Journal of Experimental Psychology: Learning, Memory, and Cognition

Coherent category training enhances generalization in prototype-based categories

--Manuscript Draft--

Manuscript Number:	XLM-2021-1967R1
Article Type:	Article
Full Title:	Coherent category training enhances generalization in prototype-based categories
Abstract:	A major question for the study of learning and memory is how to tailor learning experiences to promote knowledge that generalizes to new situations. In two experiments, we used category learning as a representative domain to test two factors thought to influence acquisition of conceptual knowledge: the number of training examples (set size) and the similarity of training examples to the category average (set coherence). Across participants, size and coherence of category training sets were varied in a fully-crossed design. After training, participants demonstrated the breadth of their category knowledge by categorizing novel examples varying in their distance from the category center. Results showed better generalization following more coherent training sets, even when categorizing items furthest from the category center. Training set size had limited effects on performance. We also tested the types of representations underlying categorization decisions by fitting formal prototype and exemplar models. Prototype models posit abstract category representations based on the category's central tendency, whereas exemplar models posit that categories are represented by individual category members. In Experiment 1, low coherence training led to fewer participants relying on prototype representations, except when training length was extended. In Experiment 2, low coherence training led to chance performance and no clear representational strategy for nearly half of the participants. The results indicate that highlighting commonalities among exemplars during training facilitates learning and generalization and may also affect the types of concept representations that individuals form.
Manuscript Classifications:	160: Categorization; 220: Computational models; 230: Concepts; 610: Knowledge representation; 970: Recognition
Keywords:	categorization, concepts, generalization, prototype model, exemplar model, category learning
Additional Information:	
Question	**Response**
I have stated whether data are available and, if so, where to access them. In both the Author Note and at the end of the Method section, I have either specified where the data will be available or noted the legal or ethical reasons for not doing so.	Yes
In both the Author Note and at the end of the Method section, I have stated whether study materials are available and, if so, where to access them.	Yes
For submissions with quantitative or simulation analytic methods, in both the Author Note and at the end of the Method section, I have stated whether the study analysis code is available, and, if so, where to access it.	No
I have stated whether or not any work was	No

preregistered and, if so, where to access the preregistration. If any aspect of the study is preregistered, I have included the registry link in the Method section and the Author Note. For example: This study’s design was preregistered; see [STABLE LINK OR DOI]; This study’s design and hypotheses were preregistered; see [STABLE LINK OR DOI]; or This study was not preregistered.	
Please disclose any changes in authorship (inclusions/exclusions/order of authors) in the revised version of your manuscript. Have any changes been made?	No, there aren't any changes
If your paper is accepted, do you have any disclosures that would need to be listed on the Full Disclosure of Interests form? This includes any interests or activities that might be seen as influencing the research (e.g., financial interests in a test or procedure, funding by pharmaceutical companies for research).	No, I have no disclosures to report.

XLM-2021-1967
Coherent category training enhances generalization and increases reliance on prototype representations
Journal of Experimental Psychology: Learning, Memory, and Cognition

We thank the Editor and Reviewers for their thoughtful and constructive comments and are excited for the opportunity to resubmit. Below, we provide a point-by-point response to each Editor and Reviewer comment. Editor and Reviewer text is in black font throughout the response, and our text is in blue font. Reprinted text from the manuscript is in italics. In the manuscript, added and significantly revised text is denoted with blue font.

We hope that the Editor and Reviewers will share our enthusiasm about the resulting paper and find it acceptable for publication in JEP:LMC.

E1. First, I was concerned about how novel these findings are. The manuscript does a nice job of delineating what is new here and what was demonstrated in your recent paper (Bowman & Zeithamova, 2020), which I appreciate. However, to my eyes the new findings were mostly about set size, and previous research had shown that set size might not matter. This experiment seems to confirm that with a better study. I don’t mean to say that’s unimportant, but at the same time, the novelty is constrained. Thus, it would help if there was a little bit more about why the novel findings here are important. Reviewer 1 agrees with this comment, but I also want to acknowledge that Reviewers 2 and 3, who see clear importance in the findings, might disagree.

We appreciate the opportunity to expand on the novelty and importance of the current work. We have revised the introduction and discussion sections to provide additional information about both the real-world significance of the research and the contribution to the field of category learning.

Namely, there has been substantial debate across multiple domains (e.g., computer science, linguistics, child development) about what kinds of training regimes best promote concept learning. While most agree that the coherence of category training sets is important, researchers do not necessarily agree on the direction of that effect. For example, a recent paper stated that, ‘A consistent finding in the literature is that increasing variability during learning allows people to gain a better understanding of a category, improving classification with novel exemplars’ (Doyle & Hourihan, 2016, pg. 1197). Our prior paper showed the exact opposite: lower variability in training examples led to better learning and generalization (Bowman & Zeithamova, 2020). But a single study is not enough to revise our understanding, of the effects of training variability given the prevalent view that high variability in training examples leads to better generalization. In machine learning, large training sets are often needed to train algorithms that generalize well with the assumption that it is the variability afforded by those training sets that is important (Zhou et al., 2017). But it seems that ‘large’ is often used to mean both ‘containing many examples’ and ‘spans the entire category space.’ Thus, teasing apart how the number of training examples versus where the examples come from in the category space has potential real-world implications beyond cognitive psychology and theories of categorization.

We also think that testing for an interaction between training set size and coherence has the potential to address novel theoretical issues in the domain of category learning. It has been suggested that exemplar representations may be particularly suited to smaller training sets with fewer dimensions and a less coherent structure, whereas prototype representations may be well suited to larger training sets with more dimensions and a more coherent structure (Minda and

Smith, 2001). While Minda and Smith (2001) demonstrated each of these effects individually (set size, dimensionality, coherence) they did not explicitly test the hypothesis that these factors might combine to robustly produce either exemplar or prototype representations. One study that did explicitly test the set size x set coherence interaction only tested it in terms of categorization performance, not in terms of the relative fit of the prototype versus exemplar models (Homa and Vosburgh, 1976). Furthermore, they trained to criterion, which would mean substantially more training in the low coherence condition. Thus, it remains unclear whether small, low coherence training sets are particularly likely to promote exemplar representations, and large, high coherence training sets are particularly likely to promote prototype representations.

Another important point is that, to our knowledge, no one (including us) had previously controlled for both the number of item repetitions AND the number of training trials when measuring effects of set size. Here, we disentangle these variables and once again show that generalization success is strongly affected by training set coherence while the effects of set size are more limited. Thus, although our current study may seem like an incremental contribution compared to our prior study, having the opportunity to test for an interaction while controlling for the length of training and the number of repetitions is a worthy new addition as the presence or absence of an interaction between set size and set coherence has theoretical implications. In the revised manuscript, we expand on rationale and significance of our study in the Introduction and Discussion to better convey the novelty and contributions of our study.

Introduction (pg. 3):

"*The ability to form new conceptual knowledge is a key function of memory, allowing individuals to organize past experiences and apply them efficiently to new situations. How to tailor learning to best promote acquisition of new conceptual knowledge has been a question of considerable interest not only in cognitive psychology (Hahn et al., 2005; Mervis & Pani, 1980; Williams & Lombrozo, 2010), but also in domains like child development (Ogren & Sandhofer, 2021; Perry et al., 2010; Twomey et al., 2013), linguistics (Bulgarelli & Weiss, 2019; Onnis et al., 2004; Plante et al., 2014), and computer science (Hart, 1968; Hernandez-Garcia & König, 2020; Roiger & Cornell, 1996). The answer to this question has practical implications for how instructors select training examples to maximize the generalizability of learning.*"

Introduction (pg. 4):

"*Nonetheless, the idea that high-variability training is especially beneficial to generalization remains widespread, with recent work claiming that, 'A consistent finding in the literature is that increasing variability during learning allows people to gain a better understanding of a category, improving classification with novel exemplars.' (Doyle & Hourihan, 2016, pg. 1197) and a recent review similarly stating the benefits of increased training variability for promoting generalization of knowledge (Raviv et al., 2022).*"

Introduction (pg. 5-6):

"*Like set coherence, the effect of training set size on subsequent generalization has been somewhat equivocal across past studies, particularly regarding whether set size affects reliance on prototype vs. exemplar representations. As suggested by Minda and Smith (2001), studies that have found better fit of the exemplar model than the prototype model tended to show only a small number of examples during training (Blair & Homa, 2003; Lamberts, 1994; Medin et al., 1978). This exemplar model advantage may arise due to the relative ease of encoding only a few items into memory compared to when the number of training examples is large. Adding low* *coherence to small set sizes may further promote exemplar representations because having less overlap among items makes it easier to individuate them in memory. Others have shown better categorization accuracy for larger training sets (Goldman & Homa, 1977; Homa et al., 1973, 1981), accompanied by better fit of the prototype model (Minda & Smith, 2001). High coherence within these large sets may be particularly well suited to promoting prototype representations because they offer ample signal to derive the category average, while exemplar-based encoding may be hindered by the difficulty of distinctively encoding many similar examples. Since either prototype or exemplar representations could support successful learning and generalization, both large, high coherence and small, low coherence training may facilitate broad category knowledge.*

Despite this theoretical motivation for an interaction between training set size and coherence, empirical evidence is lacking. Studies using continuous-dimension stimuli have shown that large training set sizes do not always lead to reliance on prototype representations. While the prototype model predicts a linear decision bound between categories, participants trained on many unique exemplars sampled from two bivariate normal distributions can learn and adopt an optimal, non-linear category boundary, indicating they must be representing the variability of each category in addition to its center (Ashby & Gott, 1988; Ashby & Maddox, 1992; McKinley & Nosofsky, 1995). Studies using binary dimension stimuli also have not shown clear support for any one effect of training set size. Minda and Smith (2001) tested both set size and set coherence, but did so across separate experiments and thus could not measure their combined effect. In a prior study (Bowman & Zeithamova, 2020), we showed that higher training set coherence tended to lead to greater reliance on prototype representations without strong effects of training set size. However, set size and coherence were somewhat correlated, which meant that we could have missed a benefit of low coherence training if it were present only for some training set sizes. Homa and Vosburgh (1976) directly tested the interaction in terms of generalization abilities and found evidence against the hypothesis that large, high coherence training and small, low coherence training promote generalization. Instead, they showed a generalization advantage for coherent training only when the set size was small (3 items). For large set sizes (6 or 9 items), lower coherence resulted in better generalization. However, this study did not assess the representations underlying generalization judgments, and subjects were trained to criterion prior the generalization test. Training to criterion rather than keeping the training fixed across conditions likely resulted in much longer training for low coherence than high coherence sets, and potentially also longer training for larger set sizes than the small set size."

Introduction (pg. 7):

"*In the present study, we conducted two experiments that aimed to disentangle the effects of training set coherence, set size, and amount of training in order to better understand the conditions that facilitate category learning and generalization, as well as the types of representations that such training conditions foster.*"

Discussion (pg. 42-43):

"*Notably, the current study was novel in that it included two versions of the large set size condition, one that matched the small set size in terms of the total number of the training trials (providing half of the repetitions per item) and one that matched the small set size in terms of the number of repetitions per item (thus providing double the number of the total training trials). Studies sometimes control for one or the other when testing the effect of set size, but not both. Here, we found that set size had little effect on generalization success regardless of whether small and large training sets were matched in terms of exposures to individual examples or total* *number of learning trials. Thus, we found no evidence that large training sets would provide a generalization advantage, not even in the condition where using a large training set meant doubling the length of training.*

In addition to assessing overall effects of set size and coherence, we were interested in the extent to which these factors interacted. More specifically, we were interested in whether set size would modulate the effect of coherence when the amount of training is controlled, as was observed previously in a study that included training to a criterion (Homa & Vosburgh, 1976) and as predicted theoretically (Minda & Smith, 2001). In contrast to the Homa and Vosburgh (1976) study, we found a strong generalization advantage for high coherence sets that was consistent across set sizes. This finding is a key addition to our prior paper, which was not a fully crossed design and did not allow us to assess the potential for an interaction between set coherence and set size (Bowman & Zeithamova, 2020). Overall, results from categorization performance suggest that training set coherence is a much stronger influence on category learning and generalization than training set size. This finding also has practical importance for those compiling training examples to teach new categories: there may be flexibility in the number of training examples needed for learners to generalize well if those training examples are coherent around the category center."

Conclusion (pg. 45-46):

"*How structure of our experience affects learning, memory and generalization is of interest to many research domains and disciplines. Here, we revisited the question how variability among category examples during training affects learning and generalization, and how training variability interacts with other aspects of training, especially the number of unique examples and their repetition. Contrary to a common assumption that high-variability training promotes generalization, we found robust benefits of training on more coherent, less variable exemplars. While coherence was a strong driver of better category learning and generalization, we found relatively limited differences in categorization performance between small and large training sets, regardless of whether they were matched for the total number of training trials or the number of exposures to individual training items. Furthermore, set size did not seem to moderate the strong effect of set coherence on generalization accuracy. Training coherence and set size jointly affected the types of category representations people formed, but the effects on representations were not consistent across experiments and may be more nuanced than predicted based on current theoretical considerations. Together, these results add to theoretical and empirical work indicating that training that highlights commonalities among exemplars promotes formation and generalization of conceptual knowledge.*"

E2. Reviewer 2 (Rob Nosofsky) points out a confound in the manipulation of coherence. I was convinced by his argument. I'm not sure whether there is a way to deal with this issue without running additional participants. One way or another, I hope you can find a convincing way to address this issue. (And by the way, if this purported confound was present in your previously published work, I don't think that necessarily makes it less problematic.)

As we wrote to the editor, Dr. Kornell, shortly after receiving the reviews, we agreed with Dr. Nosofsky's (Reviewer 2) concern about the category structure. Dr. Nosofsky was correct that some stimulus features (especially in small, low coherence training sets) were non-diagnostic of category membership. The confound was a result of using 8-dimensional stimuli where it is not mathematically possible in some conditions to have both all stimuli at distance 3 AND have each feature equally predictive of category membership. As a side note, we verified that no such confound was present in any of the category structures from our prior paper. In fact, set size and coherence were not orthogonal to each other in our prior study primarily because of the limited number of set size/coherence combinations that are possible when keeping each stimulus dimension equally predictive. We have noted this issue with the Experiment 1 training sets and the reason behind it in the revised manuscript (pg. 27):

"*The strong role of training set coherence is consistent with our prior findings (Bowman & Zeithamova, 2020), and provides new evidence that the magnitude of the coherent training performance benefit does not depend on using a large training set. However, one issue with the training set structure used in this experiment is that the low coherence training sets included features that were non-diagnostic of category membership (see Appendix). This was especially true of the small, low coherence structure in which half of the features in the training set (4 features) were non-diagnostic compared to only one feature in the large, low coherence training set. This confound occurred because it is not mathematically possible with 8-dimensional stimuli to have 4 items that each share 5 features with their category prototype and also have each feature be equally predictive of category membership.*"

We also agreed with Dr. Kornell and Dr. Nosofsky that additional data collection was needed to ensure that our effects could not be explained by this additional difference in training set structures between the high and low coherence conditions. We have collected a second experiment using training structures that avoid this confound (training set structures depicted in tables below). The findings from the newly added Experiment 2 replicated the finding of better category learning and generalization following high coherence compared to low coherence training. Effects of training set size on learning and generalization were minimal regardless of whether we equated for the number of item repetitions or number of training trials. The modelling findings were more nuanced since many participants in the low coherence training groups performed quite poorly and were not fit at above-chance levels by either the prototype or exemplar model. The revised manuscript provides a detailed discussion of each of these findings. We thank the Reviewers and Editor for their patience in awaiting this revision while we collected and analyzed the new data and believe that the addition of the second experiment resulted in a more robust and convincing paper.

New training structures

The new Experiment 2 used 10-dimensional stimuli. For high coherence categories, all training items shared 8 of 10 features with their category prototype. For low coherence categories, all training items shared 6 of 10 features with their category prototype. Small training sets included 5 training examples per category (10 total). Large training sets included 10 training examples per category (20 total). All tables depict example training sets with 1111111111 as the category A prototype. These tables have also been added to the revised manuscript Appendix.

Table 7 - Small, high coherence category structure

Category	Dimension									
	1	2	3	4	5	6	7	8	9	10
A	1	0	1	1	1	0	1	1	1	1
A	0	1	1	1	1	1	1	1	0	1
A	1	1	1	0	1	1	1	1	1	0
A	1	1	1	1	1	1	0	0	1	1
A	1	1	0	1	0	1	1	1	1	1
B	0	1	0	0	0	1	0	0	0	0
B	1	0	0	0	0	0	0	0	1	0
B	0	0	0	1	0	0	0	0	0	1
B	0	0	0	0	0	0	1	1	0	0
B	0	0	1	0	1	0	0	0	0	0

Table 8 - Small, low coherence category structure

Category	Dimension									
	1	2	3	4	5	6	7	8	9	10
A	1	0	0	1	0	1	0	1	1	1
A	1	1	1	0	1	0	0	1	0	1
A	0	1	1	1	0	1	1	0	0	1
A	0	1	0	0	1	1	1	1	1	0
A	1	0	1	1	1	0	1	0	1	0
B	0	1	1	0	1	0	1	0	0	0
B	0	0	0	1	0	1	1	0	1	0
B	1	0	0	0	1	0	0	1	1	0
B	1	0	1	1	0	0	0	0	0	1
B	0	1	0	0	0	1	0	1	0	1

Table 9 - Large, high coherence category structure

Category	Dimension									
	1	2	3	4	5	6	7	8	9	10
A	0	0	1	1	1	1	1	1	1	1
A	1	1	0	0	1	1	1	1	1	1
A	1	1	1	1	0	0	1	1	1	1
A	1	1	1	1	1	1	0	0	1	1
A	1	1	1	1	1	1	1	1	0	0
A	0	1	1	1	1	1	1	1	0	1
A	1	0	1	0	1	1	1	1	1	1
A	1	1	1	1	0	1	0	1	1	1
A	1	1	1	1	1	0	1	0	1	1
A	1	1	0	1	1	1	1	1	1	0
B	1	1	0	0	0	0	0	0	0	0
B	0	0	1	1	0	0	0	0	0	0
B	0	0	0	0	1	1	0	0	0	0
B	0	0	0	0	0	0	1	1	0	0
B	0	0	0	0	0	0	0	0	1	1
B	1	0	0	0	0	0	0	0	1	0
B	0	1	0	1	0	0	0	0	0	0
B	0	0	0	0	1	0	1	0	0	0
B	0	0	0	0	0	1	0	1	0	0
B	0	0	1	0	0	0	0	0	0	1

Table 10 - Large, low coherence category structure

Category	Dimension									
	1	2	3	4	5	6	7	8	9	10
A	0	0	0	0	1	1	1	1	1	1
A	1	1	0	0	0	0	1	1	1	1
A	1	1	1	1	0	0	0	0	1	1
A	1	1	1	1	1	1	0	0	0	0
A	0	0	1	1	1	1	1	1	0	0
A	0	1	0	1	0	1	0	1	1	1
A	1	0	1	0	1	0	1	0	1	1
A	0	1	1	0	1	1	1	0	1	0
A	1	1	0	1	0	1	0	1	0	1
A	1	0	1	1	1	0	1	1	0	0

B	1	1	1	1	0	0	0	0	0	0
B	0	0	1	1	1	1	0	0	0	0
B	0	0	0	0	1	1	1	1	0	0
B	0	0	0	0	0	0	1	1	1	1
B	1	1	0	0	0	0	0	0	1	1
B	1	0	1	0	1	0	1	0	0	0
B	0	1	0	1	0	1	0	1	0	0
B	1	0	0	1	0	0	0	1	0	1
B	0	0	1	0	1	0	1	0	1	0
B	0	1	0	0	0	1	0	0	1	1

E3. I also want to emphasize Reviewer 2's point 5. Like the reviewer, I also speculated about how these results would have come out with different stimuli. To add to the point Reviewer 2 made, I think it would be appropriate to qualify the conclusions of this study. It seems like it would be appropriate to state that the results do not apply to learning categories in general, but rather to learning this set, or type, of categories, and the results could be different with other kinds of categories. (This might seem like a generic criticism that could apply to any study, but I don't mean it that way. I think it's particularly relevant here given some of the points raised by Reviewer 2.)

We agree that these findings may not generalize to other types of stimuli and category structures. We have changed the title to make it more salient that we focus on prototype-based categories and added the following discussion to the revised manuscript (pg. 41-42):

"*The advantage of learning from high coherence training sets appears robust across the experiments presented here and our prior work with similar category structures (Bowman & Zeithamova, 2020). However, it may not be equally suitable for all types of stimuli or category structures. While we showed this effect at two different levels of high stimulus dimensionality (8 and 10 dimensional stimuli), the stimuli were always binary dimension cartoon animals, and the category structure was always prototype-based. Categories not centered around a single prototype (e.g., multiple prototypes, disjunctive or rule-plus-exception category structures) may require a very different sampling of training exemplars to be learned robustly. For example, using natural categories like rocks and birds, others have shown that training sets that span the full category space or offer a wider variety of examples may be particularly good for promoting generalization (Nosofsky et al., 2019; Wahlheim et al., 2012). Similarly, information-integration categories with non-linear boundaries require extensive training with large number of varied exemplars to provide an opportunity to learn not only the central tendency of a category but the entire distribution range (Ashby & Gott, 1988; McKinley & Nosofsky, 1995). Thus, one possibility is that high coherence training is uniquely well suited to prototype-based categories because it leads participants toward the underlying category structure. For other types of categories, especially those that are not linearly separable, greater variability and/or extended training may be quite important. Interestingly, even rule-plus-exception category structures may be easier to learn when exposure to exceptions is delayed (Heffernan et al., 2021), suggesting that training that highlights commonalities among category members not only makes it easier to learn rules or prototype-based categories but also memorize the exceptions.*"

Reviewers' comments:

Reviewer #1:
This manuscript reports one study investigating the joint impact that set size (number of training examples) and set coherence (similarity among training examples) have on generalization of categories. The authors find better generalization following training with more similar items, regardless of number of items. Moreover, the authors fit different models to participants generalization that suggest that participants are more likely to form prototypes with more coherent (highly similar) training sets.

R1.1. The manuscript has multiple strengths: well-designed experiment and tests, inclusion of computational modeling to understand the process and provide some insight into how participants represent categories and how it changes with different set coherence and size. I should also add that the paper is overall well-written, clear, and to the point and generally enjoyable to read. Despite all these, I am not positive towards acceptance of the manuscript. I have some more technical issues with the work (detailed below), but in large part my not so positive evaluation is because I'm not sure it matters or adds much to our current understanding of category learning and generalization. The main issue is that it is not clear how understanding the interaction between set size and set coherence informs our understanding of category learning or how generalization takes place. I fail to see a strong hypothesis for looking at this interaction. As I understand it, the argument seems to be that there is inconsistent results for the impact of set coherence and for the impact of set size, so let's look at the interaction. Don't get me wrong, I think looking at interactions is key, but not all interactions are theoretically relevant and there is a risk for spurious results that do not actually inform our understanding of the phenomenon and might actually add noise instead of signal. Add to this the fact that to manipulate set size one is indirectly manipulating coherence (fewer items, same number of trials means a more coherent training set), and that to equate for study length one needs to increase the number of repetitions, now having a repetition effect added to the mix, I'm just not sure this interaction happens in the real world.

We thank the Reviewer for noting the strengths of the manuscript and we appreciate their concerns about the methodological approach and the impact of the work. We have provided detailed responses to the technical questions below. Most importantly, we have revised the manuscript to better highlight what we see as the novelty and importance of investigating how the size and coherence of training sets affects category learning and generalization.

Novel contributions of the present work

As noted in our response to the Editor, we revised the Introduction and Discussion to better highlight the empirical and theoretical contributions of the current study. First, we would like to note the importance of the current empirical data, irrespective of how they may be theoretically interpreted. Early cognitive psychology and computer science research highlighted the importance of category training with varied examples, and it was thus thought that the best way to obtain category knowledge was known – a lot of examples, especially atypical, to improve generalization. We have encountered recent papers and had many conversations with colleagues who view the benefits of high-variance training as a well-established fact. For example, Doyle and Hourihan (2016) note: ‘A consistent finding in the literature is that increasing variability during learning allows people to gain a better understanding of a category, improving classification with novel exemplars’ (pg. 1197). However, a closer look indicates that empirical evidence for this claim is not nearly so clear-cut. Notably, most of this research utilized training-to-criterion approach, with less coherent training sets requiring vastly more training than more coherent training sets (e.g., Posner & Keele, 1968; Homa & Vosburgh, 1976). Further, research on the effects of variability across examples often conflates the number of examples with the heterogeneity among those examples (see Raviv, Lupyan, and Green, 2022 for a review of this issue). Here, and in our prior study (Bowman & Zeithamova, 2020), we thus revisited this question and empirically tested whether atypical examples improve generalization when the amount of training is controlled. Contrary to the commonly accepted assumption that widely varied examples provide optimal training, we find a clear advantage to learning categories from training examples typical of the category. Thus, the empirical data in themselves are important to disseminate, demonstrating an empirical phenomenon that was not previously widely known.

We revised the Introduction and Discussion to better highlight this contribution (pp. 3, 4, 39-41 and reprinted above in response to E1).

The current study also has theoretical relevance. The Reviewer noted that we failed to make the motivating hypothesis clear. To clarify, our study was directly motivated by the hypothesis postulated by Minda & Smith (2001) based on the joint consideration of prototype and exemplar theories of categorization. Specifically, current theories propose that the advantage for coherent training may be especially pronounced when set size is large and exemplar strategy ineffective. Model fitting results in our prior study indicated that training with typical exemplars promoted prototype abstraction, and the strong prototype representation then supported strong generalization, even to items that were less typical. This work suggested that the benefit of coherent training resulted from facilitated prototype abstraction. However, both prototype and exemplar strategy may support successful generalization. Thus, it is possible that under varied circumstances, coherence and prototype abstraction may be more or less critical for success, depending on the ease of using the alternative exemplar strategy. We thus wanted to test the theoretically motivated prediction that coherence and the number of exemplars would interact (Minda & Smith, 2001). Our prior study was not optimized to test for an interaction, as coherence and set size were not orthogonalized (the more coherent sets also tended to be somewhat larger). The current study orthogonalized set size and coherence in a fully crossed factorial design, making it possible to robustly test the prediction of current theories of knowledge representation that coherence and set size interact.

