own experiments, REP finishes at around 92% and NREP at around 82%.**

**More important, as we have emphasized repeatedly, there are numerous reasons why the overall performance levels would be expected to vary, due to Homa et al.’s use of their highly restricted samples of prototypes and training patterns. If I flip a set of dice only 2 times, the mean may not be very close to 7. If I flip the dice 90 times, the mean is likely to be very close to 7.**

- One possibility is that, unbeknownst to Hu and Nosofsky, their training patterns are more distinctive, more distorted, etc. than intended. This would also produce the recognition difference between old and new patterns in their non-repeat condition. The authors do not report the obtained range and mean value of the learning patterns used in the non-repeat condition.  
- Consistent with this is the large difference in the repeated condition, where the recognition responses of old and new is, estimating from their Figure 5, about .85 vs. .35; in Homa et al., this difference was about .90 vs. .60.

**In our revision, in the context of exploring variations on parameter settings in the model, we now acknowledge the possibility that our medium distortions were more distinctive, on average, than the ones used by Homa et al. (pp. 46-47). It is equally possible to argue things such as that Homa et al.s restricted sample of medium-level distortions had greater between-category dissimilarity than one would intend when generating patterns from the Posner-Keele algorithm.**

- A different concern is this. Generally, terminal levels of learning predict subsequent transfer. In the Homa et al. study, terminal levels of learning were roughly equivalent and very high for both the repeat and non-repeat conditions. Given this, these researchers found that classification accuracy, at least for the medium-level high distortions, was greater for the non-repeat condition. Had Hu and Nosofsky equated terminal levels of learning - using a common learning criterion rather than fixed blocks - then it is possible they also would have obtained superior transfer classification in the non-repeat condition.

**The issue of whether classification accuracy at time of transfer was better in the NREP vs. REP conditions is a minor one in the present context. Neither Homa et al. nor Hu and Nosofsky observed a significant main effect in one direction or the other, and the models predict very very small differences. The reviewer’s suggestion of extending number of training trials to equate terminal levels of learning across REP vs. NEP creates its own obvious confound (i.e., that number of training trials is not equated across conditions) and is clearly not what we want to do in examining the issues of central interest in the current studies.**

2. Hu and Nosofsky state that:  
"Furthermore, because the cloud of training examples is "dense" in the NREP condition, it is highly likely that each tested new medium distortion will be highly similar to at least some of the old training  
examples. A consequence is that there may be virtually no difference between the absolute  
summed-similarity signals associated with the old- and new-medium distortions in Homa et al.'s  
(2019) NREP condition." (pg. 12).  
  
- This is assumed, misleading, untested and probably wrong. It is a mistake to represent the learning patterns in a 2-dimensional display as shown in their Figure 1, which is consistent with their statement above. Mathematically, these patterns - nine vertices, each with an x & y value - can be viewed as a single point in 18-dimensional space. The surface area of this hyper-sphere is very large and the likelihood that any two, randomly-generated patterns, even of a medium-level distortion, would be highly similar to each other, is in fact rare. On the average, Homa et al. note that the average distance between any two medium-level distortions - show in their Table 1 - is greater than the distance of any pattern to its prototype. Would this be true in a psychological space? That's less clear, but when multidimensional scaling is done, the average distance between any two medium-level distortions is again greater (less similar) than to its prototype. The quotation above would be true if the stimuli were low dimensional stimuli, e.g., 2-dimensional patterns, such as linear segments in an xy-plane. If Hu and Nosofsky wish to maintain this assertion, then they need to provide some proof, e.g., generate a large sample of medium-level distortions, and compute how frequently these patterns are quite similar to each other.

**We followed the reviewer’s very constructive suggestion here of conducting these simulations. The results are reported in a new analysis on pp. 21-22 and 41 (see Table 2). The analysis confirms that the patterns in the NREP condition \*do\* tend to lie in a denser space than those in the REP condition (where density is defined as average distance to one’s nearest neighbor). (This analysis was done in our six-dimensional modeling space, not the 18-dimensional dot-coordinate one. Again, we reject the assumption that the psychological dimensions of the dot patterns align closely with the 18 physical coordinate locations of the individual dots.) Furthermore, our analyses confirm that, just as Homa et al. find using their MDS studies, the average distance between any two medium level distortions is in fact greater than the average distance of a medium distortion to its prototype. Finally, the purpose of our Figure-1 illustration is just to bring out intuitions. We emphasize even more strongly in our revision that that the two-dimensional schematic illustrations are NOT intended to capture the full set of similarity relations that exist among patterns in higher-dimensional spaces (p. 9).**

3. "Inspection of the best-fitting parameters (Table 1) reveals, as one would expect, that the between-category distance estimate greatly exceeds the within-category distance estimate." (pg. 22).  
- This should be true, but they report parameter values that differ 20-fold, which may say more about unrealistic parameter values needed to fit restricted data than what is actually true. In the paper by Homa et al., the similarity of two within-category patterns was, according to their parameters, about 2 to 1. Regardless, a converging operation of this assertion might be worthwhile. For example, would rated similarity between within- and between-category patterns reveal support for this? My guess is no.

