Thank you very much for submitting your manuscript "Exemplar-Model Account of Categorization and Recognition When Training Instances Never Repeat" for review and consideration for publication in the *Journal of Experimental Psychology: Learning, Memory, and Cognition*. I sincerely appreciate the opportunity to review the manuscript. I have now received the reviews of your manuscript and am able to make an editorial decision at this time.

The reviewers and I all thought that this paper addresses an important topic. Reviewers 1 and 3 also thought that the work was well executed and convincing, but Reviewer 2 was less positive about the paper.

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.

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).

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.

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.

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.

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.

I would like to offer you the opportunity to revise the manuscript. If you decide to revise the work, please include a cover letter that details your response to each point raised by the Editor and the reviewers (appended below) in this letter. I would likely send a revision out for further comment to the current reviewers and potentially to new ones.  
  
To submit a revision, go to https://www.editorialmanager.com/xlm/ and log in as an Author. You will see a menu item called Submission Needing Revision. You will find your submission record there.  
  
Sincerely,  
Adam Sanborn, Ph.D.  
Associate Editor, *Journal of Experimental Psychology: Learning, Memory, and Cognition*  
  
Reviewers' comments:  
  
Reviewer #1: XLM-2020-1418  
  
This article presents a convincing theoretical and empirical counterpoint to recent work by Homa, Blair, and colleagues. Using a variant of the classic dot pattern category learning paradigm, they had reported classification and recognition memory results that seemed to falsify predictions of exemplar-based models of categorization. Their key manipulation was whether training instances (dot patterns shown during training) repeated (REP: 5 training instances in each of 3 categories that repeated over all training blocks) or were non-repeating (NREP: every training instance of each of the 3 categories was novel). While Homa, Blair and colleagues suggested that their classification and recognition results at transfer were inconsistent with exemplar models, the simulation results reported in this article showed very good qualitative and quantitative accounts by an exemplar model (both of the original Homa, Blair results and in two reported replications). Homa, Blair, and colleagues also reported no difference in the speed of category learning for REP vs. NREP, which is acknowledge by the present work as being inconsistent with exemplar model predictions, the present article reports two high-powered replications of Homa, Blair, and colleagues that shows instead faster learned for REP than NREP, as predicted by exemplar models.  
  
This article is important and well done and should be accepted for publication. I have only a few suggestion:  
  
- 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).  
  
- 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.  
  
- Footnote 2 might also note the absence of a response bias parameter.  
  
- 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).  
- 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?  
  
- I would say "removed" or "eliminated" subjects from analysis (not "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)?  
  
- 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.  
  
  
  
  
  
Reviewer #2: Hu and Nosofsky ('Exemplar-Model Account of Categorization and Recognition when Training Instances Never Repeat') address a recent study by Homa et al. (2019) that produced a number of unexpected results involving category learning, subsequent classification, and recognition. Specifically, Homa et al. found that category learning was virtually unaffected regardless of whether the training patterns were repeated or not across training blocks. This occurred even though subsequent recognition of the training instances in the non-repeat condition was virtually at chance. On the classification transfer test, the classification transfer of new patterns was no worse, compared to the standard 'repeat' condition, and, in fact, was significantly higher for those instances that were medium-level distortions of the category prototype. None of these results are consistent with an exemplar model of classification; rather, a prototype model was favored, primarily because an exemplar model must predict more rapid learning when patterns repeat as well as enhanced transfer-classification, since the absence of recognition of training patterns in the non-repeat condition must predict poorer classification on a transfer test. Subsequent quantitative modeling supported the interpretation that prototype mechanisms were involved in the non-repeat condition and exemplar influences likely determined performance for their repeat condition. To simplify, Homa et al. argue that, when a few patterns are repeated in learning, subjects memorize the individual instances; when the learning floods the subject with non-repeating instances, the subject abstracts the central tendency (the prototype). The goal of the study by Hu and Nosofsky is to re-address these claims, by replicating the learning and transfer in their lab, as well as applying an exemplar model to these findings.  
Dr. Nosofsky is a prominent researcher in this field who has repeatedly argued that many of the findings taken as support for prototype theory, e.g., forgetting rates of new and prototype patterns, the shape of generalization gradients, etc., can be explained by an exemplar interpretation. A similar approach is taken here. Two experiments were conducted, and quantitative modeling was explored.  
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). 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. 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. 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 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.  
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 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.  
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.  
- 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.  
- 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.  
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.  
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.  
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.  
5. Small item, but I assume the subscripts for x (e.g., xim) on page 16 are wrong  
-------------------------------------------------------------------------------------------------------------------------  
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  
  
  
Reviewer #3:  
  
This paper represents a strong contribution consisting of a combination of model simulations and behavioral experiments to refute a recent challenge to exemplar models (Homa et al.). Broadly put, this is important work and makes a fairly convincing case. However, in my view there are substantial shortcomings that ought to be addressed.  
  
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.  
  
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).  
  
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.  
  
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.  
  
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.  
  
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.