We revised the Introduction to better convey our theoretical motivations (pg. 5-6):

"*The history of debate and mixed findings around the direction of the coherence effect leaves open questions about how coherence interacts with other aspects of training that differ across studies. One candidate factor is the number of training examples (set size). Like set coherence, the effect of training set size on subsequent generalization has been somewhat equivocal across past studies, particularly regarding whether set size affects reliance on prototype vs. exemplar representations. As suggested by Minda and Smith (2001), studies that have found better fit of the exemplar model than the prototype model tended to show only a small number of examples during training (Blair & Homa, 2003; Lamberts, 1994; Medin et al., 1978). This exemplar model advantage may arise due to the relative ease of encoding only a few items into memory compared to when the number of training examples is large. Adding low coherence to small set sizes may further promote exemplar representations because having less overlap among items makes it easier to individuate them in memory. Others have shown better categorization accuracy for larger training sets (Goldman & Homa, 1977; Homa et al., 1973, 1981), accompanied by better fit of the prototype model (Minda & Smith, 2001). High coherence within these large sets may be particularly well suited to promoting prototype representations because they offer ample signal to derive the category average, while* *exemplar-based encoding may be hindered by the difficulty of distinctively encoding many similar examples. Since either prototype or exemplar representations could support successful learning and generalization, both large, high coherence and small, low coherence training may facilitate broad category knowledge.*

Despite this theoretical motivation for an interaction between training set size and coherence, empirical evidence is lacking. Studies using continuous-dimension stimuli have shown that large training set sizes do not always lead to reliance on prototype representations. While the prototype model predicts a linear decision bound between categories, participants trained on many unique exemplars sampled from two bivariate normal distributions can learn and adopt an optimal, non-linear category boundary, indicating they must be representing the variability of each category in addition to its center (Ashby & Gott, 1988; Ashby & Maddox, 1992; McKinley & Nosofsky, 1995). Studies using binary dimension stimuli also have not shown clear support for any one effect of training set size. Minda and Smith (2001) tested both set size and set coherence, but did so across separate experiments and thus could not measure their combined effect. In a prior study (Bowman & Zeithamova, 2020), we showed that higher training set coherence tended to lead to greater reliance on prototype representations without strong effects of training set size. However, set size and coherence were somewhat correlated, which meant that we could have missed a benefit of low coherence training if it were present only for some training set sizes. Homa and Vosburgh (1976) directly tested the interaction in terms of generalization abilities and found evidence against the hypothesis that large, high coherence training and small, low coherence training promote generalization. Instead, they showed a generalization advantage for coherent training only when the set size was small (3 items). For large set sizes (6 or 9 items), lower coherence resulted in better generalization. However, this study did not assess the representations underlying generalization judgments, and subjects were trained to criterion prior the generalization test. Training to criterion rather than keeping the training fixed across conditions likely resulted in much longer training for low coherence than high coherence sets, and potentially also longer training for larger set sizes than the small set size.”

Lastly, we also think that testing the training set size x set coherence interaction has real-world significance because teachers (broadly defined) often have the opportunity to tailor their examples and way of presenting examples to facilitate category learning. For example, computer scientists compile examples for training machine learning algorithms. Radiologists-in-training learn to diagnose injuries and diseases from example images. Children learn many categories through examples selected by their teachers or authors of children’s books. Better understanding how to design category training sets to best promote category learning can help guide teachers in optimizing training materials. We have added these real-world implications to the revised manuscript (pg. 3):

“The ability to form new conceptual knowledge is a key function of memory, allowing individuals to organize past experiences and apply them efficiently to new situations. How to tailor learning to best promote acquisition of new conceptual knowledge has been a question of considerable interest not only in cognitive psychology (Hahn et al., 2005; Mervis & Pani, 1980; Williams & Lombrozo, 2010), but also in domains like child development (Ogren & Sandhofer, 2021; Perry et al., 2010; Twomey et al., 2013), linguistics (Bulgarelli & Weiss, 2019; Onnis et al., 2004; Plante et al., 2014), and computer science (Hart, 1968; Hernandez-Garcia & König, 2020; Roiger & Cornell, 1996; Zhou et al., 2017). The answer to this question has practical implications for how instructors select training examples to maximize the generalizability of learning.”

Dissociability of set size and set coherence

The Reviewer was also concerned that manipulating the set size indirectly manipulated set coherence. The term “coherence” could have a number of different meanings (Raviv, Lupyan, and Green, 2022). We therefore want to clarify that, as we define the terms, set size and set coherence *were* orthogonal in the present experiments. We defined coherence in terms of the number of features training items shared with their category prototype, consistent with our prior work (Bowman & Zeithamova, 2020). In Experiment 1, training items in high coherence sets all shared 6 of 8 features with their category prototype, and those in low coherence sets all shared 5 of 8 features with their category prototype. That was true whether the training set contained 4 items (small sets) or 8 items (large sets). Experiment 2 was similar: high coherence sets were made up of training items sharing 8 of 10 features with their category prototype and low coherence sets were made up of training items sharing 6 of 10 features with the category prototype. These two levels of set coherence were the same for training sets that contained 5 items (small sets) and those that contained 10 items (large sets). Thus, manipulating set size did *not* indirectly manipulate coherence.

We appreciate the Reviewer reminding us of this terminology issue and the myriad ways that researchers have manipulated the variability among training examples. In the revised manuscript, we acknowledge other types of manipulations like the ratio of within-category differences to between-category differences (i.e., structural ratio as suggested by R3.1). While there were differences in structural ratio between small and large sets at a given level of coherence, the differences were small in comparison to the explicit coherence manipulation (Experiment 1: small, high coherence = 0.80; large, high coherence = 0.72; small, low coherence = 1.03; large, low coherence = 0.98; Experiment 2: : small, high coherence = 0.59; large, high coherence = 0.54; small, low coherence = 1.07; large, low coherence = 0.99) . We have added the structural ratio of each training set to the revised Methods section and discuss manipulations of coherence in the Discussion.

Pg. 39:

“*Other studies showing the benefits of variability among training examples have induced variability by including more unique instances in the high variability condition and repeated individual items in the low variability condition, thus conflating the number of examples with the variability among difference examples (Doyle & Hourihan, 2016; Nosofsky et al., 2019; Wahlheim et al., 2012, 2016). Our study was novel in controlling both the amount of training and the number of examples when testing for the effects of training set coherence.*”

Pg. 40-41:

“*While our findings show a benefit of high coherence training that differs from studies that trained to criterion, they align well with a theoretical prediction based on computational modeling by Hintzman (1984) who argued that coherent training should be beneficial for generalization when the length of training is equated. They also align well with some other manipulations of category coherence. Prior work in the domain of social categorization has also shown that coherence (called ‘entitativity’ in this domain (Campbell, 1958)) is a key factor that drives naïve perceptions of what defines category membership (Haslam et al., 2000). Thus, part of the high coherence training benefit may be that it fits well with naïve intuitions about how categories are formed, allowing participants to quickly adopt strategies that are well suited to learning categories based on similarity to the category center. Another operationalization of category coherence is the structural ratio: distances between items within the same category compared* *to the distances between items from different categories (Homa et al., 1979; Minda & Smith, 2001). Prior work has shown faster learning from more coherent, 'well structured' category sets in which there is more clustering of items within vs. between categories (Minda & Smith, 2001).*"

Finally, the Reviewer notes that "to equate for study length one needs to increase the number of repetitions". It is indeed not possible to simultaneously equate the number of repetitions AND the total number of trials when varying training set size. This makes it challenging to ensure that any set size effects are indeed driven by set size specifically. We would like to point out that we are providing the first study that addressed this challenge by running the experiment both ways, equating for the total trials in one group of subjects and equating for the repetitions in another group of subjects. The fact that we are explicitly addressing this challenge and providing a new way to dissociate set size effects from repetition/total training length effects was positively noted by both Reviewer 2 and Reviewer 3.

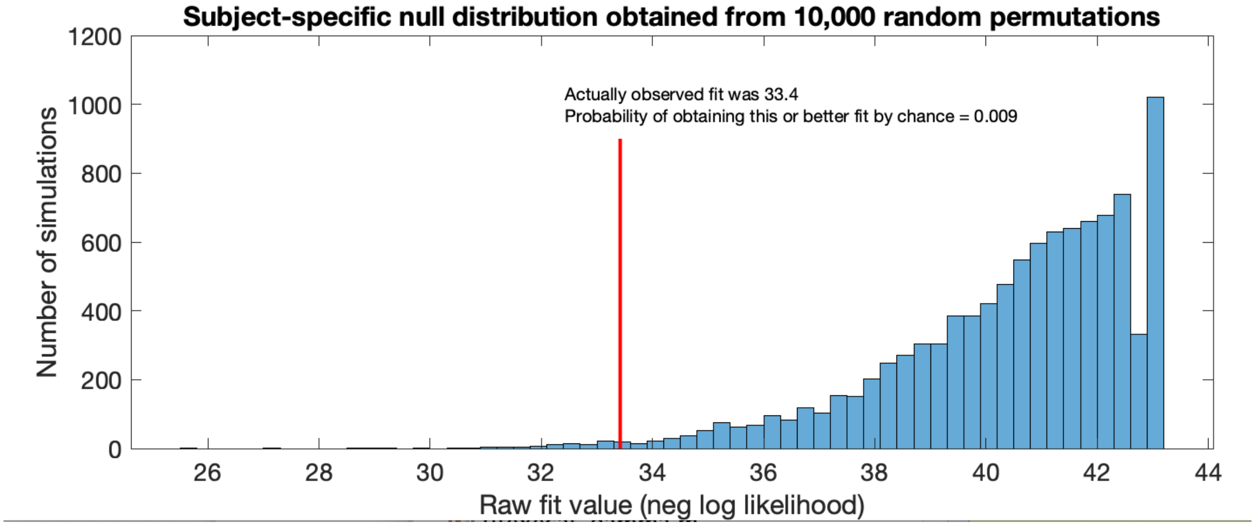
OTHER MAJOR ISSUES

R1.2. I find the modeling approach a little confusing. Why were the models not fit to the data and the model fit compared directly? I think the authors try to address this in the paragraph "We chose this alpha level for labeling participants' strategies, as in our prior study (Bowman & Zeithamova, 2020), as a compromise between a strict alpha level of 5%, which labeled many subjects as showing similar fits between the two models, and the most common no-alpha model selection approaches, such as when AIC is used to select the best fitting model based on lower fit error without addressing whether the fit value difference is reliably above chance." in p.18, but I found this paragraph opaque even after multiple reads. What is the issue with AIC/BIC again? More broadly, if I understand the approach correctly, the authors fit model A and model B to the data. Then they compare each model to the null model of permuted data (let's call them A' and B'). Then they compare A with A' and B with B'. If A is more likely than A' to a smaller rate than B is more likely than B', then A > B. But A and B could also be the same, right? Because their null models are different and A and B are actually not being compared directly. Even if their null were the same, unless A and B are directly compared, any pairwise comparison with something else does not directly inform the difference between A and B. All in all, I'm not sure about this approach, if it is OK, or why it is better than commonly accepted practices of comparing AIC/BIC (or any other measure of fit).

We apologize for the lack of clarity in our description of the model fitting procedure. In fact, to decide between the prototype model vs. the exemplar model, we *do* compare them directly with each other, as the Reviewer would expect. However, we have a step prior to the direct comparison where we check to see whether each model fits better than one would expect by chance, but the comparison to chance is not used to decide between the two models themselves. Specifically, there are two steps to our procedure. First, we assess whether each model fits a given participant's observed responses better than it fits randomized data where there is no real signal. This is the step the Reviewer refers to as comparing A with A' and B with B'. Participants for whom neither the prototype nor the exemplar model fits the observed data significantly better than the randomized data are labeled as 'chance' in our labeling scheme. The idea is that it would be misleading to label a participant as using one or the other strategy if there is no evidence that they used any strategy at all. For participants with at least one model outperforming chance (most participants), the second step is to *directly compare* the prototype and exemplar models to one another (i.e., A vs. B).

The Reviewer also asked why we didn't use a more traditional model selection method, such as AIC/BIC. First, we would like to assure the Reviewer that using AIC for model selection generates similar results and most subjects receive the same strategy label regardless of whether we use the AIC method or our permutation test method. Nonetheless, we used the permutation approach rather than AIC or BIC for two primary reasons. First, when determining whether the prototype and exemplar model fit better than you would expect by chance, AIC and BIC both penalize for the number of free parameters in the models. Since an attention parameter is estimated for each stimulus feature, we have found that this penalty can be quite severe for high dimensional stimuli like the 8- and 10-dimensional stimuli used in the present experiments. As a result, the AIC metric can label participants as responding randomly even when their classification accuracy is clearly above chance. In comparing the permutation and AIC methods in the present experiments, the AIC method labeled a higher percentage of participants as responding randomly. In Experiment 1, the AIC metric labels 25% of participants as responding randomly compared to 18% using the permutation method. Similarly, in Experiment 2, the AIC metric labels 37% of participants as responding randomly compared to 28% using the permutation method. The penalty for free parameters is even more severe with the BIC metric.

The following example illustrates that the permutation approach may provide a more appropriate comparison to chance. Below is the exemplar model fit from a subject who was best fit by a random model when AIC was used for model selection, but was considered an exemplarist when using the permutation approach. The ‘null’ distribution is in blue – how well the exemplar model fits simulated data where responses are paired with the stimuli randomly. The red line represents the actual exemplar model fit to the participant’s observed responses (33.4). Such a good fit (i.e., small negative log likelihood) only happened by chance in 95 out of 10,000 random simulations, and is thus very unlikely to happen if the subject was just responding randomly (p = 0.009). This is how we decide between each model and chance. Notably, this subject had 65% generalization accuracy. Thus, accuracy and the permutation test both suggest that this subject did NOT respond randomly, and the AIC metric is unnecessarily conservative.



In addition to allowing us to better differentiate from chance, the permutation approach allows for a more informed decision between the prototype and the exemplar model when we compare them directly. The AIC/BIC metrics only consider the number of parameters when determining whether one model outperforms the other, and not necessarily the degree of flexibility that parameters may provide (e.g., even when fitting randomized data). We have previously found that Monte Carlo null distributions for prototype and exemplar model differences are not necessarily centered on 0. In other words, even when randomized data are fit, one model can "fit better" on average. This is likely because one model or the other is more flexible to fit a wide range of responses – even ones with no real signal. In the current experiments, on average across all subjects, the exemplar model fit randomized subject data better than the prototype model in both Experiment 1 (mean exemplar advantage = 0.009; t(176) = 14.29, *p* < .001) and Experiment 2 (mean exemplar advantage = 0.005; t(275) = 17.64, *p* < .001). That is, when no real relationship exists between the stimuli and the responses, the exemplar model systematically fits better than the prototype model. The permutation approach can account for this bias when determining the threshold for labeling participants as best fit by the prototype vs. exemplar model. We provide more information about the relationship between model flexibility and model selection in our R2.2 response if the Reviewer would like more information.

We have revised the statistical analyses section to clarify and provide additional background on our model selection choices (pg. 17-19):

"*After optimization, we used Monte Carlo simulations to determine whether the prototype and exemplar models each fit better than chance and to determine whether the difference in prototype and exemplar model fits was greater than would be expected by chance (Bowman et al., 2020; Bowman & Zeithamova, 2018, 2020). To generate the subject-specific null distributions, we used the stimuli that the subject encountered and the subject's actual responses, but we randomly shuffled the stimuli and the responses with respect to each other. This produced randomized datasets in which there is no real relationship between the stimuli and the responses while maintaining any potential overall response biases. We then fit the prototype and exemplar models to this randomized data and stored the resulting prototype and exemplar model fit values. This procedure was repeated 10,000 times to generate a subject-specific null distribution of model fits for each model. The null distributions provide information about the typical range of model fit values that happen just by chance when the underlying data contain no real signal.*

We then compared each subject's observed prototype and exemplar model fits to their subject-specific prototype and exemplar null distributions to determine whether one or both models fit the participant's data better than chance at alpha = .05 (one-tailed). For example, a subject's observed exemplar model fit would be considered better than chance if fewer than 5% of the random simulations (from their exemplar null distribution) produced fits that were as good or better as the observed data fit. Participants for whom neither model outperformed chance were labeled as responding randomly ("chance").

For participants in which at least one of the models outperformed chance, we then directly compared prototype and exemplar model fits to one another using a relative difference in fits metric: (exemplar model fit – prototype model fit) / (exemplar model fit + prototype model fit). To determine whether one model reliably outperformed the other, we compared their observed difference in model fits to the null distribution of differences in model fits generated from the randomized data. One model was deemed a better fit than the other for a given participant when that difference score appeared by chance with a frequency less than 25% (75% probability that the model fit differences did not arise by chance, two-tailed test). Using this method, participants were labeled as prototype-users or exemplar-users when one model outperformed the other or as having "similar" model fits when neither model outperformed the other. We chose this alpha level for labeling participants' strategies, as in our prior study (Bowman & Zeithamova, 2020), as a compromise between a strict alpha level of 5%, which labeled many subjects as showing similar fits between the two models, and the most common no-alpha model selection approaches, such as when AIC is used to select the best fitting model based on lower fit error (or lower value of the AIC metric that accounts for the number of free parameters) without addressing whether the difference in fits is reliably above chance. It is *important to consider what constitutes a significant difference in model fits because, in practice, one model can systematically outperform the other even for randomly generated data, suggesting that one model may be more flexible (in general or for a given category structure or stimulus set). Using the simulation approach can account for this bias when comparing the models and setting a threshold for selecting one model over the other. Nevertheless, we verified that the majority of participants received the same strategy label using either approach (permutation test or AIC).*”

R1.3. Still on the modeling work, I would like to have seen best-fit values for the parameters c and w. This is important because these parameters would affect the psychological categorization space and might be influencing the results in important ways. Ideally, they would be the same across both models, why is that not the case?

Thank you for this suggestion. We have added the best-fit values for all model parameters in new Table 2 (Experiment 1) and Table 4 (Experiment 2).

Table 2 – Mean best fitting parameter values for the exemplar and prototype models in categorization

	Exemplar model										Prototype model								
Training set	C	G	W1	W2	W3	W4	W5	W6	W7	W8	C	W1	W2	W3	W4	W5	W6	W7	W8
Small, high coherence	10.0	36.2	.20	.15	.06	.21	.11	.08	.10	.10	35.3	.21	.14	.06	.18	.09	.08	.12	.12
Small, low coherence	27.7	37.5	.14	.11	.12	.15	.07	.12	.16	.12	17.4	.17	.12	.17	.07	.16	.06	.08	.17
Large, high coherence, repetition-matched	21.2	41.4	.23	.08	.08	.12	.14	.15	.12	.07	34.5	.21	.07	.10	.13	.13	.18	.11	.06
Large, low coherence, repetition-matched	17.6	52.5	.11	.06	.10	.15	.08	.15	.24	.11	20.6	.22	.09	.18	.07	.14	.10	.08	.11
Large, high coherence, trial-matched	11.9	58.2	.14	.14	.12	.19	.16	.10	.07	.09	19.6	.15	.13	.13	.18	.14	.10	.06	.09
Large, low coherence, trial-matched	31.5	.51.6	.16	.05	.11	.15	.02	.24	.22	.05	17.0	.24	.10	.19	.06	.08	.08	.12	.12

C = sensitivity, G = gamma (exemplar model only), W1-8 = attention weights to each of 8 stimulus features

Table 4 – Mean best fitting parameter values for the exemplar and prototype models in categorization

	Exemplar model												Prototype model										
Training set	C	G	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	C	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
Small, high coherence	16.1	35.6	.16	.11	.11	.12	.08	.11	.06	.03	.07	.15	34.2	.15	.09	.11	.11	.09	.11	.07	.04	.08	.16
Small, low coherence	13.7	31.9	.14	.09	.10	.16	.08	.08	.08	.08	.09	.11	7.1	.13	.10	.09	.15	.08	.11	.09	.07	.07	.10
Large, high coherence, repetition-matched	26.1	47.3	.14	.10	.10	.13	.05	.16	.09	.03	.10	.10	32.0	.13	.11	.09	.13	.07	.14	.12	.02	.08	.11
Large, low coherence, repetition-matched	15.0	33.9	.14	.08	.07	.16	.08	.12	.09	.07	.08	.11	5.3	.13	.06	.08	.14	.10	.12	.10	.07	.08	.13
Large, high coherence, trial-matched	10.9	42.8	.12	.05	.08	.11	.04	.13	.07	.06	.12	.22	33.3	.11	.04	.09	.11	.05	.14	.08	.06	.12	.20
Large, low coherence, trial-matched	11.6	46.8	.11	.08	.08	.12	.09	.07	.09	.10	.08	.16	4.6	.11	.08	.10	.10	.08	.09	.09	.10	.09	.17

C = sensitivity, G = gamma (exemplar model only), W1-10 = attention weights to each of 10 stimulus features

We did not constrain the models to come to the same best fitting parameters so that each model would have its best chance to fit the data. We did not want to fit the models with this constraint and risk one model or the other disproportionately influencing the parameter estimates, causing the other model to underperform without any ground truth as to the 'correct' values for these parameters. Yet, even without constraining the weights to match across models, the prototype and exemplar models generally returned highly correlated best fitting parameters. In addition to the similar average parameter values (see the tables above), we computed within-subject correlations of the best fitting attention weight values from each model and found that the mean r value was .66 for Experiment 1 and .91 for Experiment 2. Thus, as the Reviewer expected, the models seemed to pick up on similar features of participants' behavior, even without us intervening to make the parameters consistent across models.

We also calculated an across subject correlation for the C parameter (sensitivity) and found these values to be uncorrelated across models (Experiment 1 mean r value = .06, Experiment 2 mean r value = .07). This is likely because in the revision, we used an exemplar model that included both the C parameter and the gamma parameter, as requested by Reviewer 2. Only the C parameter is estimated for the prototype model as the gamma parameter cannot be separately estimated and is by convention fixed at 1 (see R2.2). In other words, the exemplar model allows gamma and c parameters to be estimated independently while the prototype model combines c and gamma into a single estimate (labeled as c). In our data, the c parameter from the prototype model (combining response scaling and sensitivity) was more strongly correlated with the gamma parameter from the exemplar model (Experiment 1 r = .29, Experiment 2 r = .25) and with the average of the sensitivity and gamma parameter (Experiment 1 r = .34, Experiment 2 r = .30) than with the exemplar model's c parameter. We also fit the exemplar model without the separate gamma parameter, thus combining c and gamma into a single parameter and affording a more direct comparison with the prototype model. When the parameters were the same in both models, the c estimate from the prototype model and the c estimate from the exemplar model were clearly correlated (r = 0.64).

We have added the following text about the similarity in best fitting parameters across along with Tables 2 and 4.

Experiment 1 (pg. 24):

"*Although we did not constrain the models to return the same or similar parameter estimates, there was general agreement between the prototype and exemplar models in terms of the estimated attention weights (average within-subject Pearson's r = .66). The remaining parameters were c (sensitivity) and gamma (response scaling) for the exemplar model and c (combining both sensitivity and response scaling effects in a single parameter) for the prototype model. The c parameter from the prototype model was more correlated with the gamma parameter from the exemplar model (across-subject Pearson's r = .29) than with the c parameter from the exemplar model (across-subject Pearson's r = .06).*"

Experiment 2 (pg. 36):

"*As in Experiment 1, we did not constrain the models to return the same or similar parameter estimates. Nonetheless, there was a strong average within-subject correlation of r = .91 for the attention weights estimated by the two models. Also similar to Experiment 1, the c-parameter* *from the prototype model was more correlated with the gamma parameter from the exemplar model (r = .25) than the c parameter from the exemplar model (r = .07).*"

R1.4. When comparing number of participants best fit by exemplar, prototype, or both models the authors use a regression analyses. I'm not sure that is appropriate. The data is basically count data, shouldn't a non-parametric test for count data be used?

It would indeed be problematic to use a linear regression, but we instead used a *binary logistic* regression model, as appropriate for predicting our binary outcome variable (i.e., whether or not participants were best fit by the prototype model). Logistic regression is used for categorical outcome variables. We have revised the results to add the word "binary" before "logistic regression" and report odds ratios in addition to the log odds (ß) values, which will make it more clear that the analysis was a logistic regression. Only two levels are used (prototype strategy, not a prototype strategy) as the bin counts in the exemplar and similar fit groups were too low to keep separate.

Please see updated report of this analysis in Experiment 1 (pg. 25-26) and the newly added report of binary logistic regression from Experiment 2 (pg. 37-38) in the revised manuscript.

R1.5. Previous work has shown that Exemplar models can account for concepts organized around prototypes (Medin & Shaffer, 1978; Nosofsky, 1986) because, although not seen, the prototype is maximally similar to all items seen. One would then predict that the exemplar model would be as good or better than the prototype model when there was high set coherence. That is not the case, and I cannot figure out why and the discussion does not include a good explanation why.

It is true that the exemplar model may predict better categorization of typical items (such as the prototypes) than less typical items and may even predict better categorization of prototypes than the training items themselves. It is also true that the exemplar model can predict better categorization after high coherence training than low coherence training.

In the revised manuscript, we now explicitly mention that the exemplar model can often fit prototype-based categories quite well and account for prototypicality effects (pg. 15):

Methods (p. 15):
"Although categorization performance analyses may reveal better accuracy of prototypes and the items most similar to prototypes, such prototypicality effects in categorization could result from either prototype or exemplar representations (Medin et al., 1978; Nosofsky, 1986). Thus, as in our prior study (Bowman & Zeithamova, 2020), we fit formal prototype and exemplar models to trial-by-trial categorization test data in individual subjects to estimate participants' categorization strategies."

Results (p. 24):
"*As both prototype and exemplar representations can produce typicality gradient in accuracy scores, we used formal categorization models to estimate the representations underlying categorization judgments.*"

However, it is not the case that the exemplar model should be *better* than the prototype model for high coherence sets. As a hypothetical example, consider how confidently each model would predict categorization of the prototypes (assuming all stimulus features are considered equally). As the Reviewer notes, the prototype stimulus would match the exemplar representation quite well, as it would be (e.g., in our high coherence condition in Exp. 1) two features away from all old exemplars of its category and six features away from all old exemplars of the opposing category. The exemplar model would thus predict relatively high accuracy for the prototype stimulus (perhaps even higher than for the old exemplars themselves) because it is relatively close to the training items of its category than opposing category. But it does NOT predict better accuracy than the prototype model does. Notably, a prototype stimulus will match the prototype representation even better than it matches exemplar representation. The prototype stimulus is a perfect match for the prototype representation as the distance to the prototype representation of its category is zero and the distance to the prototype representation of the opposing category is 8; a bigger contrast of distances to the two categories than the distances 2 vs. 6 per the exemplar model. The actual probabilities of responding A are just a little more complicated to compute (as perceived similarity is an exponential decay function rather than a linear function of physical distance between two stimuli) but clearly 8/0 match will lead to a stronger prediction than 6/2 match. Thus, the prototype model makes at least as strong or stronger prediction for the prototype stimulus itself than the exemplar model does, which matches participants' behavior that nears 100% accuracy for prototypes. Similarly, the prototype model makes a stronger prediction for distance 1 new items (9 features matching category A prototype vs. 1 feature matching B prototype) compared to the exemplar model (average 7.5 features matching category A exemplars vs. 2.5 features matching category B exemplars). Thus, the exemplar model can predict prototypicality effects, especially when the response scaling parameter is included (as we do now per Reviewer 2's recommendation), but it does not necessarily predict better prototype/distance 1 accuracy than the prototype model does.

5. There is no power analyses or justification of sample size. It is thus impossible to evaluate whether the study is adequately powered to find the complex interaction that is being investigated.

We apologize for this oversight and have added sample size justifications to the revised manuscript.