**We take the blame here for not carefully communicating the meaning of these parameter estimates in our initial submission. The magnitude of “between” cannot be directly compared to “within” because “within” scales the distances “low”, ”medium”, and “high” (see pp. 14-15 and Equation 1). The more important consideration is the consequence of these parameter settings on the average distance relations among the key dot-distortion types in the simulated space. As we bring out in our revision (pp. 21-22 and Table 2), the pattern of average distance relations produced by the parameters in fact accords nicely with the one described above by the reviewer.**

4. Hu and Nosofsky find that recognition of old was greater than new in the non-repeat condition. An interesting analysis of recognition by Homa et al. explored whether any distinction between old and new occurred across blocks, finding none.  
- An insight into why Hu and Nosofsky obtained significantly greater recognition of old would be to look across learning blocks like Homa et al. did. Was the effect uniform across learning blocks or was there a primacy and/or recency effect? Also, Homa et al. introduced a brief delay, e.g., 3-4 minutes, between the last learning trial and transfer. If transfer between learning and transfer was immediate, then differences in recognition between the two studies might be reflected in a positive recency effect in Hu and Nosofsky.

**We thank the reviewer for this very constructive suggestion. We conducted the suggested analysis (p. 31 and Figure 6) and find no evidence for primacy or recency effects of the form described by the reviewer.**

5. Small item, but I assume the subscripts for x (e.g., xim) on page 16 are wrong

**We thank the reviewer for pointing up this notational issue. We agree that providing a more detailed notation is helpful for explaining the precise pattern-generation procedure in our simulations (p. 15 and Equation 1). However, we then transition to our more simplified original notation in describing the mechanics of the model, because the more detailed notation starts to become unwieldly.**

-------------------------------------------------------------------------------------------------------------------------  
So, what to do when replication fails? Not uncommon in the literature, including this area. But my recommendation is that this study be rejected - it adds little to the literature as is, unless they wish to consider running additional studies involving novel manipulations and predictions that address similar theoretical issues - or seriously revise to address why this rep  
Color  
  
**We strongly disagree with the reviewer that the study adds little to the literature.**

Reviewer #3:

**Below, we address Reviewer 3’s points in reverse order, because our more major revisions were in response to his/her latter points.**

6) It appears that the behavioral experiments are conceptual rather than precise replications of Homa et al. which creates ambiguity. I would like the authors to be more clear than they are regarding what exactly is the same and what is different (and why it was made to differ). The difficulty of interpretation for the reader arises because one lab did the experiment one way and found one result and the other lab did it another way and found another result. To more satisfactorily resolve the issue, it would be ideal to include an exact replication or reach a more definitive conclusion as to why the experiment is appropriately conducted one way or the other.  
  
**As explained earlier, we are unable to conduct an exact replication because the original stimulus materials and precise methods for creating them are no longer available. In any case, as explained earlier, we think the conceptual-replication approach that we pursued offers more generality than the one used by Homa et al. in which only a small sample of prototypes was used for all subjects. We addressed the reviewer’s point by revamping the organization of our article and making clear from the outset of our new Experiment 1 (pp. 24-25) that we are conducting conceptual replications, not exact replications, of the earlier study. We clarify exactly what is different and why, and we advance arguments, both in the introductory sections of the new Experiments (pp. 24-25) and in our General Discussion (pp. 49-53), why the methodological approach we used is reasonable and has certain advantages compared to Homa et al.’s approach.**

5) While I understand that the GCM is designed to explain end-state performance, the authors' aversion to modeling the learning data strikes me as odd and somewhat problematic (including the footnote regarding the widely used ALCOVE model). This may be connected to comment #3 above. Given the repeated focus throughout the paper on learning speed as the critical issue, it would certainly make the paper stronger to include evaluation of exemplar and mixed model fits to the learning data.  
  
**We followed the reviewer’s request of also using a version of the GCM to model the learning data from our new experiments. This new modeling section is found on pp. 42-44 and Appendix A of the revision (see Figure 7).**

4) I was surprised that the authors did not include the "Mixed" account from the Homa et al paper in their simulation studies. Given the differences in how the authors pursued the modeling (as well as the behavioral experiments), it seems that the paper would be made considerably stronger if it reported how the Mixed account performs relative to the exemplar model.  
  