Experiment 1 (pg. 9-10):

"*One hundred eighty-three participants from the University of Oregon completed the experiment for course credit. Prior to data collection, we aimed to recruit approximately 30 participants for each between-subjects condition (total ~180) based on the rule of thumb from Central Limit Theorem. Five participants were subsequently excluded due to missing data in some portion of the experiment, and one other participant was excluded for failing to respond on ~60% of categorization trials, leaving data from 177 participants reported in all analyses (119 female, mean age = 19.97 years, SD age = 2.38 years, range 18-35 years). This final sample size is sufficient to detect a set size x set coherence interaction effect of η_p^2 > .055 with 80% power.*"

Experiment 2 (pg. 28):

"*Prior to data collection, we conducted a power analysis and determined that 270 participants are needed to detect a small to medium (η_p^2 > .03) set coherence x set size interaction effect with 80% power. We recruited 289 participants to ensure we could meet the target sample even after exclusions. Of those, six were excluded for failure to complete all experimental phases and one was excluded due to a deviation from the study protocol, leaving data from 282 participants reported in all analyses (205 self-reported females, 68 self-reported males, and 9 who self-reported as another gender; mean age = 19.1 years, SD age = 1.3 years, range 18-27 years).*"

6. It is not clear to me why comparisons with chance are always performed. The question of interest is whether there were condition differences, so comparison with chance seems irrelevant.

We appreciate the Reviewer’s concern about including unnecessary analyses, and we agree that the comparisons to chance are not critical for the main conclusions. We have removed them from the revised manuscript.

Reviewer #2:

Review of XLM-2021-1967 "Coherent category training enhances generalization and increases reliance on prototype representations" by Bowman and Zeithamova

Reviewed by: Rob Nosofsky (you may reveal my identity to the authors and the other reviewers)

Summary

The authors conduct category-learning experiments involving stimuli composed of 8 binary-valued dimensions that are generated from prototypes. They manipulate two main independent variables: the coherence of the training examples (extent to which they are similar to the prototypes) and category size (number of distinct training examples). Furthermore, there are two versions of the category-size manipulation, one which holds fixed across the small- and large-size conditions the number of repetitions of each individual training example, and the other that holds fixed the total number of training trials. The main empirical results are that generalization to novel transfer items is better in the high-coherence than in the low-coherence condition, with little effect of the category-size manipulation. Formal modeling indicates that more subjects rely on a prototype strategy than on an exemplar strategy following high-coherence training. Old-new recognition is also tested. In general, old-new discrimination is poor (and a variety of more nuanced empirical results involving recognition are also reported).

Overall Evaluation

From one perspective, I thought extremely highly of this manuscript: the factorial manipulation of coherence and category size is a very welcome development, and the careful distinction between equating repetitions versus total number of training trials is an extremely important one. Thus, the work has the potential to add significantly to our knowledge of factors that affect category learning and generalization and how they may interact. In addition, the evaluation of models at the individual-subject level shows formal sophistication. In general, the article was well written (although some clarification as well as qualification of conclusions is needed in some spots). For these reasons, I am hopeful that a revision of the article will be appropriate for publication in JEP:LMC.

We thank Dr. Nosofsky for his enthusiasm about the novelty and importance of our study. We have provided detailed responses to his comments below, and we are also hopeful that he will find the revised article appropriate for publication in JEP:LMC.

From other perspectives, however, I have some significant concerns that I would like to see addressed before publication takes place. My major concerns are the following:

R2.1. The manipulation of "coherence" is confounded with the variable of how many of the individual dimensions are diagnostic of category membership. As reported in the appendix, for the small-size training sets, all 8 dimensions are diagnostic of category membership in the high-coherence condition (75% of the values on all of the individual dimensions point to the correct category). But in the low-coherence condition, only 4 of the 8 dimensions provide diagnostic information (for the remaining dimensions, the contrasting binary values point half the time to one category and half the time to the other). (A similar confound occurs in the large-set-size conditions, although not as extreme.) Thus, we do not really know if the coherence effects really have to do with how variable the instances are around the prototype, or with how many dimensions the subjects have to search through before they discover dimensions that are diagnostic. As I discuss below, there is much evidence that in tasks very similar to the present one, subjects really do tend to focus attention on a very small number of dimensions that compose the objects, so the number of dimensions that the subjects need to search through to locate diagnostic ones is a potentially major confound.

We appreciate the careful review of the training set structures and for pointing out the issue with non-diagnostic features in the small, low coherence category. All our prior experiments used training structures that have equally predictive features, but this one unintentionally included non-diagnostic features. Thus, we agree that it is important to verify that our effects were not driven by this difference between high and low coherence conditions. As described in our note to the Editor, we have collected new data with training sets that manipulate set size and coherence while also ensuring that all features within a training set are equally predictive of category membership. Tables depicting the new category structures are included in Response section E1.2 and in the Appendix to the manuscript.

We elected to add this new dataset as an Experiment 2 rather than replacing our original Experiment, allowing us to show the robustness of the set coherence effect and remain transparent about the need for additional data following what is now Experiment 1.

R2.2. The "exemplar" model that the authors fit to data was abandoned by exemplar theorists three decades ago - at least when the goal involves fitting data at the level of individual subjects and when the experimental design has individual dimensions or exemplars that are probabilistically assigned to categories. There is an extremely long history here and it would take me many pages to review it. For now, I refer the authors and the editor to Nosofsky and Zaki (2002) (full reference below) for an extensive discussion. The short story is that when one uses Equation-3 for the exemplar model, it is forced to predict "probability-matching" behavior -- e.g., if the evidence is .7 that an item belongs to Cat. A and .3 that an item belongs to Cat. B, then the Equation-3 choice rule predicts that the observer will classify the item into Cat. A with probability .7. But it is well known that individual subjects often respond with more deterministic response strategies. To deal with the problem, early in the 1990s, the Equation-3 choice-rule version of the model was extended to include a response-scaling parameter "gamma". All serious models of human classification include an analogous response-scaling parameter, and process-interpretations for the emergence of the gamma parameter were developed by Nosofsky and Palmeri (1997). (What makes things complicated is that the gamma parameter is implicit in the prototype model as well, but it cannot be estimated separately from the sensitivity parameter "c" in Equation 1, so it can be arbitrarily held fixed at 1 in that model - see Nosofsky and Zaki 2002 for extensive discussion.)

Because the authors are fitting the probability-matching exemplar model, and a prototype model that allows more deterministic response rules, we have no idea if the better fits of the prototype model are really reflecting that the subjects have developed prototype representations, or if observers are storing exemplars but are responding with more deterministic response rules than is assumed in probability matching.

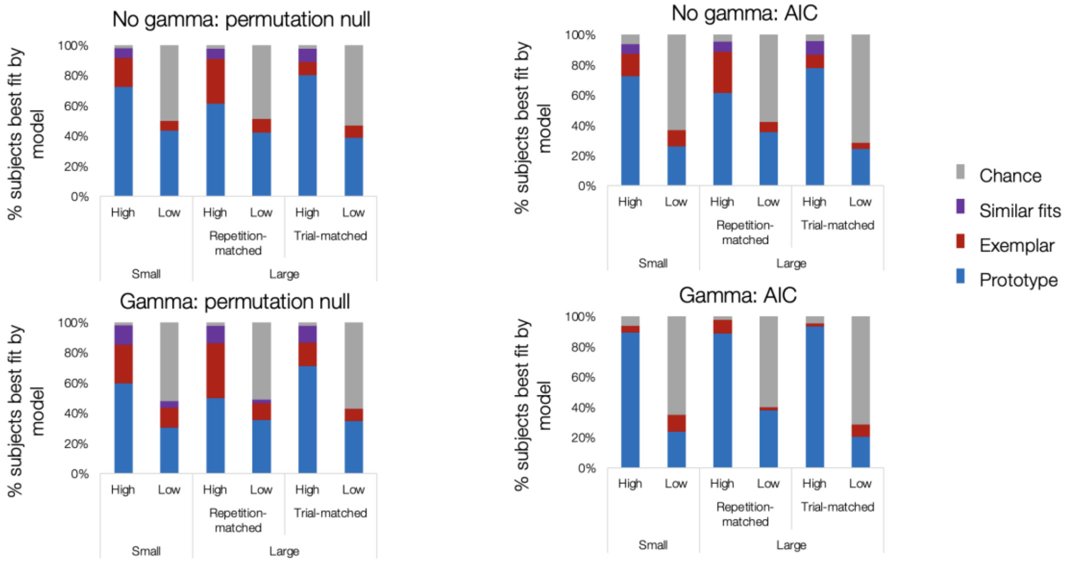
We appreciate the detailed explanation clarifying the theoretical and practical utility of the gamma parameter. We initially used the simplified exemplar model (with c but not gamma) to equate the number of parameters across the two models for ease of comparison. However, we realize that this may have differentially affected the flexibility of the two models and perhaps underestimated the fit of the exemplar model.

As requested, we have re-computed the exemplar model fits and included the additional gamma parameter. All the model fitting methods and categorization results in the revised manuscript were updated to reflect the gamma-including version of the exemplar model.

Interestingly, this provided additional evidence for the utility of our Monte Carlo permutation approach to model selection (more explanation of the permutation approach included in response to your question R2.9 and also R1.2). Notably, including the gamma parameter *does* lead to more subjects who are best fit by the exemplar model, but only when we use our permutation approach rather than AIC metric for model selection. Using the permutation approach for model selection, adding gamma to the exemplar model led to an increase of subjects best fit by the exemplar model from 11% to 20% in Experiment 1 and from 13% to 18% in Experiment 2. However, if we were to use the AIC metric, the penalization for the additional parameter in the gamma version of the exemplar model would outweigh the improvement in raw fits, such that we would see no change in the proportion of 'exemplarists' in Exp 1 (13% with or without gamma) and a *decrease* from 12% to 6% in Exp 2. To provide an intuition why this may happen, consider subjects who have perfect accuracy. They can be perfectly fit by either model, but AIC would always label the prototype model as a winner given it has one less free parameter than the gamma-including exemplar model. The no-gamma version of the exemplar model can still fit that data well, as the c parameter can stretch the similarity space enough to generate response probabilities close to 0 or 1. Thus, the no-gamma (probability matching) version of the exemplar model can be nearly as flexible as the gamma-including version, without the cost of the extra parameter. Our permutation approach seems to better reflect the real increase in flexibility of the gamma-including version that does not over-penalize the extra parameter.

For the Reviewer's information, we include below a comparison of the strategy labels obtained using different model section metrics (permutation vs. AIC) and exemplar model versions (no gamma vs. gamma). The pattern of findings with regard to set size and coherence are consistent across the four modeling approaches, as illustrated in the figure below. Thus, it is unlikely that the main conclusions would be driven by any particular modeling approach.

Categorization: Experiment 2


No gamma: permutation null

No gamma: AIC

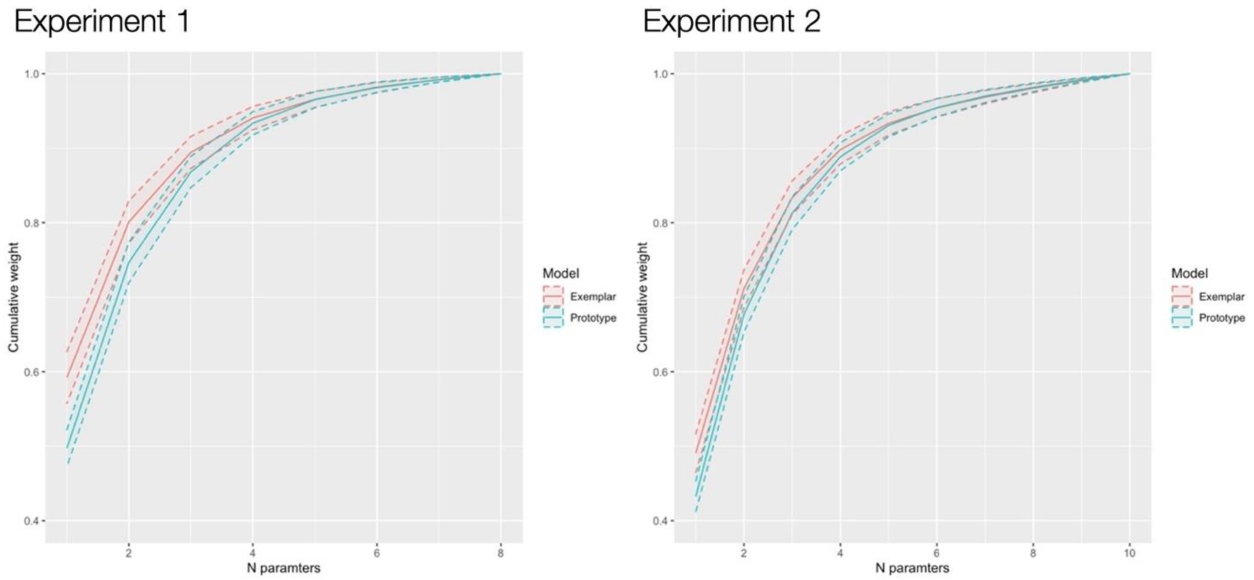
Gamma: permutation null

Gamma: AIC

R2.3. Closely related to the above issue concerns the number of dimensions to which subjects are attending in this experimental design. In fitting the models, the authors estimate 8 attention-weight parameters for each model. However, in experimental designs very similar to the present high-coherence large-size condition, Zaki and Nosofsky (2001) and Nosofsky et al. (2012) discovered that the vast majority of subjects apparently attended to only a very small subset of the 8 dimensions. This finding has important implications for the present study. First, because the subjects may have coded only a small subset of the dimensions, the low-dimensional components stored in memory really will be probabilistically associated with the alternative categories, making the probability-matching assumption in Equation 3 highly problematic. Second, such an attention strategy would explain why subjects are unable to discriminate old from new examples in the old-new recognition tests: from a psychological coding perspective, there is no distinction between the old and new examples when only a small number of the dimensions have been coded and stored in memory.

Because of this issue, and because of the prior comment, we have included the gamma parameter in the revised manuscript when fitting the categorization data. This should minimize the particular concern about the probability-matching assumption as it relates to attention to individual features.

However, to further explore this question, we wanted to assess how many features our participants seemed to consider when using our high-dimensional stimuli. To estimate the number of features participants used to make categorization judgments, we took the vector of weights estimated from each model for each subject and sorted each vector from the highest weighted feature to the lowest. We plot below the mean cumulative weight across subjects accounted for by n features, with n ranging from 1-8 in Experiment 1 and 1-10 in Experiment 2 based on the respective dimensionality of the stimuli. The dashed lines depict 95% confidence intervals.



Experiment 1

Experiment 2

Consistent with the prior work cited (Nosofsky et al., 2012; Zaki & Nosofsky, 2001), the models estimated that participants paid attention to ~3-5 features on average in Experiment 1 and ~4-6 in Experiment 2 when making their categorization judgments. Because the revised manuscript is already quite long, we elected not to include this analysis in the revised manuscript. We have, however, explicitly acknowledged that prior work has shown that subjects do not pay attention to all features when high-dimensional stimuli are used, including the suggested references (pg. 15-16). We have also added tables that include the best fitting parameters for each model and note the correlation across models in the estimated attention weights (see R1.3, pg. 24 and pg. 36).

We also agree with Dr. Nosofsky that attending to only a subset of the dimensions would affect old/new recognition in terms of corrected hit rate, as there is no difference between stimulus 1111 and 1110 when one never attends to the fourth dimension. However, further considerations suggest that the attention explanation may not provide the full account of the poor recognition. When we fit the models to the recognition data, most subjects were labeled as ‘chance’ – neither model fitting the observed data significantly better than they fit randomized data that contains no real signal. This was true whether or not we included the gamma parameter. Since the models *do* take into account attention (or lack thereof) to each feature, it seems that the low recognition scores are not solely driven by attention being limited to 3-4 features. We suspect two factors playing additional role in the poor recognition scores. First, recognition requires remembering the unique conjunction of the features, and it may be especially difficult to differentiate old from new items under such circumstance (e.g., 1100 and 0011 being old, but 0101 and 1010 being new). Second, there was an overall bias to respond ‘old’ during recognition (Experiment 1 proportion ‘old’ responses = .59; Experiment 2 = .62) that conflicted with our recognition stimuli which were unbalanced in the direction of having more new than old stimuli (65-81% new stimuli depending on training condition and experiment).

While we wanted to be transparent about the recognition data and acknowledge coherence and set size effects that emerged in Experiment 2, we also realize that the resulting paper is quite long. We thus elected to put the recognition part of the study in Supplementary Materials. We retained the non-gamma version of the models for recognition since adding it did not lead to substantial improvement in detecting subjects’ strategies (see R2.10 for details) and because it has not been used routinely when fitting recognition data (Nosofsky & Kantner, 2006; Nosofsky

& Zaki, 1998; Nosofsky & Zaki, 2003). All recognition data are also freely available on OSF. However, if the Editor and Reviewers think that it is worth including the recognition data in the main manuscript, we would be willing to put it back in.

R2.4. More generally, all of the authors' modeling conclusions are based on measures of model fit, with no attempt at developing qualitative contrasts. In addition, they display certain graphs that show poor old-new discrimination and categorization- and recognition-endorsement rates that are consistent with similarity to the prototype. But they never show what the fitted formal models themselves actually *predict*. There is a huge literature illustrating that exemplar models are quite capable of predicting poor old-new recognition as well as major prototype-enhancement effects in both classification and recognition.

In re-reading the paper, we agree that some of our framing was oversimplified. We revised the narrative to avoid implying that strong prototypicality effects are necessarily indicative of prototype representations, referencing the papers that Dr. Nosofsky suggested, especially the exemplar model predicting prototype enhancement effects (pg. 15):

“*Although categorization performance analyses may reveal better accuracy of prototypes and the items most similar to prototypes, such prototypicality effects in categorization could result from either prototype or exemplar representations (Medin et al., 1978; Nosofsky, 1986). Thus, as in our prior study (Bowman & Zeithamova, 2020), we fit formal prototype and exemplar models to trial-by-trial categorization test data in individual subjects to estimate participants’ categorization strategies.*”

We have also revised our introduction and discussion of the recognition data (now in Supplementary Materials) to avoid implying that poor recognition performance is necessarily evidence against the exemplar model.

Introduction to the Supplement (pg. S1):

"*In addition to our interest in the impact of training set coherence and training set size on categorization, we were also interested in how those factors affect other types of memory judgments – namely recognition memory. How categorization performance and category representations relate to those of recognition has been debated (Knowlton & Squire, 1993; Kumaran & McClelland, 2012; Marsh et al., 2015; Nosofsky, 1988; Varga et al., 2019; Zeithamova et al., 2012; Zeithamova & Bowman, 2020). We expected that low coherence training would make it easier to distinctly represent individual category members because they are less overlapping with one another, facilitating later recognition performance. We also expected that small set sizes would promote better recognition memory than large set sizes because large sets require encoding twice as many individual items into memory. We expected the contrast between large and small set sizes to be particularly strong when we equated the total number of training trials because participants were exposed to the individual items in those large training sets half as many times.*”

Supplementary discussion (pg. S9-10):

“*Across two experiments, we found that the multidimensional stimuli used in the current experiment posed a challenge for recognition. The vast majority of subjects showed not only poor recognition in terms of the corrected hit rate, but also model fits that were not significantly better than what can be expected by chance. This finding is notable since both models tended to fit the categorization data at above-chance levels, suggesting that prototype and/or exemplar* *representations were formed by many participants and should have been available for making other types of decisions. Nonetheless, neither representation was apparent in the recognition behavior. One possibility is that participants do use prototype and/or exemplar-based similarity in recognition, but that the decision rules in Equations S1 and S2 do not accurately reflect the way that participants use similarity information to make their judgments. Another possibility is that the quality of prototype and/or exemplar representations that participants may have formed were not detailed enough to be useful for recognition, which required memory for a unique combination of 8 or 10 features. For example, participants may have seen stimuli with values 1100 and 0011 on the first four dimensions (where 1 would indicate one value of a feature, such as a squared body and 0 would indicate the other value of a feature, such as a round body) but not encode the precise conjunction of the four features. This may have led to strong feelings of familiarity for most of the recognition stimuli and an overall bias to respond 'old', without strong differences based on similarity to prototypes or exemplars. Following the example above, stimuli with values 1100, 1010 or 1000 may all feel equally familiar since each individual feature had been presented many times before but not always in that exact combination. Participants indeed tended to respond 'old' more often than 'new' (response bias towards 'old' response = 59% in Experiment 1, 62% in Experiment 2), even though the majority of test items were new (only 19%-35% old items, depending on the set size and Experiment). Our permutation approach for model selection considered each participant's bias to respond 'old' when generating the subject-specific null distributions. The high proportion of subjects given the 'chance' label therefore indicates that the prototype and exemplar models did not offer additional explanatory power above-and-beyond knowing this response bias. Thus, the effect of training set on recognition and the categorization-recognition relationship may be more conclusively evaluated in future work, perhaps utilizing different stimuli.*"

R2.5. There are additional articles bearing on the main empirical patterns of results that I would like to bring to the authors' attention. Regarding the category-size manipulation, in studies using real-world categories (rock types in the geologic sciences), Nosofsky et al. (2019) conducted conditions in which subjects learned 10 categories and then tested generalization performance. In one condition the categories were size-3 and in a second condition the categories were size-9. Their procedure was basically the same as the present authors' "equate total trials" condition. Unlike in the present study, Nosofsky et al. found that generalization to novel transfer items was significantly better in the large-size condition than in the small-size condition. (This was despite the fact that classification of old training instances was better in the small-size condition than in the large-size condition - not surprising because the individual training instances were repeated way more often in the small-size than in the large-size condition.) See Figure 5 in their article for the summary results, and see the section "Number and variability of training instances" in their General Discussion for extensive discussion. There are a couple of other articles relevant to this issue of category size and instance variability that are also discussed in that section. I suspect that the difference in the pattern of results across the present study and the previous ones has much to do with the very different kinds of category structures and stimuli involved - cartoon animals varying along 8 binary dimensions that are generated from prototypes, versus objects that vary along continuous dimensions, that can be coded holistically, and that are sampled from real-world categories. Likewise, I strongly suspect that - even controlling for total number of training trials -- higher-variability training instances would promote better generalization than lower-variability ones for category structures and stimulus types different from these binary-dimension cartoon animals.

Thank you for pointing us to this highly relevant study and highlighting the need to be careful about not overgeneralizing our results. In the current manuscript, we have added a discussion of other category structures and included the suggested citation. (See also our response to Editor's point E3 above).

Discussion (pg. 41-42):

"*The advantage of learning from high coherence training sets appears robust across the experiments presented here and our prior work with similar category structures (Bowman & Zeithamova, 2020). However, it may not be equally suitable for all types of stimuli or category structures. While we showed this effect at two different levels of high stimulus dimensionality (8 and 10 dimensional stimuli), the stimuli were always binary dimension cartoon animals, and the category structure was always prototype-based. Categories not centered around a single prototype (e.g., multiple prototypes, disjunctive or rule-plus-exception category structures) may require a very different sampling of training exemplars to be learned robustly. For example, using natural categories like rocks and birds, others have shown that training sets that span the full category space or offer a wider variety of examples may be particularly good for promoting generalization (Nosofsky et al., 2019; Wahlheim et al., 2012). Similarly, information-integration categories with non-linear boundaries require extensive training with large number of varied exemplars to provide an opportunity to learn not only the central tendency of a category but the entire distribution range (Ashby & Gott, 1988; McKinley & Nosofsky, 1995). Thus, one possibility is that high coherence training is uniquely well suited to prototype-based categories because it leads participants toward the underlying category structure. For other types of categories, especially those that are not linearly separable, greater variability and/or extended training may be quite important. Interestingly, even rule-plus-exception category structures may be easier to learn when exposure to exceptions is delayed (Heffernan et al., 2021), suggesting that training that highlights commonalities among category members not only makes it easier to learn rules or prototype-based categories but also memorize the exceptions.*"

Other Points

R2.6. p. 5. The authors claim that studies that find better fits of the exemplar model than the prototype model tend to show only a small number of examples during training. Ashby, Maddox and their colleagues have reported innumerable studies using their "general-recognition-randomization technique" with hundreds or thousands of training examples, in which models with non-linear decision boundaries produce enormously better fits than prototype models (which produce linear boundaries). McKinley and Nosofsky (1995) showed that exemplar models predict these good-fitting non-linear boundaries, again in paradigms involving hundreds or thousands of training examples.

We thank Dr. Nosofsky for noting that our description was oversimplified. We have revised introduction and discussion sections to better acknowledge these counterexamples.

Introduction (pg. 5-6):

"*Studies using continuous-dimension stimuli have shown that large training set sizes do not always lead to reliance on prototype representations. While the prototype model predicts a linear decision bound between categories, participants trained on many unique exemplars sampled from two bivariate normal distributions can learn and adopt an optimal, non-linear category boundary, indicating they must be representing the variability of each category in addition to its center (Ashby & Gott, 1988; Ashby & Maddox, 1992; McKinley & Nosofsky, 1995).*"

Discussion (pg. 41-42):

"*Similarly, information-integration categories with non-linear boundaries require extensive training with large number of varied exemplars to provide an opportunity to learn not only the central tendency of a category but the entire distribution range (Ashby & Gott, 1988; McKinley & Nosofsky, 1995). Thus, one possibility is that high coherence training is uniquely well suited to prototype-based categories because it leads participants toward the underlying category structure. For other types of categories, especially those that are not linearly separable, greater variability and/or extended training may be quite important.*"

R2.7. Related to the above point, I think it's important for the authors to qualify the generality of their conclusions. In the present designs, the categories are being PRODUCED by starting with prototypes and creating instances that are statistical distortions of the prototype. To optimize generalization performance, an ideal observer would therefore in fact adopt the prototype strategy (i.e., that is what is being used to create the categories). It is a completely open question, however, whether categories in the real world have structures analogous to those produced by perturbing a large number of binary-valued dimensions around their prototype values. As was found by Nosofsky et al. (2019), very different patterns of results may be observed in designs using different kinds of category structures and stimuli.