**For reasons we explained in the introduction to our article, we do not think that this domain is a good one in which to conduct rigorous quantitative modeling comparisons between competing models, especially classes of models with differing functional complexities (for which complicated model-complexity adjustments in fit would be required). In our revised General Discussion (p. 54), we acknowledge that a variety of mixed models could undoubtedly also provide good fits our data. As brought out in our revised General Discussion, the central message of our paper is not to argue AGAINST mixed models; instead, it is to point up the fact that a single-system exemplar model accounts well for the qualitative pattern of results observed in this REP vs. NREP dot-pattern paradigm, thereby addressing the challenges that Homa et al. set forth in their article.**

3) I have some concerns about the treatment of free parameters. I would like to see the authors strive to minimize individual tailoring of free parameter settings across tasks, conditions, tests, etc. I believe it goes without saying that the explanatory power of the model is highest when it does not rely on different parameterizations to fit different data points within an experiment or series of experiments. Therefore, the authors may chose to report fits that include such tailoring (with appropriate justifications and caveats), but I think it is critical to report simulation results with the model restricted to a single set of parameter values (except possibly where it is clearly and explainably the modelers' design intent that a parameter functionally address a difference between experimental conditions). Additional points: the usage and rationale for the within and between free parameters seems non-standard and could be better explained; the differing treatment of free parameters between the initial and later simulations is somewhat jarring.

**We found this line of criticism surprising because we DID strive to hold parameters fixed across those conditions and tests where it made sense to do so – and provided clear explanations for cases in which the parameters varied across conditions. Consider our fits to our classification and recognition transfer data across our Experiments 1 and 2 in the REP and NREP conditions (see Table 3). The parameters *between*, *within*, and *c* are all held fixed across classification and recognition \*and\* across the REP and NREP conditions. The criterion parameters k(REP) and k(NREP) are relevant only for recognition, and we explained clearly why the recognition criterion settings are expected to vary across the REP and NREP conditions. The only parameter that we allowed to vary across classification and recognition was “gamma” and we provided an extensive discussion of why it made sense to do so. With regard to the use of the parameters “between” and “within”, we hope that our improved terminology and notation on pp. 14-15 will clarify how these parameters are used to generate the simulated patterns in our modeling. In addition, we have now included a more extended discussion of the consequences of those parameter settings for average distance relations among key pattern types in our model space (pp. 21-22 and Table 2). The resulting average distance relations are sensible and in qualitative accord with previous distance estimates obtained using MDS methods.**

2) Assuming I followed correctly, the GCM is applied with summing of similarity over every single stimulus experienced rather than over the set of unique exemplars. Is this a choice in the application of GCM? My sense has been that in a typical category learning experiment, the GCM (and certainly ALCOVE) operate by comparing a test item to each unique exemplar that appeared during training, not to each instance of each exemplar. Perhaps in traditional experiments with repeated blocks of the same training items this issue does not come up? This seems like a key concern in the present work, so I would like to see it treated more comprehensively (particularly since the exemplar view is often criticized for assuming that each and every example is independently stored and accessed for every category decision-- let alone assuming this to be true for every exposure to every example). I do see the advantage that storing every exposure gets around the problem of having to decide on each learning trial whether the current item is new (and therefore requires addition to the category representation) and speaks to issues of frequency of item exposure and possibly recognition (admittedly, I haven't looked recently at those papers).

**We stated clearly in our initial submission of this article that the GCM is summing similarities to the individual instances presented on each individual trial (e.g., see pp. 17-18 of revision). This assumption has been true of the GCM since its inception, and we address the reviewer’s question simply by providing an historical reference as a form of documentation (Nosofsky, 1988b). If the memory strength of the repeated examples did not change with their repetitions, then the model would have little basis for predicting performance differences across the REP and NREP conditions of the present experiments. (The reviewer is correct that for certain types of experimental designs, the issue is not relevant and so is often not discussed explicitly; in this design the issue is highly relevant and it was discussed explicitly.)**

1) The authors discuss the difficulty of finding a non-compromising way to model findings based on dot-pattern stimuli. It is my understanding that Nosofsky's group recently developed a technique based on deep learning to address just this type of challenge. I would very much like to see this technique applied in the present project because there is something distinctly unsatisfying about the simulated version of the experiment employed by the authors, and it is not particularly clear that this approach is superior to that used in the Homa et al. paper.  
  
**We have added a new footnote (Footnote 2) to our article that mentions the Sanders and Nosofsky (2018, 2020) proposal for deep-learning approaches to multidimensional scaling. Although the approach is a promising one, Sanders and Nosofsky have clearly acknowledged that it still has major limitations and a great deal of future research would be needed before we could apply it in a rigorous manner to a study like the present one. We are disappointed that the reviewer found there to be something unsatisfying about our current simulation-based modeling approach – in our view, it is conceptually sensible and produced very good accounts of the reported data with relatively few free parameters. Also, as we have documented in the revision, the estimated parameters varied in sensible ways. Please also see our response to Reviewer 1’s final point.**