This is a good point, and we have included it as part of the revised discussion (also part of R2.5; pg. 41-42):

"*The advantage of learning from high coherence training sets appears robust across the experiments presented here and our prior work with similar category structures (Bowman & Zeithamova, 2020). However, it may not be equally suitable for all types of stimuli or category structures. While we showed this effect at two different levels of high stimulus dimensionality (8 and 10 dimensional stimuli), the stimuli were always binary dimension cartoon animals, and the category structure was always prototype-based. Categories not centered around a single prototype (e.g., multiple prototypes, disjunctive or rule-plus-exception category structures) may require a very different sampling of training exemplars to be learned robustly. For example, using natural categories like rocks and birds, others have shown that training sets that span the full category space or offer a wider variety of examples may be particularly good for promoting generalization (Nosofsky et al., 2019; Wahlheim et al., 2012). Similarly, information-integration categories with non-linear boundaries require extensive training with large number of varied exemplars to provide an opportunity to learn not only the central tendency of a category but the entire distribution range (Ashby & Gott, 1988; McKinley & Nosofsky, 1995). Thus, one possibility is that high coherence training is uniquely well suited to prototype-based categories because it leads participants toward the underlying category structure. For other types of categories, especially those that are not linearly separable, greater variability and/or extended training may be quite important. Interestingly, even rule-plus-exception category structures may be easier to learn when exposure to exceptions is delayed (Heffernan et al., 2021), suggesting that training that highlights commonalities among category members not only makes it easier to learn rules or prototype-based categories but also memorize the exceptions.*"

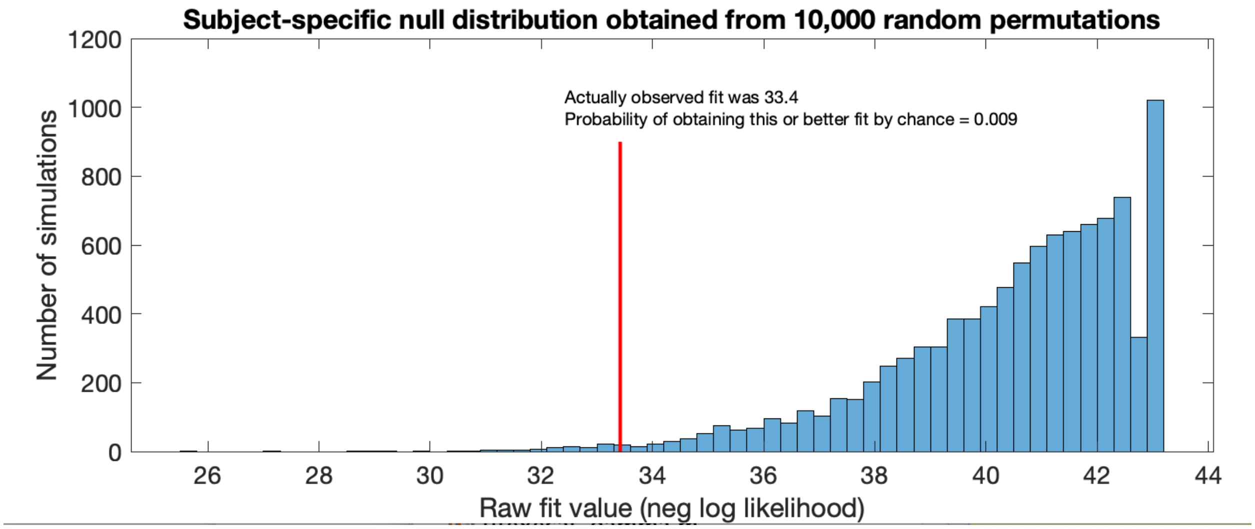
R2.8. p. 13. What was the purpose of the 8-second fixation periods that followed the responses?

The long fixation periods were included to be comparable to fMRI paradigms we were collecting at a similar time, which have timing constraints in order to estimate the brain signal in response to a trial. In our newly added Experiment 2, we use a more typical 1-second intertrial interval. We have revised the text in Experiment 1 to explain the reason for the long ITI (pg. 14):

"*In both tests, each exemplar was presented on the screen for 4 seconds followed by an 8 second fixation period. The relatively long fixation period was included so that the paradigm would be comparable to recent fMRI studies (Bowman et al., 2020; Bowman & Zeithamova, 2018).*"

R2.9. p. 17. Honestly, I was unable to understand the shuffling procedure that was used here in evaluating the model fits. A better explanation is needed.

We apologize for the lack of clarity. To generate the subject-specific null distributions, we keep the stimuli that the subject encountered and their responses, but we shuffle the stimuli and the responses randomly with respect to each other. This produces a randomized dataset in which there is no real relationship between the stimuli and the responses while still maintaining in the randomized data any overall response bias the participant may have had. We then fit the models to the randomized data and store the resulting prototype and exemplar model fits. We repeat the randomization and model fitting 10,000 times. The main goal is to estimate the 'null' distribution, or the typical range of model fits to data that do not include any real signal – the range of values that happen simply by chance. We then compare the actual observed fit value to the null distribution of fits to see how likely the observed fit happens just by chance. The inserted figure illustrates this approach and how we compared each model to chance. The blue distribution represents the distribution of random exemplar model fits (exemplar model fits to randomized data for that subject), with raw fit values centered around 40 (mean random fit = 40.4, median random fit = 40.9), with a clear skew. The observed fit to the actual subject's data (not randomized) was 33.4. A fit of 33.4 or better (lower negative log likelihood) happened in only 94 out of 10,000 simulations by chance, making the probability of the fit value as low as 33.4 to be p = 0.009 (very unlikely). We would thus consider the model fit to be better than chance.



We have revised the description of the shuffling procedure to make it more clear (pg. 17-18):

"*After optimization, we used Monte Carlo simulations to determine whether the prototype and exemplar models each fit better than chance and to determine whether the difference in* *prototype and exemplar model fits was greater than would be expected by chance (Bowman et al., 2020; Bowman & Zeithamova, 2018, 2020). To generate the subject-specific null distributions, we used the stimuli that the subject encountered and the subject's actual responses, but we randomly shuffled the stimuli and the responses with respect to each other. This produced randomized datasets in which there is no real relationship between the stimuli and the responses while maintaining any potential overall response biases. We then fit the prototype and exemplar models to this randomized data and stored the resulting prototype and exemplar model fit values. This procedure was repeated 10,000 times to generate a subject-specific null distribution of model fits for each model. The null distributions provide information about the typical range of model fit values that happen just by chance when the underlying data contain no real signal.*"

While this is computationally intensive, it provides some advantages over traditional model comparison approaches, such as using AIC or BIC metric. We discussed some of the reasons in our response to R1.2 and will reiterate here. First, the comparison of the exemplar or prototype model to a random model would involve quite dramatic penalization for free parameters, given our high-dimensional stimuli. Using AIC for model selection (prototype, exemplar, random) leads to a larger portion of subjects being best fit by the random model than when we use the permutation approach (25% instead of 16% in Experiment 1; 36% instead of 28% in Experiment 2). Every subject who is labeled as responding randomly per the permutation approach is also best fit by the random model when using AIC model comparison. However, some subjects who are labeled as "chance" by AIC get assigned a strategy when using the permutation test for model selection (their model fits are considered above chance). For example, the histogram of the exemplar null distribution provided above is from a subject who was best fit by a random model when using AIC for model selection, but who was labeled as an "exemplarist" using the permutation approach. We believe the permutation-based classification makes more sense. As noted above, their simulated null distribution indicates that their observed exemplar model fit (33.4) does not seem likely to happen by chance. Furthermore, the subject had 65% accuracy for categorization of new items. Both categorization accuracy and the permutation test indicate this subject did not respond randomly, so the AIC is unnecessarily strict in this case. Thus, even just for the comparison to chance, we believe that the permutation approach is better than AIC/BIC.

The second reason is that the permutation approach also offers a more principled (and more realistic) comparison between the prototype and exemplar models themselves. When using the no-gamma version of the exemplar model, both models have the same number of parameters and thus AIC picks whichever model has the better raw fit. However, for very small differences in fits, one would either have to call one model a winner irrespective of how small the fit difference is or set an arbitrary threshold without knowing what fit differences are likely just noise and which are large enough to be meaningful. Using our permutation approach, we can calculate the difference in fits for each random simulation and consider the size of the observed fit difference with respect to the null distribution of fit differences one would expect by chance. It may not be perfect, but it is less arbitrary than the alternative. Another reason to consider what model fit differences happen just by chance is the fact that the two models are not equally flexible even though they had the same number of parameters. For example, when fit to random categorization data, the exemplar model usually happens to fit a little better than the prototype model. In our Experiment 2, the average fit to random data was 38.00 for the exemplar model and 38.17 for the prototype model. In other words, in the current task, the exemplar model can fit pure noise slightly better than the prototype model can, and the permutation approach can take that into account.

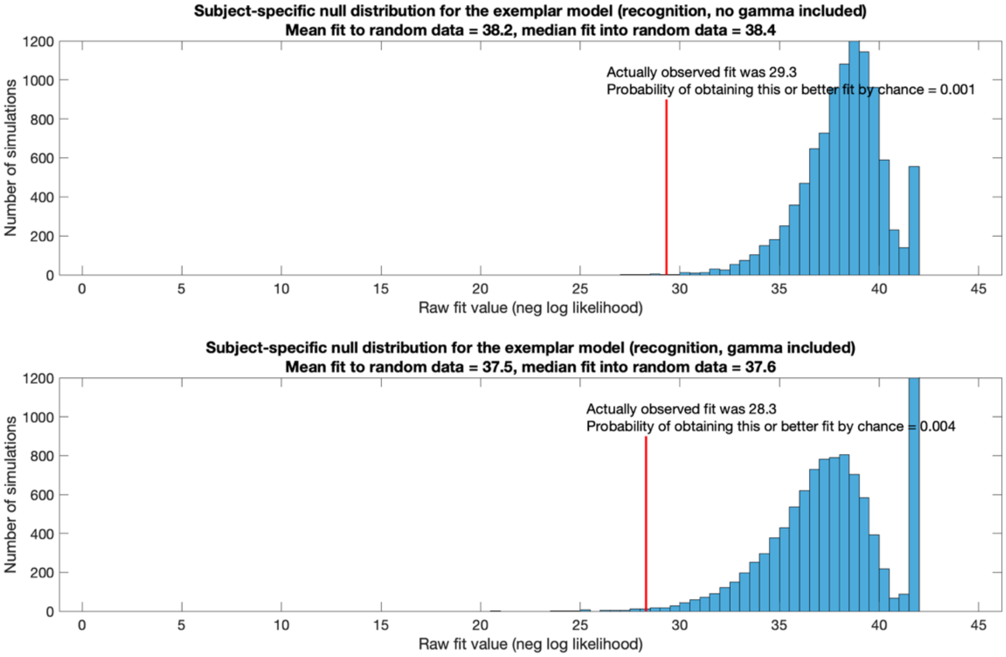
As we responded in R2.2., even more reasons for using the permutation approach emerged when we switched to the gamma-including version of the exemplar model during revision. The gamma-including model was indeed generating better fits than the no-gamma version, but the improvement was not large enough to make up for the AIC penalization for the extra parameter. We were able to see the advantage of the gamma-version of the exemplar model only when we used the permutation approach, indicating that the permutation approach more realistically reflects the extra flexibility due to the added parameter, avoiding over-penalization.

R2.10. p. 18. Possibly, the reason for the poor fits to the recognition data (of both models) may have to do with the lack of the response-scaling parameter in the Equation-4 choice rule. Although the response-scaling parameter cannot be estimated separately from the sensitivity parameter when the prototype model is used to predict classification, the same is not true for the Equation-4 recognition response rule.

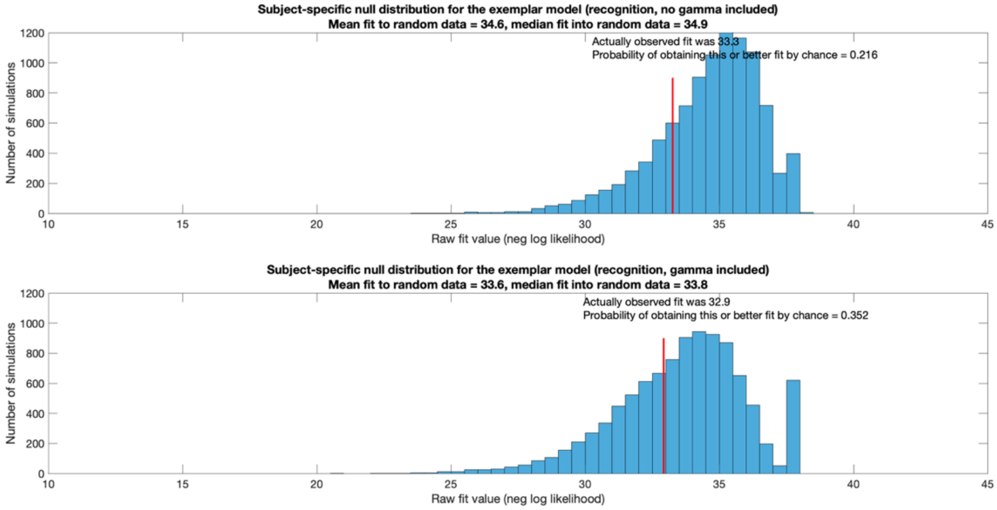
We have re-run the modeling analyses using the gamma-including versions of both models. While the raw fits indeed improved, the strategy classification results of Experiment 1 did not change much (permutation: 86% still best fit by random model, down from 88% without gamma; AIC: 92% still best fit by random model, down from 95% without gamma). The recognition results from the new Experiment 2 were a little more encouraging, with “only” 77% subjects best fit by random model. However, the effect of including gamma was even smaller (permutation: 77% at chance, irrespective of if gamma is included in the prototype and exemplar models; AIC: 88% best fit by random model, down from 91% without gamma in the prototype and exemplar models).

In general, few subjects improved their fits beyond the improvement would one expect by chance from a model with an extra parameter. To illustrate, below are two figures illustrating the effect of adding gamma to the exemplar model for one subject who was classified as an exemplarist and one subject that was classified as “chance”. The relative goodness of fit remained about the same once we considered that the gamma-including versions also produce better fits into pure noise (random permutations).

Example 1. Recognition exemplar model fit and the subject-specific null distribution for one subject who was best fit by the exemplar model (top: no gamma included; bottom: gamma included). The blue histogram shows the distribution of fit values when modeling randomized (no real signal) data. The red line shows actual fits (fit value and p-value of the fit in the text above the red line).



Example 2: Recognition exemplar model fit and the subject-specific null distribution for one subject who was classified as “chance”. (Prototype model fits also did not exceed chance, but are not shown here). As you can see, the exemplar model fit improved with the addition of the gamma parameter (from raw 33.3 to raw 32.9), but the null distribution also shifted slightly left.



Similarly, when we used the traditional AIC method for model selection, the explicit penalization for the free parameter mostly offsets the small improvements in fits. Although there was a 3% decrease in subjects classified as “chance” by the AIC model selection criterion when we included gamma (from 95% to 92%), these were all subjects who were already classified as not-chance using the permutation test.

To summarize, we found a small improvement in recognition fits when we included gamma, but the improvement was modest and mostly aligned with what one would expect by chance after increasing model flexibility with an additional parameter. We suspect that the poor recognition fits are not just due to mismodeling the behavior. The recognition task is difficult, with the need to remember the unique combination of many binary features. Furthermore, the unequal base rate may have also affected performance. As we noted in response to R2.3., we minimized mention of the recognition data in the main manuscript and moved it instead to supplementary materials, retaining the no-gamma version of both models as seems to be most common when fitting recognition data (Nosofsky & Kantner, 2006; Nosofsky & Zaki, 1998; Nosofsky & Zaki, 2003).

R2.11. p. 20, Figure 2B. Training performance in the small low-coherence condition certainly looks much better than in the large low-coherence condition. Are you sure this effect is not significant?

The set size x coherence effect in Experiment 1 that Dr. Nosofsky points to is marginally significant (p = .08). We did not discuss it in the original submission so as to not make too much of non-significant effects, but given our a priori interest in this interaction and how strong the effect appears in the figure, we have revised the results section to specifically comment on the nature of that effect (pg. 20):

"*The main effect of set size was not significant when we compared the small and large, repetition-matched sets [F(1,113), = 1.13, p = .29, η_p^2 = .01], but there was a marginal benefit to training on small sets when small and large sets were trial-matched [F(1,116), = 3.36, p = .06, η_p^2 = .03]. There was also a marginal set size x set coherence interaction for the trial-matched sets [F(1,116) = 3.12, p = .08, η_p^2 = .03]. Visual inspection of Figure 2B indicates that when coherence is low, there is an advantage of learning from small compared to large sets, with no set size effect for high coherence sets. No other interaction effects were significant or marginally significant (all F's < 1.9, p's > .1, η_p^2 < .02]. Taken together, training data showed better category learning from high coherence than low coherence sets without a strong effect of training set size. There was a strong effect of training set coherence regardless of whether small and large sets were matched in terms of the number of item repetitions or the total number of training trials, with some suggestion that set size may emerge when training is matched in terms of the number of training trials.*"

With that said, that effect did not approach significance in Experiment 2. We include a discussion of this difference across experiments in the revised manuscript (pg. 32-33):

"*In Experiment 1, there were hints of a benefit of learning from small compared to large sets in the low coherence condition, but only when sets were equated in terms of the total number of trials. Here, neither the main effect of set size (F(1,186) = 0.06, p = .80, η_p^2 < .001) nor the set size x coherence interaction (F(1,186) = 0.20, p = .66, η_p^2 = .001) approached significance in the trial-matched comparison. These effects were also not significant when we controlled for the number of item repetitions (set size main effect: F(1,182) = 2.46, p = .12, η_p^2 = .01; set size x set coherence interaction: F(1,182) = .35, p = .55, η_p^2 = .002). Across both trial-matched and repetition-matched comparisons, no other interaction effect approach significance either (all F's < 1.8, p's > .11, η_p^2 < .009). Taken together, there was a clear advantage of learning from high coherence training sets without a strong effect of training set size regardless of whether small* *and large sets were equated for the number of item repetitions or the total number of training trials*."

R2.12. pp. 23-25. I only skimmed through this section, given that the probability-matching exemplar model that is being fitted is not a serious candidate, and given my difficulties in understanding the model-fit shuffling procedure.

As noted above, we have added the response scaling parameter to the exemplar model in the revised manuscript and thus no longer make the probability matching assumption. We have also clarified the shuffling procedure as detailed in R2.9 above. We hope that these revisions will allow Dr. Nosofsky to fully evaluate the modeling findings.

R2.13. p. 35, top. The authors argue that, according to exemplar models, the ability to generalize should be positively related to recognition-memory accuracy. This may or may not be true, and it will all depend on the particular experimental design. The Nosofsky et al. (1988) article that the authors already reference in the article provides examples of all sorts of dissociations between categorization and recognition that are predicted by the exemplar model, e.g., because of the different decision rules used across the two tasks (as well as other factors). Whether the prediction holds true for the present study needs to be evaluated by actually cranking out the predictions from the model.

We thank Dr. Nosofsky for pointing to the exemplar model's flexibility to predict a variety of relationships between categorization and recognition performance. For this and other reasons (very poor recognition that may not carry reliable information about individual differences), we removed the correlation between recognition and categorization from the revised manuscript. However, we remain interested in evaluating that relationship, and we have taken note of this point and will consider it in our future endeavors.

To reiterate, although I have raised a number of serious concerns, I do feel there is much that is excellent in this article, and I hope the authors will consider my criticisms to be constructive ones.

We once again thank Dr. Nosofsky for his careful consideration of our work and fully agree that his comments were constructive. We think that addressing his concerns have improved the manuscript considerably and hope that it will be suitable for publication.

References

Nosofsky, R. M., Sanders, C. A., Zhu, X., & McDaniel, M. A. (2019). Model-guided search for optimal natural-science-category training exemplars: A work in progress. Psychonomic bulletin & review, 26(1), 48-76.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(5), 924.

Zaki, S. R., & Nosofsky, R. M. (2001). A single-system interpretation of dissociations between recognition and categorization in a task involving object-like stimuli. Cognitive, Affective, & Behavioral Neuroscience, 1(4), 344-359.

Nosofsky, R.M., Denton, S.E., Murphy-Knudsen, A.F., & Unverzagt, F.W. (2012). Studies of implicit prototype extraction in patients with mild cognitive impairment and early Alzheimer's disease. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38, 860-880.

Reviewer #3: TITLE
Coherent category training enhances generalization and increases reliance on prototype representations

AUTHORS
Caitlin R. Bowman and Dagmar Zeithamova

SUMMARY
This paper reports the results of a single study with multiple components that compared participants' ability to learn new category sets. The authors manipulated training set size and training set category coherence. They analyzed learning rate, strategy, recognition, and generalization. Their results suggested that high-coherence categories were acquired more readily and they were more likely to encourage prototype abstraction. High-coherence training also encouraged the endorsement of prototypical items in a transfer test phase.

Overall, this paper fits within a rich tradition of research examining concept acquisition and category learning. The authors situate the work within their own recent research as well as within the broader context of earlier research in the same area. I think this paper makes a solid contribution in that regard. My review and suggestions are below, but I am enthusiastic about this paper's eventual publication.

REVIEW

The introduction was well written and concise. It highlighted the incremental nature of this work. It's a great example of building on the team's prior study, connecting it to earlier research, and attempting to push the field forward and generalize. The researchers identify a key problem with prior research: "However, no prior study has addressed the challenge of how test for set size effects while separating them from potential training length or item repetition effects" and designed research to address this. They devised a way to compare the small and large categories on number of repetitions and number of trials.

I like that the stimulus set and data are available on OSF, which is a great resource for modellers and especially for students and trainees learning about this work. The modelling work is also well described, and these are well established procedures. The permutation test and simulations are novel and are described in some of the author's earlier work.

All in all, I really like this paper, it's straightforward and the work would be easy to reproduce. The data are available online, the stimuli are available online, and the results are interesting. I think with some minor changes, this paper could be published in JEP:LMC and it will be well-received in the field.

We appreciate the Reviewer's enthusiasm for the manuscript. As noted above, we made substantial edits to the manuscript, including collecting new data and adding an Experiment 2. Despite these changes, we think that the elements the Reviewer was excited by remain intact, including our manipulation of training length vs. item repetition, the data availability on OSF, and the modelling approach. Below we provide detailed replies to the Reviewer's comments and corresponding changes to the manuscript.

SPECIFIC COMMENTS

R3.1. Page 10 "The training set coherence factor had two values (high and low) that were defined by the average percentage of shared features between training stimuli and their respective prototypes." Other researchers have quantified this differently. "Structural ratio" is one measure that has been used in the past (Homa, 1979); "Entitativity" is another (Haslam, Rothschild, & Ernst, 2000). Of the two, structural ratio is easy to calculate directly, and it might be worth including it here.

We thank the Reviewer for this suggestion. We have computed structural ratios in each condition and added this information to the revised manuscript. We also include a discussion of both entitativity and structural ratio as variability metrics in the discussion section.

Experiment 1 methods (pg. 11-12):

"*Another index used in prior work to describe category coherence is the 'structural ratio': the number of features that differ between items in the same category versus those in opposing categories (Homa et al., 1979). For comparison with this prior work, we provide the mean structural ratio for each training set: small, high coherence = 0.80; small, low coherence = 1.03; large, high coherence (both repetition- and trial-matched) = 0.72; large, low coherence (both repetition- and trial-matched) = 0.98. Higher values on this metric indicate greater within relative to between-category differences among stimuli and are thus larger for the low coherence categories in which individual examples within a category differ significantly from one another while being relatively close to members of the opposing category.*"

Experiment 2 methods (pg. 30):

"*As in Experiment 1, training sets varied in coherence (high and low). All training stimuli in the low coherence sets had 60% typical features, sharing 6 of 10 features with their category prototype. All the training stimuli in the high coherence sets had 80% typical features, sharing 8 of 10 features with their category prototype. The structural ratio (average differing features within vs. between categories) for each training set was as follows: small, high coherence = 0.59; small, low coherence = 1.07; large, high coherence (both repetition- and trial-matched) = 0.54; large, low coherence (both repetition- and trial-matched) = 0.99.* "

Discussion (pg. 40-41):

"*While our findings show a benefit of high coherence training that differs from studies that trained to criterion, they align well with a theoretical prediction based on computational modeling by Hintzman (1984) who argued that coherent training should be beneficial for generalization when the length of training is equated. They also align well with some other manipulations of category coherence. Prior work in the domain of social categorization has also shown that coherence (called 'entitativity' in this domain (Campbell, 1958)) is a key factor that drives naïve perceptions of what defines category membership (Haslam et al., 2000). Thus, part of the high coherence training benefit may be that it fits well with naïve intuitions about how categories are formed, allowing participants to quickly adopt strategies that are well suited to learning categories based on similarity to the category center. Another operationalization of category coherence is the structural ratio: distances between items within the same category compared to the distances between items from different categories (Homa et al., 1979; Minda & Smith, 2001). Prior work has shown faster learning from more coherent, 'well structured'* *category sets in which there is more clustering of items within vs. between categories (Minda & Smith, 2001). That high coherence training produced highly generalizable category knowledge in our study is also consistent with prior work suggesting that stability and consistency in input facilitates broad category knowledge (Carvalho et al., 2019; Horst et al., 2011), and that learning from an easier training set can facilitate later categorization of more difficult items (Edmunds et al., 2019). Finally, the findings also align with our prior aging study showing that when training contains a mixture of typical and atypical items, older adults have difficulty learning the atypical items but their successful acquisition of typical items is sufficient to support subsequent generalization at levels comparable to young adults (Bowman et al., 2022).*"

R3.2. Page 12 "Each exemplar was presented for 2.5 seconds before the response options appeared on the screen" beginning here, and elsewhere in the procedure section, if the authors have any information about their justification for decisions like how long the exemplar appears, how long feedback takes to appear, etc. it would be worth including.

We appreciate the need to justify choices about our experimental procedure and have made revisions to the manuscript to do so:

Experiment 1 training (pg. 13):

"*Each exemplar was presented for 2.5 seconds before the response options appeared on the screen to give participants time to evaluate the high-dimensional stimuli without a need to respond. When the response options appeared, participants were asked to indicate their response with a keyboard press. Response timing was self-paced following the initial display period to ensure that participants made a response and received corrective feedback on every trial. Two seconds following the response, participants were told if their answer was correct or wrong, and to which family the stimulus belonged. The delay in providing feedback was not intentional, but instead an error in the experiment program. Feedback appeared for 1.5 seconds, which piloting showed was typically enough time to read the full feedback text.*"

Experiment 1 categorization (pg. 14):

"*During the categorization test (Figure 1F), individual exemplars were presented, and participants indicated which family they belong to. No feedback was given. Each exemplar was presented on the screen for 4 seconds followed by an 8 second fixation period. The relatively long fixation period was included so that the paradigm would be comparable to recent fMRI studies (Bowman et al., 2020; Bowman & Zeithamova, 2018). There were two blocks of the categorization test, and stimuli were pseudorandomly ordered so that no more than three exemplars from each category were presented consecutively.*"

Experiment 2 (pg. 30-31):

"*We adjusted the trial timing in each phase to make the experiment more efficient and reduce the passive time for the participant. During the feedback-based training, each cartoon animal was presented on the screen for 2 seconds before the response options appeared on the screen and the participant could make a self-paced response. Feedback was displayed for 1.5 seconds immediately after the response, followed by a 1.5 second inter-trial fixation. In the categorization test, test items were presented for 4 seconds as in Experiment 1, but the inter-trial fixation was reduced to 1 second. The order of trials within a block was completely randomized.*"

We also note that, although the timing differed across the two experiments, the pattern of findings in terms of performance was quite similar, showing that the positive effect of training set coherence is relatively robust across these details in the training procedure.

R3.3. Page 18 do the authors have any ideas as to why the prototype and exemplar models fit the recognition data so poorly? The poor fit is acknowledged, but the authors don't discuss the data much beyond that. Why do these models fit so poorly? If participants did not rely on prototype and exemplar representations to make the recognition judgments, how did they make the recognition judgement then? I thought this section could use some additional exploration and improvement/

We agree with the Reviewer that it is interesting that neither the prototype nor the exemplar model fit the data very well and we did not find any systematicity in the responses that would hint to any specific strategy. The newly added Experiment 2 had more subjects fit better than chance, but it was still only about a quarter of the participants.

Importantly, although both models fit poorly, we would not go so far as to claim that participants did not use prototype or exemplar representations during recognition. Both models tended to fit the categorization data, implying that one or both representations were formed and presumably available during the recognition phase. We suspect that the main challenge for finding evidence for one or the other model in the recognition data was the tendency of subjects to endorse stimuli as “old”. In Experiment 1, participants responded ‘old’ on 59% of trials on average, despite only 21-35% of stimuli actually being old (depending on training set size condition). In Experiment 2, an average of 62% of responses were ‘old’, despite only 19-32% of stimuli actually being old. As the bias is also accounted for in the null model (see R1.2 and R2.9 for details on the null distribution generation), the prototype and exemplar models may have difficulty accounting for recognition responses above-and-beyond participants’ overall response bias during recognition. To elaborate, all stimuli feel fairly familiar and it’s difficult to remember the unique combination of the many binary features that differentiate one stimulus from another. For example, it would not be surprising for someone to endorse e.g. stimulus 11101110 after seeing 11001111 and 11111100, leading to a bias to respond ‘old’ to most stimuli. But that also means that simply knowing the response bias (which is included in the null model) already provides a good ground for guessing what the response would be, without any relationship to the stimulus. In other words, knowing the prototype- or exemplar-based similarity did not offer much additional explanatory power beyond knowing how likely a participant was to respond ‘old’ across the entire recognition test. This may mean that participants responded randomly or used a strategy not resembling either model, but it is also possible that the negative finding is driven by the limitation of using the current stimuli for recognition tasks.

As noted above in response to R2.3, we ultimately decided to include the recognition data as a supplement rather than in the main manuscript for space purposes. Nonetheless, we are glad for the opportunity to better discuss the interpretation of the poor recognition fits as part of the newly added supplement (pg. S10-11):

“*Across two experiments, we found that the multidimensional stimuli used in the current experiment posed a challenge for recognition. The vast majority of subjects showed not only poor recognition in terms of the corrected hit rate, but also model fits that were not significantly better than what can be expected by chance. This finding is notable since both models tended to fit the categorization data at above-chance levels, suggesting that prototype and/or exemplar representations were formed by many participants and should have been available for making other types of decisions. Nonetheless, neither representation was apparent in the recognition* *behavior. One possibility is that participants do use prototype and/or exemplar-based similarity in recognition, but that the decision rule in Equation S1 does not accurately reflect the way that participants use similarity information to make their judgments. Another possibility is that the quality of prototype and/or exemplar representations that participants may have formed were not detailed enough to be useful for recognition, which required memory for a unique combination of 8 or 10 features. For example, participants may have seen stimuli with values 1100 and 0011 on the first four dimensions (where 1 would indicate one value of a feature, such as a squared body and 0 would indicate the other value of a feature, such as a round body) but not encode the precise conjunction of the four features. This may have led to strong feelings of familiarity for most of the recognition stimuli and an overall bias to respond 'old', without strong differences based on similarity to prototypes or exemplars. Following the example above, stimuli with values 1100, 1010 or 1000 may all feel equally familiar since each individual feature had been presented many times before but not always in that exact combination. Participants indeed tended to respond 'old' more often than 'new' (response bias towards 'old' response = 59% in Experiment 1, 62% in Experiment 2), even though the majority of test items were new (only 19%-35% old items, depending on the set size and Experiment). Our permutation approach for model selection considered each participant's bias to respond 'old' when generating the subject-specific null distributions. The high proportion of subjects given the 'chance' label therefore indicates that the prototype and exemplar models did not offer additional explanatory power above-and-beyond knowing this response bias. Thus, the effect of training set on recognition and the categorization-recognition relationship may be more conclusively evaluated in future work, perhaps utilizing different stimuli.*"

Coherent category training enhances generalization in prototype-based categories

Caitlin R. Bowman[1,2] and Dagmar Zeithamova[1]

[1] Department of Psychology, University of Oregon

[2] Department of Psychology, University of Wisconsin-Milwaukee

Author Note

Caitlin R. Bowman, https://orcid.org/0000-0002-5833-3591

Dagmar Zeithamova, https://orcid.org/0000-0002-0310-5030

Acknowledgments. This work was supported by the National Institute on Aging Grant F32-AG-054204 awarded to Caitlin R. Bowman and National Institute for Neurological Disease and Stroke Grant R01-NS- 112366 awarded to Dagmar Zeithamova.

Data available through the Open Science Framework:

https://osf.io/snqd5/?view_only=f56e3da5d1a54633aafe4a8ad0ec051b

Stimuli available through the Open Science Framework: osf.io/8bph2

The authors declare no competing financial interests.

Correspondence concerning this article should be addressed to:

Dagmar Zeithamova

dasa@uoregon.edu

1227 University of Oregon

Eugene, OR 97403

Abstract

A major question for the study of learning and memory is how to tailor learning experiences to promote knowledge that generalizes to new situations. In two experiments, we used category learning as a representative domain to test two factors thought to influence acquisition of conceptual knowledge: the number of training examples (set size) and the similarity of training examples to the category average (set coherence). Across participants, size and coherence of category training sets were varied in a fully-crossed design. After training, participants demonstrated the breadth of their category knowledge by categorizing novel examples varying in their distance from the category center. Results showed better generalization following more coherent training sets, even when categorizing items furthest from the category center. Training set size had limited effects on performance. We also tested the types of representations underlying categorization decisions by fitting formal prototype and exemplar models. Prototype models posit abstract category representations based on the category's central tendency, whereas exemplar models posit that categories are represented by individual category members. In Experiment 1, low coherence training led to fewer participants relying on prototype representations, except when training length was extended. In Experiment 2, low coherence training led to chance performance and no clear representational strategy for nearly half of the participants. The results indicate that highlighting commonalities among exemplars during training facilitates learning and generalization and may also affect the types of concept representations that individuals form.

The ability to form new conceptual knowledge is a key function of memory, allowing individuals to organize past experiences and apply them efficiently to new situations. How to tailor learning to best promote acquisition of new conceptual knowledge has been a question of considerable interest not only in cognitive psychology (Hahn et al., 2005; Mervis & Pani, 1980; Williams & Lombrozo, 2010), but also in domains like child development (Ogren & Sandhofer, 2021; Perry et al., 2010; Twomey et al., 2013), linguistics (Bulgarelli & Weiss, 2019; Onnis et al., 2004; Plante et al., 2014), and computer science (Hart, 1968; Hernandez-Garcia & König, 2020; Roiger & Cornell, 1996; Zhou et al., 2017). The answer to this question has practical implications for how instructors select training examples to maximize the generalizability of learning.

Category learning is a well-established concept-learning domain with well-established formal models of how individuals represent categories and generalize to new instances (Bowman & Zeithamova, 2020; Kruschke, 1992; Love et al., 2004; Medin et al., 1978; Nosofsky, 1986; Nosofsky et al., 1994; Posner & Keele, 1968; Smith et al., 1997). There has been a longstanding debate about whether it is better to learn categories from a coherent set of examples that are relatively similar to one another and to the category center or to learn from a more variable set that exposes the learner to the breadth of the category during training (Homa & Cultice, 1984; Homa & Vosburgh, 1976; Minda & Smith, 2001; Peterson et al., 1973; Posner & Keele, 1968). Early work indicated that learning from highly varied exemplars led to better transfer of category knowledge (Posner & Keele, 1968) with some studies showing a particular benefit for generalizing to items near the category boundary (Peterson et al., 1973). This also aligned with early intuition from computer science that machine learning algorithms would benefit from training on instances near decision boundaries (Hart, 1968). However, the empirical studies that showed a benefit for more variable (less coherent) training over more coherent training trained learners to a performance criterion (e.g., Homa & Cultice, 1984; Posner & Keele, 1968), which tended to take much longer for those trained on more variable sets. Modeling work suggested that once the amount of training was equated, the coherent sets would likely lead to better transfer (Hintzman, 1984). More recent empirical studies have also indicated that variability in training exemplars does not lead to better generalization once the amount of training is equated (Bowman & Zeithamova, 2020; Minda & Smith, 2001). For example, our recent study (Bowman & Zeithamova, 2020) tested six training sets that varied in coherence and found increasing accuracy in a subsequent generalization test with increasing coherence, even when we limited analyses to only test items closest to the category boundary. Inclusion of atypical, widely-varying training instances in a training set has been also been shown to lower generalization performance of some machine learning algorithms (Roiger & Cornell, 1996). Nonetheless, the idea that high-variability training is especially beneficial to generalization remains widespread, with recent work claiming that, ‘A consistent finding in the literature is that increasing variability during learning allows people to gain a better understanding of a category, improving classification with novel exemplars.’ (Doyle & Hourihan, 2016, pg. 1197) and a recent review similarly stating the benefits of increased training variability for promoting generalization of knowledge (Raviv et al., 2022).

Besides affecting the accuracy of categorization judgments, the coherence of training examples may affect how categories are represented. Formal models of categorization make different assumptions about the nature of representations underlying category knowledge. The prototype model, illustrated in Figure 1A, posits that categories are represented by their central tendency – an abstract category prototype that includes all the most typical features across the individual category members (Homa & Little, 1985; Minda & Smith, 2002, 2011; Posner & Keele, 1968). Generalization then involves consideration of this prototype. In contrast, the exemplar model, illustrated in Figure 1B, posits that categories are represented by individual category members encountered in the past (Kruschke, 1992; Lamberts, 1994; Medin et al., 1978; Nosofsky, 1988). Generalization involves jointly considering the individual members of relevant categories rather than an abstraction. Our prior study showed that, along with increasing generalization accuracy, participants trained with more coherent examples were more likely to rely on prototype representations during generalization, suggesting that coherence benefits learning by facilitating prototype extraction (Bowman & Zeithamova, 2020).

The history of debate and mixed findings around the direction of the coherence effect leaves open questions about how coherence interacts with other aspects of training that differ across studies. One candidate factor is the number of training examples (set size). Like set coherence, the effect of training set size on subsequent generalization has been somewhat equivocal across past studies, particularly regarding whether set size affects reliance on prototype vs. exemplar representations. As suggested by Minda and Smith (2001), studies that have found better fit of the exemplar model than the prototype model tended to show only a small number of examples during training (Blair & Homa, 2003; Lamberts, 1994; Medin et al., 1978). This exemplar model advantage may arise due to the relative ease of encoding only a few items into memory compared to when the number of training examples is large. Adding low coherence to small set sizes may further promote exemplar representations because having less overlap among items makes it easier to individuate them in memory. Others have shown better categorization accuracy for larger training sets (Goldman & Homa, 1977; Homa et al., 1973, 1981), accompanied by better fit of the prototype model (Minda & Smith, 2001). High coherence within these large sets may be particularly well suited to promoting prototype representations because they offer ample signal to derive the category average, while exemplar-based encoding may be hindered by the difficulty of distinctively encoding many similar examples. Since either prototype or exemplar representations could support successful learning and generalization, both large, high coherence and small, low coherence training may facilitate broad category knowledge.

Despite this theoretical motivation for an interaction between training set size and coherence, empirical evidence is lacking. Studies using continuous-dimension stimuli have shown that large training set sizes do not always lead to reliance on prototype representations. While the prototype model predicts a linear decision bound between categories, participants trained on many unique exemplars sampled from two bivariate normal distributions can learn and adopt an optimal, non-linear category boundary, indicating they must be representing the variability of each category in addition to its center (Ashby & Gott, 1988; Ashby & Maddox, 1992; McKinley & Nosofsky, 1995). Studies using binary dimension stimuli also have not shown clear support for any one effect of training set size. Minda and Smith (2001) tested both set size and set coherence, but did so across separate experiments and thus could not measure their combined effect. In a prior study (Bowman & Zeithamova, 2020), we showed that higher training set coherence tended to lead to greater reliance on prototype representations without strong effects of training set size. However, set size and coherence were somewhat correlated, which meant that we could have missed a benefit of low coherence training if it were present only for some training set sizes. Homa and Vosburgh (1976) directly tested the interaction in terms of generalization abilities and found evidence *against* the hypothesis that large, high coherence training and small, low coherence training promote generalization. Instead, they showed a generalization advantage for coherent training only when the set size was small (3 items). For large set sizes (6 or 9 items), lower coherence resulted in better generalization. However, this study did not assess the representations underlying generalization judgments, and subjects were trained to criterion prior the generalization test. Training to criterion rather than keeping the training fixed across conditions likely resulted in much longer training for low coherence than high coherence sets, and potentially also longer training for larger set sizes than the small set size.

Indeed, a key challenge in assessing training sets of different sizes is that one can either equate the number of repetitions of each item during training or equate the total number of training trials between conditions but cannot equate both simultaneously. For example, one condition in our prior study included a small set with 10 total training items and one condition included a large set with 20 total training items (Bowman & Zeithamova, 2020). We equated the training for all sets in terms of the number of repetitions of each individual training item (16 presentations per item), which led to 160 training trials for the smaller set and 320 trials for the larger set. We found minimal effects of set size, but the difference in the total number of training trials could have masked differences between conditions if it is more difficult to learn from large sets but additional training trials in this condition compensated for that difficulty. While one can instead decrease the number of repetitions of individual items for large sets in order to equate the total number of trials, this approach also comes with potential drawbacks. Namely, the representational quality for individual items will likely depend on how many times they are presented during training, leading to poorer quality item representations for conditions with fewer presentations of each item. The quality of item representations is likely to affect exemplar-based categorization judgments as well as other types of memory judgements like old/new recognition. However, no prior study has addressed the challenge of how to test for set size effects while separating them from potential training length or item repetition effects.

In the present study, we conducted two experiments that aimed to disentangle the effects of training set coherence, set size, and amount of training in order to better understand the conditions that facilitate category learning and generalization, as well as the types of representations that such training conditions foster. Participants learned to classify novel cartoon animals into two categories (Romeo’s family and Juliet’s family; Figure 1D). Participants were randomly assigned to learn from training sets that differed in the number of examples per category (small or large) and the set coherence (high or low). These two factors were fully crossed, allowing us to test whether any effect of set coherence depended on set size. We also equated small and large set sizes in two ways: for the number of repetitions of each training item and the total number of training trials. Following training on one of these training sets, participants completed a category generalization task, which was structured identically across all training groups. We compared training regimes to determine how learning from small vs.

large and high coherence vs. low coherence sets affects the ability to learn and generalize category labels and the types of representations underlying categorization decisions.

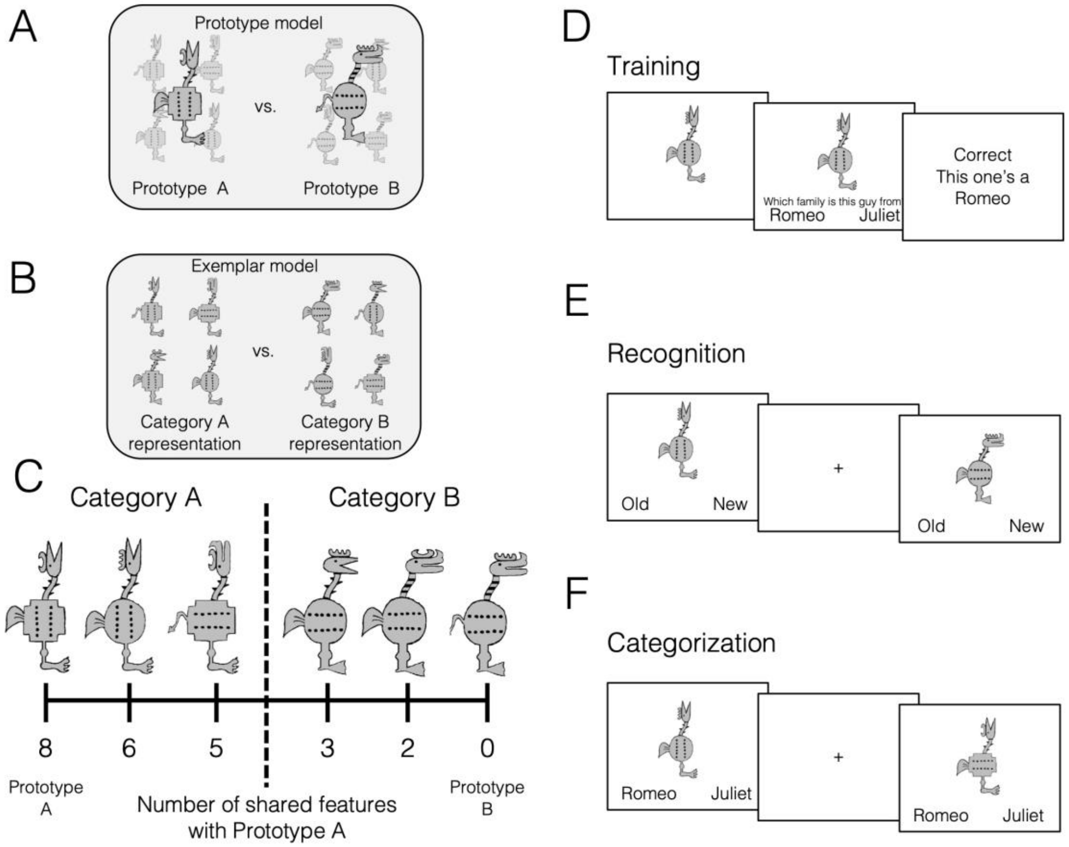


Figure 1. Category-learning task. Category representations under the assumptions of the **A.** prototype model and **B.** exemplar model. Prototype: categories are represented by their central tendencies (prototypes). New items are classified into the category with the most similar prototype. Exemplar: categories are represented as individual exemplars. New items are classified into the category with the most similar exemplars. **C.** Example stimuli. The leftmost stimulus is the prototype of category A and the rightmost stimulus is the prototype of category B, which shares no features with prototype A. Members of category A share more features with prototype A than prototype B and vice versa for members of category B. **D**. Participants underwent feedback-based training with one of six possible training sets that varied the size of the training set and the coherence of the examples. **E.** In recognition, participants were shown training (old) items and never seen category members and made old/new judgments (see Supplemental Materials for detailed methods and results). **F.** In categorization, participants were shown training (old) items and never seen category members and made categorization judgments without feedback.

Experiment 1

In this experiment, we tested the relative roles of training set size and set coherence on category learning and generalization using cartoon stimuli with 8 binary dimensions. We generated training sets consisting of items sharing 6 of 8 features with their category prototype (high coherence training) or items sharing 5 of 8 features with their category prototype (low coherence training). The training sets could either be small (containing 4 items per category) or large (containing 8 items per category). As it is not possible to equate large and small set conditions for both the number of repetitions per item and the total number of training trials, there were two large set conditions: one equated with small sets in terms of the number of repetitions of each item (large, repetition-matched) and one equated with the small sets in terms of the total number of training trials (large, trial-matched). By comparing small training sets to two types of large sets, we could determine whether any effect of set size was 1) driven by differences in the amount of training, 2) driven by differences in the amount of exposure to individual examples, or 3) consistent regardless of which aspect of training we equated across set sizes. Once crossed, these variables created six between-subjects training conditions. Following training, participants were tested on their ability to differentiate between old and new category examples in a recognition test and their ability to generalize category labels to new examples. Our main analyses focus only on the data from the category training and generalization phases, and the detailed methods and results from the recognition phase are available in the Supplemental Materials.

Method

Participants

One hundred eighty-three participants from the University of Oregon completed the experiment for course credit. Prior to data collection, we aimed to recruit approximately 30 participants for each between-subjects condition (total ~180) based on the rule of thumb from

Central Limit Theorem. Five participants were subsequently excluded due to missing data in some portion of the experiment, and one other participant was excluded for failing to respond on ~60% of categorization trials, leaving data from 177 participants reported in all analyses (119 female, mean age = 19.97 years, SD age = 2.38 years, range 18-35 years). This final sample size is sufficient to detect a set size x set coherence interaction effect of η_p^2 > .055 with 80% power. Participants were randomly assigned to one of six training groups (see Table 1 for demographic information separated by training condition). All participants completed written informed consent. All procedures were approved by the University of Oregon's Institutional Review Board.

Table 1

Experiment 1 training sets and demographic information

Training Set	n females / n males	Mean age (SD age, age range)
Small, high coherence	23 / 7	20.2 (2.7, 18-30)
Small, low coherence	23 / 7	19.2 (1.5, 18-24)
Large, high coherence (repetition-matched)	21 / 9	19.3 (1.4, 18-23)
Large, low coherence (repetition-matched)	14 / 13	20.4 (2.5, 18-26)
Large, high coherence (trial-matched)	20 / 10	20.6 (3.2, 18-35)
Large, low coherence (trial-matched)	18 / 12	20.2 (2.4, 18-28)

Materials

The complete stimulus set is freely available through the Open Science Framework (https://osf.io/8bph2). Stimuli consisted of cartoon animals that varied along 8 binary dimensions: foot shape (clawed/webbed), body shape (squared/circular), tail shape (devil tail/feather tail), body dot orientation (vertical/horizontal), neck pattern (stripes/thorns), head shape (with beak/with horn), crown shape (crescent/comb), and head orientation (forward/up). One stimulus was chosen randomly for each subject from a set of four possible prototypes to be the prototype of category A. The stimulus that shared no features with the category A prototype served as the category B prototype. The two possible versions of each feature can be seen across the two prototypes shown on Figure 1C. Physical similarity between all pairs of stimuli was defined based on their number of shared features. Stimuli with 5-8 prototypical A features were considered category A members, stimuli with 0-3 prototypical A features (thus 5-8 prototypical B features) were considered category B members. Items equidistant from the prototypes (sharing 4 features with each) were not used in any phase of the experiment.

Training sets. Training sets were created that varied in two main factors of interest: training set size and training set coherence. Training set size had two possible values (small or large) and was defined by the number of individual category exemplars from each category presented during training. 'Small' sets including 4 training items per category (8 unique training instances) and 'large' sets including 8 training items per category (16 unique training instances). The training set coherence factor also had two values (high and low) and were defined by the average percentage of shared features between training stimuli and their respective prototypes. All the training stimuli in the 'low coherence' sets had 63% typical features, sharing 5 out of 8 features with their category prototype and differing on 3 features from their respective prototype. All the training stimuli in the 'high coherence' sets had 75% typical features, sharing 6 out of 8 features with their category prototype and differing on 2 features from their respective prototype. Another index used to describe category coherence in prior work is the 'structural ratio': the number of features that differ for items in the same category versus those in opposing categories (Homa et al., 1979). For comparison with this prior work, we provide the mean structural ratio for each training set: small, high coherence = 0.80; small, low coherence = 1.03; large, high coherence (both repetition- and trial-matched) = 0.72; large, low coherence (both repetition- and trial-matched) = 0.98. Values on this metric are higher for the low coherence compared to high coherence categories because, for low coherence training sets, individual examples within a category differ significantly from one another while also being closer to members of the opposing category.

In addition to training set coherence and training set size, the training sets also differed in the total number of training trials. In the small set size condition, each of the 8 training stimuli (4 per category) was repeated 16 times across training, for a total of 128 training trials. We then included two large set size conditions because it is not possible to simultaneously equate the number of repetitions per training item and the number of total trials when increasing the training set size from 4 to 8 items per category. Large, ‘repetition-matched’ sets included the same number of repetitions of each training item as the small training sets (16 repetitions per training stimulus), leading to equivalent exposure to individual training items but more total training trials (256 total training trials). Because differences in the total number of training trials could mask any effect of set size on subsequent categorization, we also included large, ‘trial-matched’ sets with half as many repetitions of individual items as the small training sets. This procedure leads to less exposure to each training item (8 repetitions per stimulus) but equates the total number of training trials between large, trial-matched sets and small training sets (both 128 total trials). Including both versions of the ‘large set size’ condition allowed us to test the effect of training size set while controlling for the potential role of differential number of repetitions of individual items and the potential role of differential total amount of training.

Combining the training set coherence (high, low) and training set size (small, large and repetition matched, large and trial matched) factors created the following six training groups: small, high coherence training; small, low coherence training; large and repetition-matched, high coherence training; large and repetition-matched, low coherence training; large and trial-matched, high coherence training; and large and trial-matched, low coherence training. We did not use the prototypes themselves or items sharing 7 of 8 features with prototypes in any of the training sets. Training set structures for each set size x set coherence condition are in the Appendix.

Categorization stimuli. In addition to old (training) items that differed based on the initial training condition, the categorization test included 30 new stimuli: two prototypes, 8 new items sharing 7 out of 8 features with their category prototype (the remaining 8 items sharing 7 out of 8 features with their category prototype were used for recognition test, see Supplemental Materials), 10 new items sharing 6 out of 8 features with their category prototype, and 10 new items sharing 5 out of 8 features with their category prototype. Half of the stimuli at each prototypicality level were from category A and half were from category B. Importantly, the novel test stimuli had the same structure across all conditions, allowing us to compare all conditions in terms of generalization success to both typical and near-boundary category members.

Procedure

Participants completed the three phases of the experiment in the following order: training (Figure 1D), recognition (Figure 1E), and categorization (Figure 1F), with self-paced breaks in between. In each trial of the feedback-based training (Figure 1D), an individual exemplar was presented on the screen and participants were instructed to decide which of two families (Romeo's or Juliet's) it belonged to. Participants were told that they would have to start by guessing, but that it was possible to learn to sort the items accurately over time. Each exemplar was presented for 2.5 seconds before the response options appeared on the screen to give participants time to evaluate the high-dimensional stimuli without a need to respond. When the response options appeared, participants were asked to indicate their response with a keyboard press. Response timing was self-paced following the initial display period to ensure that participants made a response and received corrective feedback on every trial. Two seconds following the response, participants were told if their answer was correct or wrong, and to which family the stimulus belonged. The delay in providing feedback was not intentional, but instead an error in the experiment program. Feedback appeared for 1.5 seconds, which piloting showed was typically enough time to read the full feedback text.

For small sets and large sets that were repetition-matched, training trials were split evenly across 8 blocks. For large sets that were trial-matched, training trials were split evenly across 4 blocks. Each block consisted of two presentations of each training item, pseudorandomly ordered so that no more than three exemplars from the same category were presented consecutively and each training item was presented once in the first half and once in the second half of a block, in a new pseudorandom order. All groups took self-paced breaks in between blocks. Immediately after training, participants completed an old/new recognition test (Figure 1E). For conciseness and because recognition performance was not consistently above chance, we have reported the detailed methods and results of the recognition test in the Supplemental Materials.

During the categorization test that followed recognition (Figure 1F), individual exemplars were presented, and participants indicated which family they belong to. No feedback was given. Each exemplar was presented on the screen for 4 seconds followed by an 8 second fixation period. The relatively long fixation period was included so that the paradigm would be comparable to recent fMRI studies (Bowman et al., 2020; Bowman & Zeithamova, 2018). There were two blocks of the categorization test, and stimuli were pseudorandomly ordered so that no more than three exemplars from each category were presented consecutively.

Statistical analyses

Training accuracy. Training accuracy was computed as the proportion of correct classifications for each block of training. The different number of training trials and blocks across groups prevented all three set size groups from being compared simultaneously. Instead, we computed two separate 2 (set coherence: high, low) x 2 (set size: small, large) ANOVAs: one comparing small sets to large, repetition-matched sets and one comparing small sets to large, trial-matched sets. For the ANOVA including large, repetition-matched sets, training accuracy was computed across blocks 1-8 for both the small and large sets. For the ANOVA including large, trial-matched sets, we averaged over pairs of blocks (e.g., blocks 1-2, 3-4, etc.) for the small sets so that both groups had 4 training scores and each score included 32 trials.

Categorization accuracy. Overall accuracy was computed as the proportion of correct classifications. We were interested in how training conditions influenced the ability to generalize broadly across the category space. Thus, we included test item typicality (sharing 5-8 features with the respective category prototype) as a within-subjects factor when computing a set size x set coherence ANOVA on generalization scores (accuracy for new items).

Across training and categorization ANOVAs that included a within-subject factor, Greenhouse-Geisser corrections for violations of the sphericity assumption were applied as needed (denoted with ‘GG’). A Bonferroni correction for multiple comparisons was applied when multiple independent statistical tests were computed, such as when conducting a separate t-test for each group or following an omnibus ANOVA with multiple pairwise comparisons.

Prototype and exemplar model fitting in categorization. Although categorization performance analyses may reveal better accuracy of prototypes and the items most similar to prototypes, such prototypicality effects in categorization could result from either prototype or exemplar representations (Medin et al., 1978; Nosofsky, 1986). Thus, as in our prior study (Bowman & Zeithamova, 2020), we fit formal prototype and exemplar models to trial-by-trial categorization test data in individual subjects to estimate participants’ categorization strategies.

Prototype similarity**.** The conceptual representation of the prototype model is depicted in Figure 1A. Prototype models assume that categories are represented by their prototypes (i.e., the combination of typical category features from all training items in each category). Consistent with prior modeling literature (Maddox et al., 2011; Minda & Smith, 2001), the similarity of each test stimulus to each prototype was computed, assuming that perceptual similarity is an exponential decay function of physical similarity (Shepard, 1957) and taking into account potential differences in attention to individual features since subjects may not attend to all features with the present high-dimensional stimuli (Nosofsky et al., 2012; Zaki & Nosofsky, 2001). Formally:

(1) $$S_A(x) = \exp\left[-c \sum_{i=1}^{m} (w_i |x_i - proto_{Ai}|^r)^{1/r}\right]$$

where $S_A(x)$ is the similarity of item x to category A, x_i represents the value of the item x on the i-th dimension of its *m* binary dimensions (m=8 in this experiment), $proto_A$ is the prototype of category A, r is the distance metric (fixed at 1 for city-block metric for the binary-dimension stimuli), w is a vector with weights for each of the 8 stimulus features with weight values estimated from the data (fixed to sum to 1), and c is sensitivity (rate at which similarity declines with distance), also estimated from the data (constrained to be 0-100).

***Exemplar similarity*.** The conceptual representation of the exemplar model is depicted in Figure 1B. The exemplar model assumes that categories are represented by their exemplars, and that summed similarity across category exemplar drives exemplar-based decision-making. Formally (Nosofsky, 1987; Zaki et al., 2003), similarity of an item x to category A is computed as:

(2) $$S_A(x) = \sum_{y \in A} \exp\left[-c \sum_{i=1}^{m} (w_i |x_i - y_i|^r)^{1/r}\right]$$

where y represents the individual training stimuli from category A, and the remaining notation and parameters as in Equation 1.

***Parameter estimation*.** For both models, the probability of assigning a stimulus *x* to category A is equal to the similarity to category A divided by the summed similarity to categories A and B, with the response scaling factor gamma:

(3) $$P(A|x) = \frac{S_A(x)^\gamma}{S_A(x)^\gamma + S_B(x)^\gamma}$$

The response scaling factor can account for participants who respond more deterministically than predicted by probability-matching assumptions (Nosofsky & Zaki, 2002). In the case of the prototype model, gamma is fixed at 1 since it cannot be estimated separately from the sensitivity parameter (*c*). For the exemplar model, the gamma parameter was constrained to be 0-100 to match the constraint from the sensitivity parameter. In our prior work (Bowman & Zeithamova, 2020), we did not include the response scaling parameter in the primary results that were presented in tables and figures, but we did ensure that inclusion/exclusion of that parameter did not drive our findings. Here, we have included the gamma parameter because it is the most commonly used formulation of the exemplar model but again verified that the inclusion/exclusion does not drive any conclusions.

For each trial, the probability of the participant's response under the assumptions of each model was computed. An error metric (negative log-likelihood of the whole sequence of responses) was then computed for each model by summing the negative of log-transformed probabilities. This summed value was minimized by adjusting the attention weights, sensitivity parameter, and response scaling parameter (exemplar model only) using standard maximum likelihood methods with the "fminsearch" function in MATLAB (Mathworks, Natick, MA). Parameters were optimized separately for each model (prototype/exemplar) and for each participant.

Determining participants' strategies using a permutation approach for model comparison. After optimization, we used Monte Carlo simulations to determine whether the prototype and exemplar models each fit better than chance and to determine whether the difference in prototype and exemplar model fits was greater than would be expected by chance (Bowman et al., 2020; Bowman & Zeithamova, 2018, 2020). To generate the subject-specific null distributions, we used the stimuli that the subject encountered and the subject's actual responses, but we randomly shuffled the stimuli and the responses with respect to each other. This produced randomized datasets in which there is no real relationship between the stimuli and the responses while maintaining any potential overall response biases. We then fit the prototype and exemplar models to this randomized data and stored the resulting prototype and exemplar model fit values. This procedure was repeated 10,000 times to generate a subject-specific null distribution of model fits for each model. The null distributions provide information about the typical range of model fit values that happen just by chance when the underlying data contain no real signal.

We then compared each subject's observed prototype and exemplar model fits to their subject-specific prototype and exemplar null distributions to determine whether one or both models fit the participant's data better than chance at alpha = .05 (one-tailed). For example, a subject's observed exemplar model fit would be considered better than chance if fewer than 5% of the random simulations (from their exemplar null distribution) produced fits that were as good or better as the observed data fit. Participants for whom neither model outperformed chance were labeled as responding randomly ("chance").

For participants in which at least one of the models outperformed chance, we then directly compared prototype and exemplar model fits to one another using a relative difference in fits metric: (exemplar model fit – prototype model fit) / (exemplar model fit + prototype model fit). To determine whether one model reliably outperformed the other, we compared their observed difference in model fits to the null distribution of differences in model fits generated from the randomized data. One model was deemed a better fit than the other for a given participant when that difference score appeared by chance with a frequency less than 25% (75% probability that the model fit differences did not arise by chance, two-tailed test). Using this method, participants were labeled as prototype-users or exemplar-users when one model outperformed the other or as having "similar" model fits when neither model outperformed the other. We chose this alpha level for labeling participants' strategies, as in our prior study (Bowman & Zeithamova, 2020), as a compromise between a strict alpha level of 5%, which labeled many subjects as showing similar fits between the two models, and the most common no-alpha model selection approaches, such as when AIC is used to select the best fitting model based on lower fit error (or lower value of the AIC metric that accounts for the number of free parameters) without addressing whether the difference in fits is reliably above chance. It is important to consider what constitutes a significant difference in model fits because, in practice, one model can systematically outperform the other even for randomly generated data, suggesting that one model may be more flexible (in general or for a given category structure or stimulus set). Using the simulation approach can account for this bias when comparing the models and setting a threshold for selecting one model over the other. Nevertheless, we verified that the majority of participants received the same strategy label using either approach (permutation test or AIC).

Comparing categorization strategies across conditions. After determining each participant's best fitting model, we tested whether categorization strategy use varied across conditions. Because a larger proportion of participants were classified as using a prototype strategy than an exemplar strategy, we combined "exemplar" and "similar fit" participants into a single bin and used a binary logistic regression with prototype strategy as one outcome and the combination of exemplar and similar fit as the second outcome. Predictor variables were set size and coherence, where the set size variable was separated into two dummy coded predictors, one for the large, repetition-matched set size and one for the large, trial-matched set size, with small sets as the reference group for each. We also included an interaction term for each set size variable x set coherence to test whether the magnitude or direction of any set coherence effect differed based on the training set size.

Data availability

Data for the present study is freely available through the Open Science Framework at https://osf.io/snqd5/?view_only=f56e3da5d1a54633aafe4a8ad0ec051b, and stimuli are available at https://osf.io/8bph2.

Results

Training accuracy

Accuracies across the training phase separated by training group are presented in Figure 2. We computed two set size x set coherence x block mixed-factors ANOVAs on training accuracy. One ANOVA compared small sets and large sets that were repetition-matched (Figure 2A) and one ANOVA compared small sets and large sets that were matched for the total number of training trials (Figure 2B). We used a Bonferroni corrected α = .025 to account for the two separate ANOVAs. The effects that were similar for the two ANOVAs are reported jointly for brevity. First, there was a significant main effect of block (both F's > 28, p's < .001, η_p^2 > .20) along with a significant linear effect of block (both F's > 81, p's < .001, η_p^2 > .42), driven by increasing accuracy as training progressed. There was a significant main effect of set coherence (both F's > 16, p's < .001, η_p^2 > .12) with higher overall training accuracy for those trained on high coherence sets (repetition-matched M = .71, SD = .10; trial-matched M = .69, SD = .11) compared to those trained on low coherence sets (repetition-matched M = .64, SD = .11; trial-matched M = .61, SD = .11). The main effect of set size was not significant when we compared the small and large, repetition-matched sets [*F*(1,113), = 1.13, *p* = .29, η_p^2 = .01], but there was a marginal benefit to training on small sets when small and large sets were trial-matched [*F*(1,116), = 3.36, *p* = .06, η_p^2 = .03]. There was also a marginal set size x set coherence interaction for the trial-matched sets [*F*(1,116) = 3.12, *p* = .08, η_p^2 = .03]. Visual inspection of Figure 2B indicates that when coherence is low, there is an advantage of learning from small compared to large sets, with no set size effect for high coherence sets. No other interaction effects were significant or marginally significant (all F's < 1.9, p's > .1, η_p^2 < .02]. Taken together, training data showed better category learning from high coherence than low coherence sets without a strong effect of training set size. There was a strong effect of training set coherence regardless of whether small and large sets were matched in terms of the number of item repetitions or the total number of training trials, with some suggestion that a set size effect may emerge when training is matched in terms of the number of training trials.

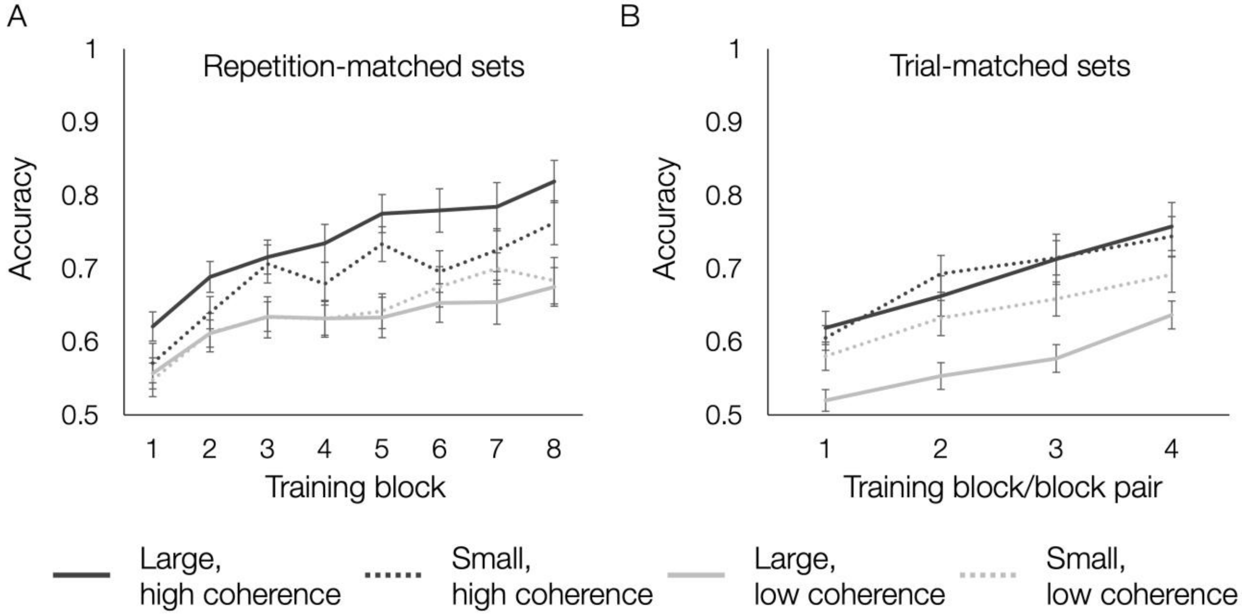


Figure 2. **A**. Training accuracy by block for the small set size conditions (dotted lines) and large set size conditions (solid lines) that matched the small set size in terms of the number of repetitions of each item. High-coherence conditions in black, low-coherence conditions in gray. **B.** Same as A, but depicting the large set size conditions (solid lines) that matched the small set size in terms of the total number of trials during training. For small training sets, the same data are depicted in A and B, but collapsed from 8 blocks of 16 trials each in A into 4 block-pairs of 32 trials in B to match 32 trials per block in the large set size (trial matched) conditions. Error bars depict the standard error of the mean across subjects.

Categorization test accuracy

See Figure 3 for categorization accuracy separated by training group and test item type. To test whether training set size and/or coherence influenced subsequent categorization accuracy, we computed a 3 (set size: small, large and repetition-matched, large and trial-matched) x 2 (set coherence: high, low) x 4 (test item typicality: 5-8 prototypical features) mixed-factors ANOVA. There was a significant main effect of item typicality [*F*(2.7,464.2) = 94.54, *p* < .001, η_p^2 = 0.36, GG] accompanied by a significant linear effect [*F*(1,171) = 201.18, *p* < .001, η_p^2 = .54] driven by better categorization for items closest to the prototypes and decreasing accuracy towards the category boundary. The main effect of training set size was not significant

[*F*(2,171) = 0.22, *p* = .81, η_p^2 = .003] nor were there any significant interactions between set size and the other factors (all F's < 1.6, p's > .16, η_p^2 < .02).

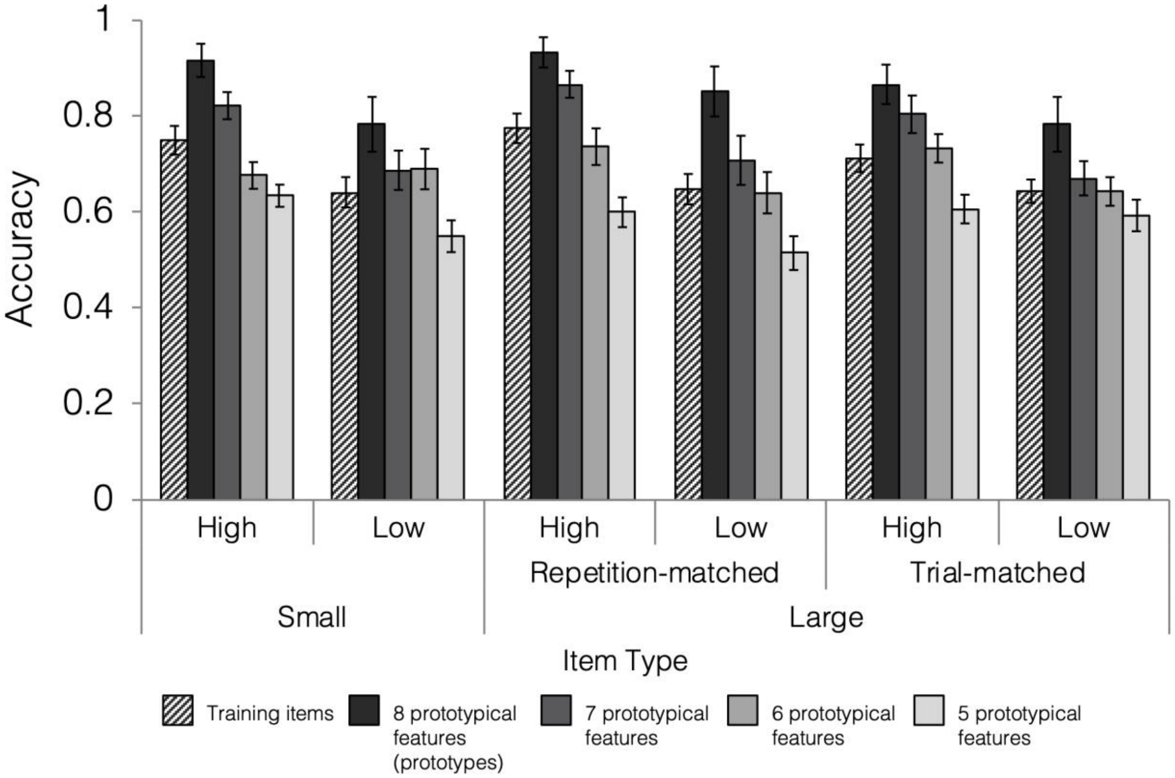


Figure 3. Categorization accuracy for each item type separated by training set. Small = sets with 4 items/category. Large = sets with 8 items/category. High = high coherence: 6 prototypical features for all training examples. Low = low coherence: 5 prototypical features for all training examples. Accuracies on training (old) items re-presented during the categorization test are depicted with striped bars. Accuracies for new items varying in their similarity to category prototypes are depicted with solid bars. Accuracies for prototypes are depicted in the darkest bars with increasingly lighter bars for new items sharing fewer features with the prototypes. Error bars depict standard error of the mean across subjects.

There was, however, a significant main effect of training set coherence [*F*(1,171) = 14.80, *p* < .001, η_p^2 = .08] with better accuracy for those trained on high coherence sets (M = .77, SD = .15) compared to low coherence sets (M = .68, SD = .16). Training set coherence also interacted significantly with test item typicality [*F*(2.7,464.2) = 2.79, *p* = .046, η_p^2 = .02, GG]. This interaction is depicted in Figure 4. To better understand the nature of this interaction, we compared high and low coherence training groups (collapsed across training set size groups) in terms of categorization accuracy at each level of test item typicality using two-sample t-tests. We used a Bonferroni-corrected $\alpha = .0125$ to account for the four separate tests. There was an advantage for high compared to low coherence training at each level of test item typicality, but the advantage was larger and significant for prototypes [t(175) = 2.66, p = .008, d = 0.40] and items with 7 prototypical features [t(175) = 4.68, p < .001, d = 0.70], whereas the difference between groups did not pass correction for multiple comparisons for items with 6 prototypical features [t(175) = 1.97, p = .05, d = 0.30] and 5 prototypical features [t(175) = 2.36, p = .02, d = 0.35]. Thus, the advantage for learning from a high coherence training set was smaller for items closer to the category boundary compared to items close to prototypes. Notably, however, there was never an *advantage* for the low coherence training set, which is inconsistent with the hypothesis that high variability among examples helps promote breadth of category knowledge.

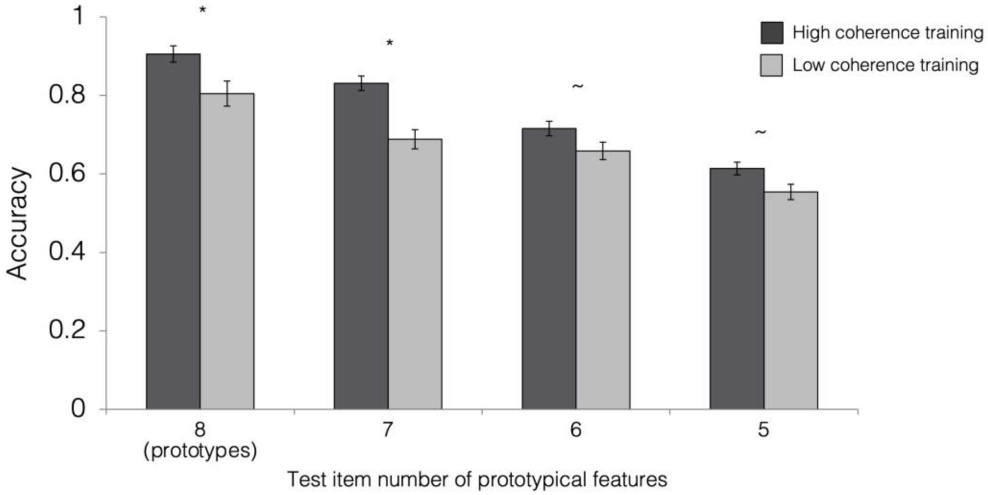


Figure 4. Training set coherence x test item typicality interaction effect in generalization. Test item typicality is depicted on the x-axis with the prototypes on the far left and the items closest to the category boundary (5 prototypical features) on the far right. Dark bars depict generalization accuracy for those trained with high coherence sets (6 prototypical features for all training examples), and light bars depict generalization accuracy for those trained on low coherence sets (5 prototypical features for all training examples). Both coherence groups were collapsed across set size conditions. Stars indicate differences between groups of p < .05 corrected, and tildes indicate differences between groups of p < .05 uncorrected. Error bars depict the standard error of the mean across subjects.

Prototype and exemplar model fits in categorization

As both prototype and exemplar representations can produce typicality gradient in accuracy scores, we used formal categorization models to estimate the representations underlying categorization judgments. The proportion of participants best fit by each model in each condition is presented in Figure 5. Table 2 presents the average best fit values for each estimated parameter separated by model and by training group. Although we did not constrain the models to return the same or similar parameter estimates, there was general agreement between the prototype and exemplar models in terms of the estimated attention weights (average within-subject Pearson's r = .66). The remaining parameters were c (sensitivity) and gamma (response scaling) for the exemplar model and c (combining both sensitivity and response scaling effects in a single parameter) for the prototype model. The c parameter from the prototype model was more correlated with the gamma parameter from the exemplar model (across-subject Pearson's r = .29) than with the c parameter from the exemplar model (across-subject Pearson's r = .06). This suggest that the addition of the gamma parameter to the exemplar model altered how the parameters were used to account for behavior.

Table 2 – Mean best fitting parameter values for the exemplar and prototype models in categorization

	Exemplar model										Prototype model								
Training set	C	G	W1	W2	W3	W4	W5	W6	W7	W8	C	W1	W2	W3	W4	W5	W6	W7	W8
Small, high coherence	10.0	36.2	.20	.15	.06	.21	.11	.08	.10	.10	35.3	.21	.14	.06	.18	.09	.08	.12	.12
Small, low coherence	27.7	37.5	.14	.11	.12	.15	.07	.12	.16	.12	17.4	.17	.12	.17	.07	.16	.06	.08	.17
Large, high coherence, repetition-matched	21.2	41.4	.23	.08	.08	.12	.14	.15	.12	.07	34.5	.21	.07	.10	.13	.13	.18	.11	.06
Large, low coherence, repetition-matched	17.6	52.5	.11	.06	.10	.15	.08	.15	.24	.11	20.6	.22	.09	.18	.07	.14	.10	.08	.11
Large, high coherence, trial-matched	11.9	58.2	.14	.14	.12	.19	.16	.10	.07	.09	19.6	.15	.13	.13	.18	.14	.10	.06	.09
Large, low coherence, trial-matched	31.5	.51.6	.16	.05	.11	.15	.02	.24	.22	.05	17.0	.24	.10	.19	.06	.08	.08	.12	.12

C = sensitivity, G = gamma (exemplar model only), W1-8 = attention weights to each of 8 stimulus features

Our first step in determining the best fitting model for each subject was to compare each subjects' model fits to their respective null distributions to determine whether the fit of each model was better than would be expected by chance. Across the entire sample, regardless of training condition, 140 subjects (79%) showed an exemplar model fit that reliably outperformed chance, and 143 subjects (81%) showed a prototype model fit that reliably outperformed chance. Thirty-one subjects (18%) showed prototype and exemplar model fits that did not outperform chance and were thus labeled with 'chance' for their strategy. Comparing prototype and exemplar model fits to each other in the 146 subjects where at least one model outperformed chance, 35 subjects (24%) were best fit by the exemplar model, 93 subjects (64%) were best fit by the prototype model, and 18 subjects (12%) had comparable exemplar and prototype fits and were labeled with 'similar fits' for their strategy.

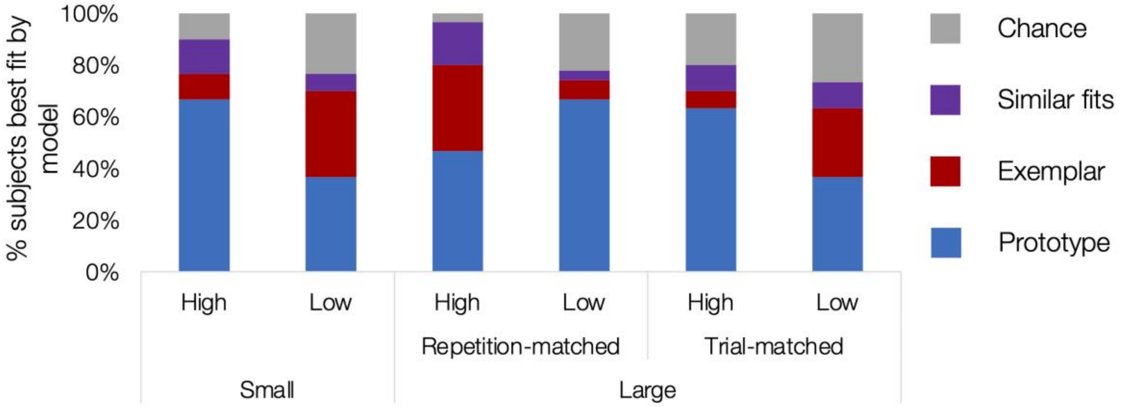


Figure 5. Prototype and exemplar model fits to categorization data. Small = sets with 4 items/category. Large = sets with 8 items/category. High = high coherence: 6 prototypical features for all training examples. Low = low coherence: 5 prototypical features for all training examples. Subjects better fit by the prototype than exemplar model are indicated in blue. Those better fit by the exemplar than the prototype model are indicated in red. Those will comparable fits across models are in purple ('similar fits'). Those whose fits did not outperform the random model are in grey. Model fits are separated by training set typicality, set size, and whether large sets were matched to small sets in terms of the number of repetitions of each item (repetition-matched) or the number of total training trials (trial-matched).

To determine whether training set size or coherence affected the number of prototype users during categorization, we computed a binary logistic regression with set size (small, large repetition-matched, large trial-matched), set coherence (high, low) and their interaction as categorical predictors of whether individuals relied on a prototype strategy, excluding the ‘chance’ participants. Only two levels of the outcome variable were used (prototype strategy, not a prototype strategy) because counts in the exemplar and similar fit bins were too low to keep separate. This model explained 14% of variance (Nagelkerke R^2). There was a marginal positive effect of training set coherence [Odds ratio = 3.12, ß = 1.14, *p* = .06], showing some evidence that high coherence training tended to lead to a higher proportion of prototype users than low coherence training. There was also a significant effect for one of the two dummy-coded set size variables where large, repetition-matched sets were coded as 1 (Odds ratio = 6.55, ß = 1.88, *p* = .01), indicating that the large, repetition-matched sets had the highest proportion of prototype-users overall. There was also a significant interaction between training set coherence and the large, repetition-matched dummy coded variable (Odds ratio = .05, ß = -3.00, *p* = .002). High coherence training led to a higher proportion of subjects best fit by the prototype model compared to low coherence training for both the small sets (high = 74% prototype, low = 48% prototype) and the large, trial-matched sets (high = 79% prototype, low = 50% prototype). That was not the case for large, repetition-matched sets where a higher proportion of subjects were best fit by the prototype model following low coherence (86% prototype) compared to high coherence training (48% prototype). Thus, there is some evidence that the tendency to use prototype representations during category generalization may depend on the coherence of training as well as both the number of training examples and how often they are repeated during training.

Discussion

In Experiment 1, we manipulated training set coherence and set size to test which of these most strongly impact the formation of generalizable category knowledge and the representations underlying category generalization. Overall, we found that higher coherence training sets facilitated category learning and broad generalization compared to training on low coherence sets. This difference in performance was accompanied by a higher proportion of participants best fit by the prototype model following high coherence compared to low coherence training in two of the three set size conditions: small set size, and large, trial-matched set size condition. Unexpectedly, the exception was the large, repetition-matched condition, which showed a higher proportion of prototype users following low coherence training. Thus, while set size had minimal effects on accuracy, it did impact the types of representations that individuals used for category generalization.

The strong role of training set coherence is consistent with our prior findings (Bowman & Zeithamova, 2020), and provides new evidence that the magnitude of the coherent training performance benefit does not depend on using a large training set. However, one issue with the training set structure used in this experiment is that the low coherence training sets included features that were non-diagnostic of category membership (see Appendix). This was especially true of the small, low coherence structure in which half of the features in the training set (4 features) were non-diagnostic compared to only one feature in the large, low coherence training set. This confound occurred because it is not mathematically possible with 8-dimensional stimuli to have 4 items that each share 5 features with their category prototype and also have each feature be equally predictive of category membership. For high coherence training sets, all features were equally predictive of category membership. Including these non-diagnostic features only in the low coherence categories may have added an additional demand – determining relevant versus irrelevant features – that was not present for the high coherence training. We thus wanted to ensure that the strong effect of coherence and minimal effect of set size in Experiment 1 were not driven by differences in the predictiveness of features across training conditions.

Experiment 2

Experiment 2 manipulated training set coherence and training set size while keeping all stimulus features within a condition equally predictive of category membership. The number of stimulus features that varied independently was increased from 8 to 10, allowing us to equate feature predictiveness within each coherence group and also test whether our findings from Experiment 1 extend to even higher dimensional stimuli. As in Experiment 1, we focused on whether training set coherence and training set size have independent and/or combined effects on the ability to learn categories, the ability to generalize category labels to new examples, and the representations underlying those generalization judgments. Old/new recognition was also tested, and those data are presented in the Supplementary Materials.

Method

Participants

Prior to data collection, we conducted a power analysis and determined that 270 participants are needed to detect a small to medium (η_p^2 > .03) set coherence x set size interaction effect with 80% power. We recruited 289 participants to ensure we could meet the target sample even after exclusions. Of those, six were excluded for failure to complete all experimental phases, six were excluded for failing to respond on a majority of trials in one or both of the test phases (categorization, recognition), and one was excluded due to a deviation from the study protocol, leaving data from 276 participants reported in all analyses (199 self-reported females, 68 self-reported males, and 9 who self-reported as another gender; mean age = 19.2 years, SD age = 1.3 years, range 18-27 years). Participants were randomly assigned to one of six training groups (see Table 3 for demographic information separated by training condition). All participants were recruited from the University of Oregon and received course credit. All participants completed written informed consent. All procedures were approved by the University of Oregon’s Institutional Review Board.

Table 3

Experiment 2 training sets and demographic information

Training Set	n females / n males / n other[1]	Mean age in years (SD age, age range)
Small, high coherence	32 / 14 / 1	19.2 (1.2, 18-23)
Small, low coherence	30 / 13 / 3	19.0 (0.9, 18-21)
Large, high coherence (repetition-matched)	33 / 9 / 2	19.0 (1.0, 18-22)
Large, low coherence (repetition-matched)	32 / 13 / 0	18.9 (1.0, 18-21)
Large, high coherence (trial-matched)	34 / 10 / 1	19.4 (1.5, 18-26)
Large, low coherence (trial-matched)	38 / 9 / 2	19.4 (1.7, 18-27)

Materials

The stimuli were cartoon animals that varied along 10 binary features (as in Bowman & Zeithamova, 2020; Zeithamova et al., 2008) rather than 8 binary features as in Experiment 1. Specifically, in Experiment 1, the width of the leg (thin/wide) and the shape of the foot (clawed/webbed) co-varied perfectly to create a single, combined foot and leg feature (thin clawed vs. wide webbed). The crown shape (crescent/comb) and the face shape (beak/snout) also covaried to create a combined head shape feature. In Experiment 2, these features were separated and no longer co-varied together as one. The process for generating the category structure was similar to Experiment 1 but took into account the added dimensions: stimuli with 6-10 prototypical A features were considered category A members. Stimuli with 0-4 prototypical A features (thus 6-10 prototypical B features) were considered category B members.

[1] In Experiment 1, participants were given binary male/female options to describe their gender. In Experiment 2, the gender question was open-ended, and we have separated responses into females, males, and some other gender identity.

Training sets. As in Experiment 1, training sets varied in coherence (high and low). All training stimuli in the low coherence sets had 60% typical features, sharing 6 of 10 features with their category prototype. All the training stimuli in the high coherence sets had 80% typical features, sharing 8 of 10 features with their category prototype. The structural ratio (average differing features within vs. between categories) for each training set was as follows: small, high coherence = 0.59; small, low coherence = 1.07; large, high coherence (both repetition- and trial-matched) = 0.54; large, low coherence (both repetition- and trial-matched) = 0.99. Training sets also varied in size. Small sets included 5 items per category (10 unique training items), and large sets included 10 items per category (20 unique training items). In the small training set conditions, each item was presented 16 times during training. As in Experiment 1, large training sets could either be matched to small sets in terms of the number of presentations of each training item (repetition-matched condition) or in terms of the number of total training trials (160 total trials; trial-matched condition). Training set structures for each set size x set coherence condition are in the Appendix.

Categorization stimuli. Selection of test stimuli was similar to Experiment 1, taking into account the use of 10-dimensional rather than 8-dimensional stimuli. In addition to old (training) stimuli that differed based on the initial training condition, the categorization test included 42 new stimuli: two prototypes and 5 new items for each level of prototypicality (i.e., sharing 9, 8, 7, 6, 4, 3, 2, or 1 feature with prototype A), excluding equidistant items.

Procedure

The procedure followed the structure from Experiment 1 with participants completing phases in the following order: training, recognition, and categorization. We adjusted the trial timing in each phase to make the experiment more efficient and reduce the passive time for the participant. During the feedback-based training, each cartoon animal was presented on the screen for 2 seconds before the response options appeared on the screen and the participant could make a self-paced response. Feedback was displayed for 1.5 seconds immediately after the response, followed by a 1.5 second inter-trial fixation. In the categorization test, test items were presented for 4 seconds as in Experiment 1, but the inter-trial fixation was reduced to 1 second. The order of trials within a block was completely randomized.

Statistical analyses

The statistical approach was the same as in Experiment 1 but adjusted for 10-dimensional rather than 8-dimensional stimuli. For example, analyses including test item typicality as a within-subject factor included an additional level of typicality afforded by higher dimensionality stimuli, and the vector of attention weights estimated by prototype and exemplar models included 10 rather than 8 values.

Data availability

Data for the present study are freely available alongside the data from Experiment 1 through the Open Science Framework at https://osf.io/snqd5/?view_only=f56e3da5d1a54633aafe4a8ad0ec051b.

Results

Training accuracy

We first tested whether there were differences in how well participants learned from training sets that varied in their number of examples and the coherence of those examples. We computed two set size x set coherence x block mixed-factors ANOVAs on training accuracy. One ANOVA compared small sets and large sets that were repetition-matched (Figure 6A), and one ANOVA compared small sets and large sets that were matched for the total number of training trials (Figure 6B). We used a Bonferroni corrected α = .025 to account for the two separate ANOVAs. In both cases, there was a significant effect of training block (both F's > 43, p's < .001, η_p^2 > .19), driven by increases in accuracy over the course of the training phase (both linear effects F's > 131, p's < .001, η_p^2 > .41). Both also showed a significant main effect of set coherence (both F's > 637, p's < .001, η_p^2 > .78) and a significant training block x set coherence interaction (both F's > 15, p's < .001, η_p^2 > .07). These effects were driven by better overall learning for those trained on high coherence sets (repetition-matched M = .80, SD = .08; trial-matched M = .79, SD = .08) compared to low coherence sets (repetition-matched M = .52, SD = .08; trial-matched M = .51, SD = .08), and a steeper learning curve for training on high coherence sets (repetition-matched single group linear effect: $F(1,90) = 164.71$, $p < .001$, η_p^2 = .65; trial-matched, single group linear effect: $F(1,91) = 149.37$, $p < .001$, η_p^2 = .62) compared to low coherence sets (repetition-matched single group linear effect: $F(1,90) = 14.38$, $p < .001$, η_p^2 = .14; trial-matched, single group linear effect: $F(1,94) = 13.41$, $p < .001$, η_p^2 = .13). Thus, there was a distinct disadvantage to learning categories from a low coherence set of examples.

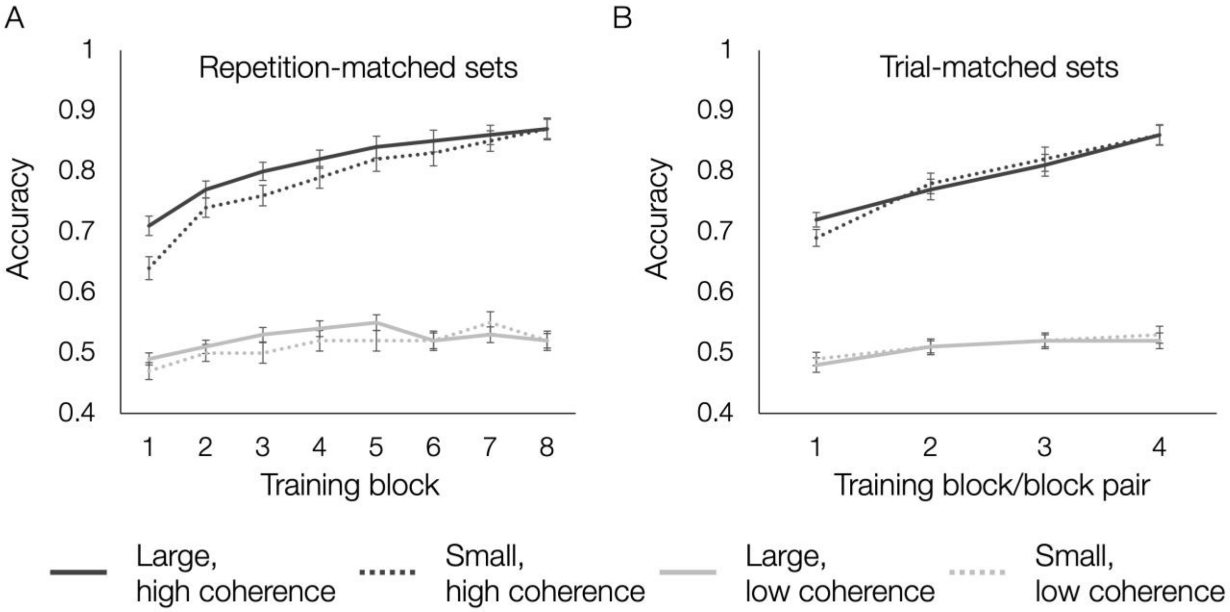


Figure 6. **A**. Training accuracy by block for the small set size conditions (dotted lines) and large set size conditions (solid lines) that matched the small set size in terms of the number of repetitions of each item. High-coherence conditions in black, low-coherence conditions in gray. **B.** Same as A, but depicting the large set size conditions (solid lines) that matched the small set size in terms of the total number of trials during training. For small training sets, the same data are depicted in A and B, but collapsed from 8 blocks of 20 trials each in A into 4 block-pairs of 40 trials in B to match 40 trials per block in the large set size (trial matched) conditions. Error bars depict the standard error of the mean across subjects.

In Experiment 1, there were hints of a benefit of learning from small compared to large sets in the low coherence condition, but only when sets were equated in terms of the total number of trials. Here, neither the main effect of set size (*F*(1,183) = 0.04, *p* = .84, η_p^2 < .001) nor the set size x coherence interaction (*F*(1,183) = 0.23, *p* = .63, η_p^2 = .001) was significant in the trial-matched comparison. There was a marginal main effect of set size when we controlled for the number of item repetitions (*F*(1,178) = 2.96, *p* = .09, η_p^2 = .02) that was driven by numerically better performance in the large set size that involved twice as many training trials. The set size x set coherence interaction was not significant (*F*(1,178) = 0.45, *p* = .50, η_p^2 = .003). Across both trial-matched and repetition-matched comparisons, no other interaction effect approach significance either (all F's < 1.5, p's > .19, η_p^2 < .009). Taken together, there was a clear advantage of learning from high coherence training sets without a strong effect of training set size regardless of whether small and large sets were equated for the number of item repetitions or the total number of training trials. Notably, participants in the large, repetition matched set size condition received double the training compared to the small set size condition or large, trial matched condition. Yet, this was not enough to rescue their performance when the coherence of training exemplars was low, with the final block training accuracy not exceeding 55% accuracy in any of the 3 low coherence groups.

Categorization test accuracy

See Figure 7 for categorization accuracy separated by training group and test item type.

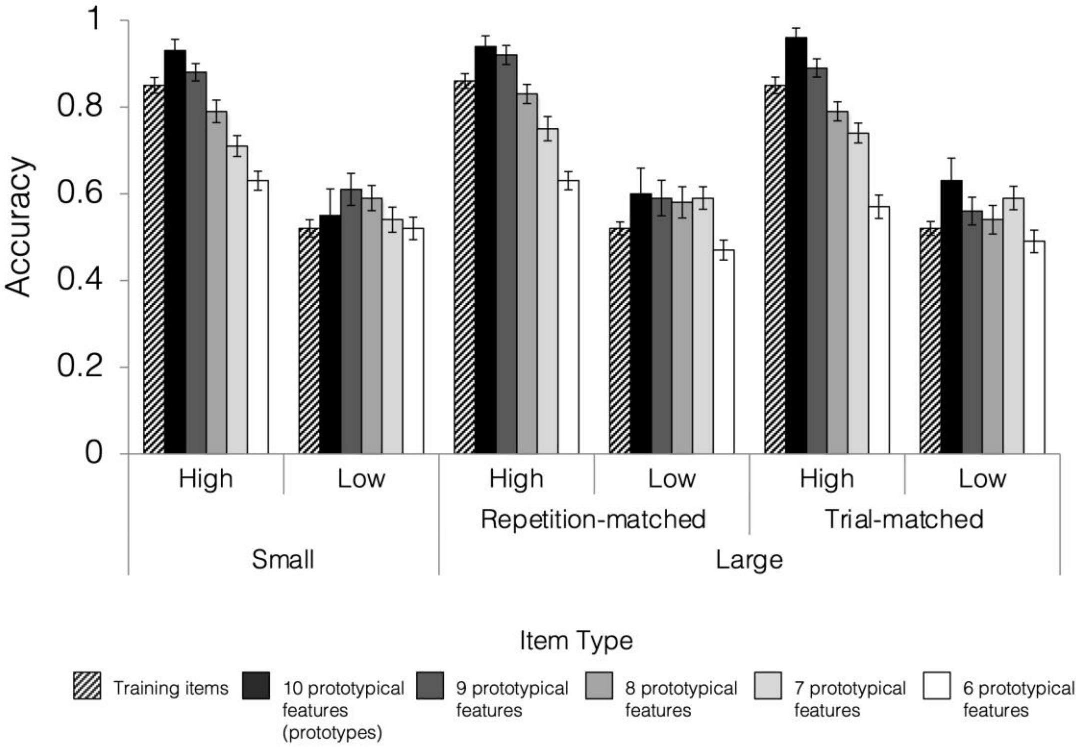


Figure 7. Categorization accuracy for each item type separated by training set. Small = sets with 5 items/category. Large = sets with 10 items/category. High = high coherence: 8 prototypical features for all training examples. Low = low coherence: 6 prototypical features for all training examples. Accuracies on training (old) items re-presented during the categorization test are depicted with striped bars. Accuracies for new items varying in their similarity to category prototypes are depicted with solid bars. Accuracies for prototypes are depicted in the darkest bars with increasingly lighter bars for new items sharing fewer features with the prototypes. Error bars depict standard error of the mean across subjects.

To understand how training set coherence affected the ability to classify items that varied in their distance from category prototypes and to test for any effect of training set size on generalization, we computed a 3 (set size: small, large and repetition-matched, large and trial-matched) x 2 (set coherence: high, low) x 5 (test item typicality: 6-10 prototypical features) mixed-factors ANOVA. There was a significant main effect of item typicality [$F(3.1,824.4) = 65.86$, $p < .001$, $\eta_p^2 = 0.20$, GG] accompanied by a significant linear effect [$F(1,270) = 142.17$, $p < .001$, $\eta_p^2 = .35$] driven by better categorization for items closest to the prototypes and decreasing accuracy towards category boundary. The main effect of training set size was not significant [*F*(2,270) = 0.27, *p* = .76, η_p^2 = .002] nor were there any significant interactions between set size and the other factors (all F's < 1.7, p's > .13, η_p^2 < .02).

There was, however, a significant main effect of training set coherence [*F*(1,270) = 173.23, *p* < .001, η_p^2 = .39] with better accuracy for those trained on high coherence sets (M = .80, SD = .12) compared to low coherence sets (M = .56, SD = .16). Training set coherence also interacted significantly with test item typicality [*F*(3.1,824.4) = 21.42, *p* < .001, η_p^2 = .07, GG]. This interaction is depicted in Figure 8. To better understand the nature of this interaction, we compared high and low coherence training groups (collapsed across training set size groups) in terms of categorization accuracy at each level of test item typicality using two-sample t-tests. We used a Bonferroni-corrected α = .01 to account for the five separate tests. There was a significant advantage for high compared to low coherence training at each level of test item typicality (all t's > 5.9, p's < .001), but the effect size decreased for items with fewer prototypical features (prototype *d* = 1.15, 9 prototypical features *d* = 1.52, 8 prototypical features *d* = 1.21, 7 prototypical features *d* = 0.92, 6 prototypical features *d* = 0.71). Notably, however, those trained on low coherence sets never outperformed those trained on high coherence sets. Further, those trained on low coherence sets performed significantly worse on the category boundary items (6 prototypical features) compared to all other generalization items (all t's > 2.9, p's < .005, d's > 0.24). Thus, our findings are *inconsistent* with the hypothesis that high variability among training examples promotes breadth of category knowledge.

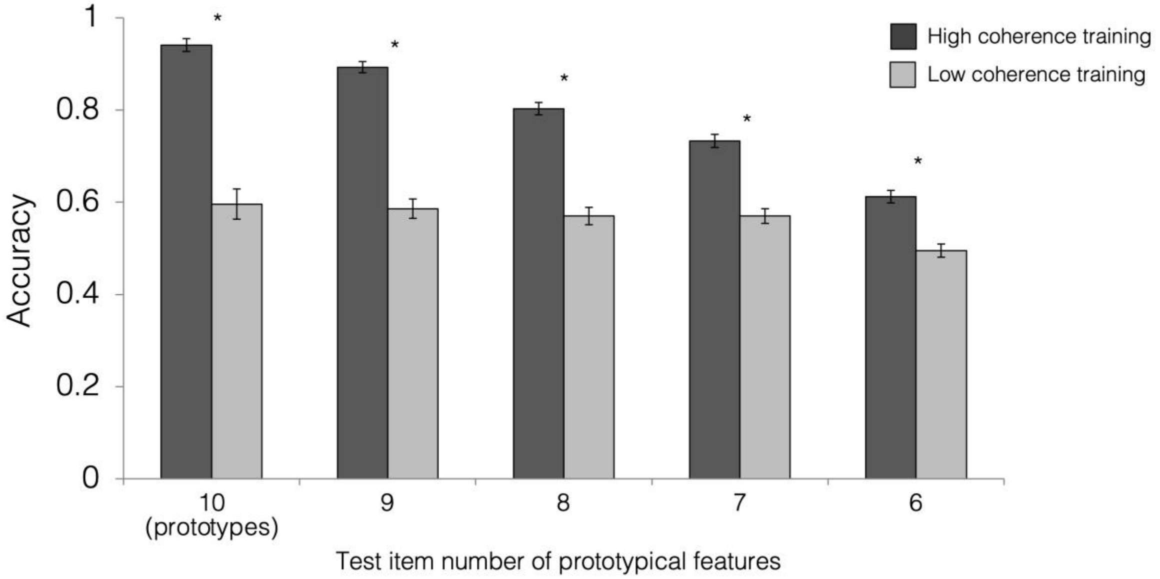


Figure 8. Training set coherence x test item typicality interaction effect in generalization. Test item typicality is depicted on the x-axis with the prototypes on the far left and the items closest to the category boundary (6 prototypical features) on the far right. Dark bars depict generalization accuracy for those trained with high coherence sets (8 prototypical features for all training examples), and light bars depict generalization accuracy for those trained on low coherence sets (6 prototypical features for all training examples). Both coherence groups were collapsed across set size conditions. Stars indicate differences between groups of p < .05 corrected. Error bars depict the standard error of the mean across subjects.

Prototype and exemplar model fits in categorization

The proportion of participants best fit by each model in each condition is presented in Figure 9. Table 4 presents the average best fit values for each estimated parameter separated by model and by training group. As in Experiment 1, we did not constrain the models to return the same or similar parameter estimates. Nonetheless, there was a strong average within-subject correlation of r = .91 for the attention weights estimated by the two models. Also similar to Experiment 1, the c-parameter from the prototype model was more correlated with the gamma parameter from the exemplar model (r = .25) than the c parameter from the exemplar model (r = .07).

Table 4 – Mean best fitting parameter values for the exemplar and prototype models in categorization

	Exemplar model												Prototype model										
Training set	C	G	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	C	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
Small, high coherence	16.1	35.6	.16	.11	.11	.12	.08	.11	.06	.03	.07	.15	34.2	.15	.09	.11	.11	.09	.11	.07	.04	.08	.16
Small, low coherence	13.7	31.9	.14	.09	.10	.16	.08	.08	.08	.08	.09	.11	7.1	.13	.10	.09	.15	.08	.11	.09	.07	.07	.10
Large, high coherence, repetition-matched	26.1	47.3	.14	.10	.10	.13	.05	.16	.09	.03	.10	.10	32.0	.13	.11	.09	.13	.07	.14	.12	.02	.08	.11
Large, low coherence, repetition-matched	15.0	33.9	.14	.08	.07	.16	.08	.12	.09	.07	.08	.11	5.3	.13	.06	.08	.14	.10	.12	.10	.07	.08	.13
Large, high coherence, trial-matched	10.9	42.8	.12	.05	.08	.11	.04	.13	.07	.06	.12	.22	33.3	.11	.04	.09	.11	.05	.14	.08	.06	.12	.20
Large, low coherence, trial-matched	11.6	46.8	.11	.08	.08	.12	.09	.07	.09	.10	.08	.16	4.6	.11	.08	.10	.10	.08	.09	.09	.10	.09	.17

C = sensitivity, G = gamma (exemplar model only), W1-10 = attention weights to each of 10 stimulus features

Our first step in determining the best fitting model for each subject was to compare each subjects' model fits to their respective null distributions to determine whether the fit of each model was better than would be expected by chance. Across the entire sample, regardless of training condition, 193 subjects (70%) showed an exemplar model fit that reliably outperformed chance, and 197 subjects (71%) showed a prototype model fit that reliably outperformed chance. Seventy-eight subjects (28%) showed prototype and exemplar model fits that were not reliably above chance and were thus labeled with 'chance' for their strategy. Comparing prototype and exemplar model fits to each other in the remaining 198 subjects where at least one model outperformed chance, 50 subjects (25%) were best fit by the exemplar model, 129 subjects (65%) were best fit by the prototype model, and 19 subjects (10%) had comparable exemplar and prototype fits and were labeled with 'similar fits' for their strategy.

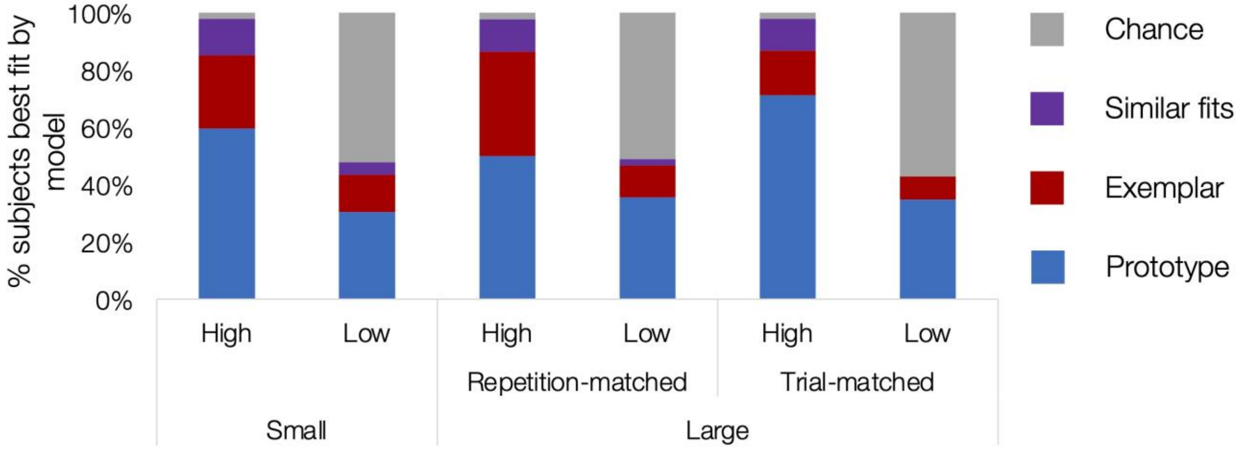


Figure 9. Prototype and exemplar model fits to categorization data. Small = sets with 5 items/category. Large = sets with 10 items/category. High = high coherence: 8 prototypical features for all training examples. Low = low coherence: 6 prototypical features for all training examples. Subjects better fit by the prototype than exemplar model are indicated in blue. Those better fit by the exemplar than the prototype model are indicated in red. Those will comparable fits across models are in purple ('similar fits'). Those whose fits did not outperform the random model are in grey. Model fits are separated by training set typicality, set size, and whether large sets were matched to small sets in terms of the number of repetitions of each item (repetition-matched) or the number of total training trials (trial-matched).

To determine whether training set size or coherence affected the number of prototype users during categorization, we computed a binary logistic regression with set size (small, large repetition-matched, large trial-matched), set coherence (high, low) and their interaction as categorical predictors of whether individuals relied on a prototype strategy, excluding 'chance' subjects. This model explained only 6% of variance (Nagelkerke R^2) and none of the individual predictors reached significance (all odds ratios < 2.5, |ß's| < .9, p's > .2). Thus, unlike in Experiment 1, neither training set size nor set coherence reliably affected the tendency for subjects to rely on one categorization strategy over another. When we instead used the proportion of 'chance' subjects as the dependent variable, we saw a strong effect of training set coherence (Odds ratio = .02, ß = -3.92, p < .001; Model Nagelkerke R^2 = .46), with more 'chance' subjects following low coherence training (54% of subjects) compared to high coherence training (2% of subjects). No other predictor reached significance (all odds ratios < 1.3, |ß's| < .3, p's > .6). Taken together, these results indicate that low coherence training tended to make it difficult for participants to use either an exemplar or prototype strategy successfully rather than biasing participants to one strategy or the other.

Discussion

In Experiment 2, we sought to determine whether the benefits of high coherence training identified in Experiment 1 were still present when all features were diagnostic. Category learning and generalization results were broadly consistent with both Experiment 1 and our prior findings (Bowman & Zeithamova, 2020): high coherence training was associated with better learning and generalization compared to low coherence training. However, unlike in Experiment 1 and prior work (Bowman & Zeithamova, 2020), high coherence training did not lead to greater reliance on prototype representations during generalization. Instead, poorer categorization performance for those trained on low compared to high coherence sets was accompanied by an increase in the proportion of subjects who were not fit well by either model, with approximately half of the participants labeled as “chance”. Effects of training set size were minimal in learning and generalization performance and in the representations underlying categorization judgments.

General Discussion

Across two experiments, we manipulated the number of category training examples and their coherence, and we measured rates of learning, category generalization, and the representations (prototype vs. exemplar) underlying categorization judgments. Consistent with our prior study (Bowman & Zeithamova, 2020), we found better acquisition of category knowledge *and* better category generalization when participants were trained on high coherence compared to low coherence sets. Some early studies of categorization found that learning from low coherence sets led to better generalization (Homa & Vosburgh, 1976; Peterson et al., 1973; Posner & Keele, 1968) and suggested that learning from low coherence sets helps with learning category breadth and particularly benefits generalization to the category boundary (Dukes & Bevan, 1967; Perry et al., 2010; Peterson et al., 1973). The idea that high variability in training examples leads to better generalization is still a prominent one (for review see Raviv et al., 2022). However, the benefits of low coherence training were likely driven by the increased exposure during training necessary to meet a learning criterion set for those studies (Hintzman, 1984). Other studies showing the benefits of variability among training examples have induced variability by including more unique instances in the high variability condition and repeated individual items in the low variability condition, thus conflating the number of examples with the variability among difference examples (Doyle & Hourihan, 2016; Nosofsky et al., 2019; Wahlheim et al., 2012, 2016). Our study was novel in controlling both the amount of training and the number of examples when testing for the effects of training set coherence. Our findings show that there was never a generalization advantage for the low coherence training, even for new items at the category boundary. Instead, generalization to new items was always better—numerically or significantly—for all levels of typicality of the test items. This finding is particularly striking given that those in the low coherence groups were trained entirely on items nearest the category boundary, yet they are no better than the high coherence training group at generalizing to new items near category boundary.

While our findings show a benefit of high coherence training that differs from studies that trained to criterion, they align well with a theoretical prediction based on computational modeling by Hintzman (1984) who argued that coherent training should be beneficial for generalization when the length of training is equated. They also align well with some other manipulations of category coherence. Prior work in the domain of social categorization has also shown that coherence (called ‘entitativity’ in this domain (Campbell, 1958)) is a key factor that drives naïve perceptions of what defines category membership (Haslam et al., 2000). Thus, part of the high coherence training benefit may be that it fits well with naïve intuitions about how categories are formed, allowing participants to quickly adopt strategies that are well suited to learning categories based on similarity to the category center. Another operationalization of category coherence is the structural ratio: distances between items within the same category compared to the distances between items from different categories (Homa et al., 1979; Minda & Smith, 2001). Prior work has shown faster learning from more coherent, ‘well structured’ category sets in which there is more clustering of items within vs. between categories (Minda & Smith, 2001). That high coherence training produced highly generalizable category knowledge in our study is also consistent with prior work suggesting that stability and consistency in input facilitates broad category knowledge (Carvalho et al., 2019; Horst et al., 2011), and that learning from an easier training set can facilitate later categorization of more difficult items (Edmunds et al., 2019). Finally, the findings also align with our prior aging study showing that when training contains a mixture of typical and atypical items, older adults have difficulty learning the atypical items but their successful acquisition of typical items is sufficient to support subsequent generalization at levels comparable to young adults (Bowman et al., 2022).

The advantage of learning from high coherence training sets appears robust across the experiments presented here and our prior work with similar category structures (Bowman & Zeithamova, 2020). However, it may not be equally suitable for all types of stimuli or category structures. While we showed this effect at two different levels of high stimulus dimensionality (8 and 10 dimensional stimuli), the stimuli were always binary dimension cartoon animals, and the category structure was always prototype-based. Categories *not* centered around a single prototype (e.g., multiple prototypes, disjunctive or rule-plus-exception category structures) may require a very different sampling of training exemplars to be learned robustly. For example, using natural categories like rocks and birds, others have shown that training sets that span the full category space or offer a wider variety of examples may be particularly good for promoting generalization (Nosofsky et al., 2019; Wahlheim et al., 2012). Similarly, information-integration categories with non-linear boundaries require extensive training with large number of varied exemplars to provide an opportunity to learn not only the central tendency of a category but the entire distribution range (Ashby & Gott, 1988; McKinley & Nosofsky, 1995). Thus, one possibility is that high coherence training is uniquely well suited to prototype-based categories because it leads participants toward the underlying category structure. For other types of categories, especially those that are not linearly separable, greater variability and/or extended training may be quite important. Interestingly, even rule-plus-exception category structures may be easier to learn when exposure to exceptions is delayed (Heffernan et al., 2021), suggesting that training that highlights commonalities among category members not only makes it easier to learn rules or prototype-based categories but also memorize the exceptions.

In contrast to training set coherence, we found little evidence that training set size affected the ability to learn and generalize novel categories. This lack of a set size effect is consistent with our prior study (Bowman & Zeithamova, 2020), but differs from some previous work showing better generalization for larger categories, even when they used small and large set sizes comparable with the current study (Goldman & Homa, 1977; Homa et al., 1973, 1981). However, these previous studies involved each subject learning two or more categories that differed in the number of examples during learning, whereas our studies involved learning two categories of equal size and the set size manipulation occurred across subjects. Thus, the contrast across different categories within a single learning session may partially drive benefits larger set sizes identified in prior studies.

Notably, the current study was novel in that it included two versions of the large set size condition, one that matched the small set size in terms of the total number of the training trials (providing half of the repetitions per item) and one that matched the small set size in terms of the number of repetitions per item (thus providing double the number of the total training trials). Studies sometimes control for one or the other when testing the effect of set size, but not both. Here, we found that set size had little effect on generalization success regardless of whether small and large training sets were matched in terms of exposures to individual examples or total number of learning trials. Thus, we found no evidence that large training sets would provide a generalization advantage, not even in the condition where using a large training set meant doubling the length of training.

In addition to assessing overall effects of set size and coherence, we were interested in the extent to which these factors interacted. More specifically, we were interested in whether set size would modulate the effect of coherence when the amount of training is controlled, as was observed previously in a study that included training to a criterion (Homa & Vosburgh, 1976) and as predicted theoretically (Minda & Smith, 2001). In contrast to the Homa and Vosburgh (1976) study, we found a strong generalization advantage for high coherence sets that was consistent across set sizes. This finding is a key addition to our prior paper, which was not a fully crossed design and did not allow us to assess the potential for an interaction between set coherence and set size (Bowman & Zeithamova, 2020). Overall, results from categorization performance suggest that training set coherence is a much stronger influence on category learning and generalization than training set size. This finding also has practical importance for those compiling training examples to teach new categories: there may be flexibility in the number of training examples needed for learners to generalize well if those training examples are coherent around the category center.

We were also interested in whether high coherence training tended to promote categorization judgements based on prototype representations as suggested by some theoretical accounts (Minda & Smith, 2001) and as we found in our prior work (Bowman & Zeithamova, 2020). While the effect of training coherence was robust in terms of category learning and generalization, the model estimates of underlying representations produced a more nuanced pattern. Experiment 1 showed increased prototype fit for high compared to low coherence in two out of the three training set size groups: the small set size and the large, trial matched group. This finding is similar to our prior study (Bowman & Zeithamova, 2020) and suggests that high coherence training may promote generalization by helping participants to detect commonalities across training examples in support of prototype formation. This finding also supports the idea that making prototype information more salient during training increases the likelihood that participants will rely on prototype representations during generalization

(Medin et al., 1984). But the pattern was reversed in the large, repetition-matched group: low coherence training led to a larger proportion of subjects who relied on prototype representations than the high coherence group. The direction of this interaction was unexpected. If anything, we expected large, high coherence training to produce a particularly strong prototype model advantage. This expectation was based on prior theoretical proposals (Minda & Smith, 2001) and work showing that larger training sets help participants uncover prototype-based category structures compared to small sets (Homa et al., 1979). One possible explanation for this unexpected finding is that the large, repetition-matched training condition allowed participants more flexibility to explore a variety of strategies across an extended training regime. This would then provide those in the low coherence group more time to "discover" the prototype strategy during training.

However, we hesitate to strongly interpret the interaction between training set coherence and set size from Experiment 1 because it did not replicate in Experiment 2. In Experiment 2, only about half of subjects in low coherence training groups had model fits reliably better than chance, while most subjects in the high coherence training had model fits clearly above chance. Thus, it was clear both from the accuracy results and the model fitting that low coherence training in Experiment 2 made it difficult to form high quality category representations of any kind. Once we excluded subjects who were given the 'chance' label in model fitting, we did not see differences across conditions in prototype vs. exemplar strategy use based on either training coherence or set size. Thus, subjects who were able to form categories from the low coherence training did not show a different strategy bias than those in the high coherence training. Taken together, the present experiments provide mixed support for a prototype bias following high coherence compared to low coherence training.

Notably, however, we did not find evidence of an exemplar representation bias for any training condition in either of the experiments presented here nor in our prior study (Bowman & Zeithamova, 2020). One might expect that low coherence training would facilitate memory for individual category members given that these members should be relatively distinct from one another and thus potentially easier to encode separately from one another (Konkle et al., 2010; Vokey & Read, 1992; Winograd, 1981). One might similarly expect that small set sizes might be well suited to exemplar representations given the limited number of individual items to encode into memory, and that the combination of a small, low coherence set might be particularly conducive to exemplar representations. It is possible that we could not see the exemplar model advantage for small sets because the current stimuli make the formation of exemplar representations challenging. Notably, most participants in low coherence training conditions in Experiment 2 failed to learn the category labels for just 10 (small set) or 20 (large sets) training items, even after 16 repetitions, which they could have theoretically done through rote memorization. A replication study testing the separate and joint effects of training coherence and set size in a task in which subjects are consistently able to form representations of individual items across conditions would be helpful to determine how training conditions affect the representations supporting subsequent generalization judgments.

Conclusion

How structure of our experience affects learning, memory and generalization is of interest to many research domains and disciplines. Here, we revisited the question how variability among category examples during training affects learning and generalization, and how training variability interacts with other aspects of training, especially the number of unique examples and their repetition. Contrary to a common assumption that high-variability training promotes generalization, we found robust benefits of training on more coherent, less variable exemplars. While coherence was a strong driver of better category learning and generalization, we found relatively limited differences in categorization performance between small and large training sets, regardless of whether they were matched for the total number of training trials or the number of exposures to individual training items. Furthermore, set size did not seem to moderate the strong effect of set coherence on generalization accuracy. Training coherence and set size jointly affected the types of category representations people formed, but the effects on representations were not consistent across experiments and may be more nuanced than predicted based on current theoretical considerations. Together, these results add to theoretical and empirical work indicating that training that highlights commonalities among exemplars promotes formation and generalization of conceptual knowledge.

References

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53. https://doi.org/10.1037/0278-7393.14.1.33

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 50. https://doi.org/10.1037/0096-1523.18.1.50

Blair, M., & Homa, D. (2003). As easy to memorize as they are to classify: the 5-4 categories and the category advantage. *Memory & Cognition*, *31*(8), 1293–1301. https://doi.org/10.3758/BF03195812

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *ELife*, *e59360*, e59360. https://doi.org/10.7554/elife.59360

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2022). The effects of age on prototype- and exemplar-based categorization. *Psychology and Aging*, *37*(7), 800–815. https://doi.org/10.31234/OSF.IO/A3VUJ

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *The Journal of Neuroscience*, *38*(10), 2605–2614. https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Bowman, C. R., & Zeithamova, D. (2020). Training set coherence and set size effects on concept generalization and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(8), 142–1464.

Bulgarelli, F., & Weiss, D. J. (2019). The More the Merrier? The Impact of Talker Variability on Artificial Grammar Learning in Preschoolers and Adults. In M. M. Brown & B. Dailey (Eds.), *Proceedings of the 43rd Boston University Conference on Language Development* (pp. 123–136). Cascadilla Press.

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science*, *3*(1), 14–25.

Carvalho, P. F., Chen, C., & Chen, Y. (2019). Rethinking the input: Skewed distributions of exemplars result in broad generalization in category learning. *PsyArXiv*.

Doyle, M. E., & Hourihan, K. L. (2016). Metacognitive monitoring during category learning: how success affects future behaviour. *Memory*, *24*(9), 1197–1207. https://doi.org/10.1080/09658211.2015.1086805

Dukes, W. F., & Bevan, W. (1967). Stimulus variation and repetition in the acquisition of naming responses. *Journal of Experimental Psychology*. https://doi.org/10.1037/h0024575

Edmunds, C., Wills, A. J., & Milton, F. (2019). Initial training with difficult items does not facilitate category learning. *Quarterly Journal of Experimental Psychology (2006)*. https://doi.org/10.1080/17470218.2017.1370477

Goldman, D., & Homa, D. (1977). Integrative and metric properties of abstracted information as a function of category discriminability, instance variability, and experience. *Journal of Experimental Psychology: Human Learning and Memory*. https://doi.org/10.1037/0278-7393.3.4.375

Hahn, U., Bailey, T. M., & Elvin, L. B. C. (2005). Effects of category diversity on learning, memory, and generalization. *Memory & Cognition 2005 33:2*, *33*(2), 289–302. https://doi.org/10.3758/BF03195318

Hart, P. E. (1968). The Condensed Nearest Neighbor Rule. *IEEE Transactions on Information Theory*. https://doi.org/10.1109/TIT.1968.1054155

Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology*, *39*(1), 113–127. https://doi.org/10.1348/014466600164363

Heffernan, E. M., Schlichting, M. L., & Mack, M. L. (2021). Learning exceptions to the rule in human and model via hippocampal encoding. *Scientific Reports*, *11*, 21429. https://doi.org/10.1038/s41598-021-00864-9

Hernandez-Garcia, A., & König, P. (2020). Data augmentation instead of explicit regularization. *ArXiv Preprint*.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101. https://doi.org/10.3758/BF03202365

Homa, D., Cross, J., Cornell, D., Goldman, D., & Shwartz, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology*, *101*(1), 116–122. https://doi.org/10.1037/h0035772

Homa, D., & Cultice, J. C. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 83–94. https://doi.org/10.1037/0278-7393.10.1.83

Homa, D., & Little, J. (1985). The abstraction and long-term retention of ill-defined categories by children. *Bulletin of the Psychonomic Society*, *23*(4), 325–328. https://doi.org/10.3758/BF03330172

Homa, D., Rhoads, D., & Chambliss, D. (1979). Evolution of conceptual structure. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(1), 11–23. https://doi.org/10.1037/0278-7393.5.1.11

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *J Exp Psychol Hum Learn Mem*, *7*(6), 418–439. https://doi.org/10.1037//0278-7393.7.6.418

Homa, D., & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(3), 322–330. https://doi.org/10.1037/0278-7393.2.3.322

Horst, J. S., Parsons, K. L., & Bryan, N. M. (2011). Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2011.00017

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/a0019165

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. https://doi.org/10.1037/0033-295X.99.1.22

Lamberts, K. (1994). Flexible tuning of similarity in exemplar-based categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*(5), 1003–1021. https://doi.org/10.1037//0278-7393.20.5.1003

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309

Maddox, W. T., Glass, B. D., Zeithamova, D., Savarie, Z. R., Bowen, C., Matthews, M. D., & Schnyer, D. M. (2011). The effects of sleep deprivation on dissociable prototype learning systems. *Sleep*, *34*(3), 253–260. https://doi.org/10.1093/sleep/34.3.253

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of Exemplar and Decision Bound Models in Large, Ill-Defined Category Structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 128–148. https://doi.org/10.1037/0096-1523.21.1.128

Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/0278-7393.10.3.333

Medin, D. L., Schaffer, M. M., & College, B. (1978). Context Theory of Classification Learning.

Psychological Review, *85*(3), 207–238.

Mervis, C. B., & Pani, J. R. (1980). Acquisition of basic object categories. *Cognitive Psychology*, *12*(4), 496–522. https://doi.org/10.1016/0010-0285(80)90018-3

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal Of Experimental Psychology-Learning Memory And Cognition*, *27*(3), 775–799. https://doi.org/10.1037/0278-7393.27.3.775

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 275–292. https://doi.org/10.1037//0278-7393.28.2.275

Minda, J. P., & Smith, J. D. (2011). Prototype models of categorization: Basic formulation, predictions, and limitations. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 40–64). Cambridge University Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *13*(1), 87–108. https://doi.org/10.1037/0278-7393.13.1.87

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708. https://doi.org/10.1037/0278-7393.14.4.700

Nosofsky, R. M., Denton, S. E., Zaki, S. R., Murphy-Knudsen, A. F., & Unverzagt, F. W. (2012). Studies of implicit prototype extraction in patients with mild cognitive impairment and early Alzheimer’s disease. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 860–880. https://doi.org/10.1037/a0028064

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79. https://doi.org/10.1037/0033-295x.101.1.53

Nosofsky, R. M., Sanders, C. A., Zhu, X., & Mcdaniel, M. A. (2019). Model-guided search for optimal natural-science-category training exemplars: A work in progress. *Psychonomic Bulletin & Review*, *26*, 48–76. https://doi.org/10.3758/s13423-018-1508-8

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 924–940. https://doi.org/10.1037//0278-7393.28.5.924

Ogren, M., & Sandhofer, C. M. (2021). Toddler word learning is robust to changes in emotional context. *Infant and Child Development*, *30*(6), e2270. https://doi.org/10.1002/ICD.2270

Onnis, L., Monaghan, P., & Christiansen, M. H. (2004). Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*.

Perry, L. K., Samuelson, L. K., Malloy, L. M., & Schiffer, R. N. (2010). Learn locally, think globally: Exemplar variability supports higher-order generalization and word learning. *Psychological Science*. https://doi.org/10.1177/0956797610389189

Peterson, M. J., Meagher, R. B. J., Herschel, C., & Gillie, S. (1973). The abstraction and generalization of dot patterns. *Cognitive Psychology*, *4*(3), 378–398. https://doi.org/https://doi.org/10.1016/0010-0285(73)90019-4

Plante, E., Ogilvie, T., Vance, R., Aguilar, J. M., Dailey, N. S., Meyers, C., Lieser, A. M., & Burton, R. (2014). Variability in the Language Input to Children Enhances Learning in a Treatment Context. *American Journal of Speech-Language Pathology*, *23*(4), 530–545. https://doi.org/10.1044/2014_AJSLP-13-0038

Posner, M. I., & Keele, S. W. (1968). On the Genesis of Abstract Ideas. *Journal of Experimental Psychology*, *77*(3, Pt.1), 353–363. https://doi.org/10.1037/h0025953

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*, *26*(6), 462–483. https://doi.org/10.1016/j.tics.2022.03.007

Roiger, R., & Cornell, L. (1996). Selecting training instances for supervised classification. *Proceedings of the Joint Conference on Intelligent Systems/ISAI/IFIS*.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345. https://doi.org/10.1007/BF02288967

Smith, J. D., Murray, M. J., & Minda, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/0278-7393.23.3.659

Twomey, K. E., Ranson, S. L., & Horst, J. S. (2013). *That's More Like It: Multiple Exemplars Facilitate Word Learning*. https://doi.org/10.1002/icd.1824

Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition*. https://doi.org/10.3758/BF03199666

Wahlheim, C. N., Finn, B., & Jacoby, L. L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: evidence for variability neglect. *Memory & Cognition*, *40*, 703–716. https://doi.org/10.3758/s13421-011-0180-2

Wahlheim, C. N., McDaniel, M. A., & Little, J. L. (2016). Category learning strategies in younger and older adults: Rule abstraction and memorization. *Psychology and Aging*, *31*(4), 346–357. https://doi.org/10.1037/pag0000083

Williams, J. J., & Lombrozo, T. (2010). The Role of Explanation in Discovery and Generalization: Evidence From Category Learning. *Cognitive Science*, *34*(5), 776–806. https://doi.org/10.1111/J.1551-6709.2010.01113.X

Winograd, E. (1981). Elaboration and distinctiveness in memory for faces. *Journal of Experimental Psychology: Human Learning and Memory*. https://doi.org/10.1037/0278-7393.7.3.181

Zaki, S. R., & Nosofsky, R. M. (2001). A single-system interpretation of dissociations between recognition and categorization in a task involving object-like stimuli. *Cognitive, Affective, & Behavioral Neuroscience 2001 1:4*, *1*(4), 344–359. https://doi.org/10.3758/CABN.1.4.344

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and Exemplar Accounts of Category Learning and Attentional Allocation: A Reassessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1160–1173. https://doi.org/10.1037/0278-7393.29.6.1160

Zeithamova, D., Maddox, W. T., & Schnyer, D. M. (2008). Dissociable prototype learning systems: evidence from brain imaging and behavior. *Journal of Neuroscience*, *28*(49), 13194–13201. https://doi.org/10.1523/JNEUROSCI.2915-08.2008

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*, 350–361. https://doi.org/10.1016/J.NEUCOM.2017.01.026

Appendix

Category structures for each training condition

Experiment 1

For each participant, the prototype of category A was chosen by randomly selecting from one of four possible prototypes (Table 1). The prototype of category B was defined as the stimulus sharing no features with the prototype of category A. Tables 2-5 present structures of stimuli for each training group coded with respect to the first possible category A prototype (i.e., 1111111111, with 0000000000 serving as the category B prototype). For each training group, the authors pre-defined training stimuli, which were recoded for each participant based on their selected category prototypes.

Table 1. Possible category A prototypes

Dimension							
1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1
1	1	1	1	1	0	0	0
1	1	0	0	1	1	0	1
0	1	1	1	0	0	1	0

Table 2. Stimuli for small, high coherence training sets

Category	Dimension							
	1	2	3	4	5	6	7	8
A	1	1	0	1	1	1	0	1
A	1	0	1	0	1	1	1	1
A	0	1	1	1	0	1	1	1
A	1	1	1	1	1	0	1	0
B	0	0	1	0	0	0	1	0
B	0	1	0	1	0	0	0	0
B	1	0	0	0	1	0	0	0
B	0	0	0	0	0	1	0	1

Table 3. Stimuli for small, low coherence training sets

Category	Dimension							
	1	2	3	4	5	6	7	8
A	0	0	0	1	1	1	1	1
A	1	1	1	0	0	0	1	1
A	1	1	1	1	1	0	0	0
A	1	0	1	0	1	1	0	1
B	0	1	0	0	1	0	0	1
B	1	0	0	1	0	1	0	0
B	0	1	0	0	0	1	1	0
B	0	0	1	1	0	0	1	0

Table 4. Stimuli for large, high coherence training sets

Category	Dimension							
	1	2	3	4	5	6	7	8
A	1	1	0	1	1	1	0	1
A	1	0	1	0	1	1	1	1

A	0	1	1	1	0	1	1	1
A	1	1	1	1	1	0	1	0
A	0	0	1	1	1	1	1	1
A	1	1	1	0	0	1	1	1
A	1	1	0	1	1	0	1	1
A	1	1	1	1	1	1	0	0
B	0	0	1	0	0	0	1	0
B	0	1	0	1	0	0	0	0
B	1	0	0	0	1	0	0	0
B	0	0	0	0	0	1	0	1
B	1	0	0	0	0	0	0	1
B	0	1	0	0	0	0	1	0
B	0	0	1	0	0	1	0	0
B	0	0	0	1	1	0	0	0

Table 5. Stimuli for large, low coherence training sets

Category				Dimension				
	1	2	3	4	5	6	7	8
A	0	0	0	1	1	1	1	1
A	1	1	1	0	0	0	1	1
A	1	1	1	1	1	0	0	0
A	1	0	1	0	1	1	0	1
A	0	1	1	0	1	1	1	0
A	1	1	0	1	0	1	0	1
A	1	0	1	1	1	0	1	0
A	1	1	1	0	1	0	0	1
B	0	1	0	0	1	0	0	1
B	1	0	0	1	0	1	0	0
B	0	1	0	0	0	1	1	0
B	0	0	1	1	0	0	1	0
B	0	0	0	1	0	1	0	1
B	0	1	1	0	0	0	1	0
B	1	0	0	1	0	0	0	0
B	0	0	0	0	0	1	1	1

Experiment 2

As in Experiment 1, the category A prototype was randomly selected for each participant from a list of possible prototypes (Table 6). For each training group, the authors pre-defined training stimuli, which were recoded for each participant based on their selected category prototypes. The base training structure for each training condition is presented in Tables 7-10.

Table 6. Possible category A prototypes

				Dimension					
1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	1	1	1	1	1
1	1	0	0	1	1	0	0	1	1
0	1	1	1	0	0	0	1	0	0

Table 7 - Small, high coherence category structure

Category	Dimension									
	1	2	3	4	5	6	7	8	9	10
A	1	0	1	1	1	0	1	1	1	1
A	0	1	1	1	1	1	1	1	0	1
A	1	1	1	0	1	1	1	1	1	0
A	1	1	1	1	1	1	0	0	1	1
A	1	1	0	1	0	1	1	1	1	1
B	0	1	0	0	0	1	0	0	0	0
B	1	0	0	0	0	0	0	0	1	0
B	0	0	0	1	0	0	0	0	0	1
B	0	0	0	0	0	0	1	1	0	0
B	0	0	1	0	1	0	0	0	0	0

Table 8 - Small, low coherence category structure

Category	Dimension									
	1	2	3	4	5	6	7	8	9	10
A	1	0	0	1	0	1	0	1	1	1
A	1	1	1	0	1	0	0	1	0	1
A	0	1	1	1	0	1	1	0	0	1
A	0	1	0	0	1	1	1	1	1	0
A	1	0	1	1	1	0	1	0	1	0
B	0	1	1	0	1	0	1	0	0	0
B	0	0	0	1	0	1	1	0	1	0
B	1	0	0	0	1	0	0	1	1	0
B	1	0	1	1	0	0	0	0	0	1
B	0	1	0	0	0	1	0	1	0	1

Table 9 - Large, high coherence category structure

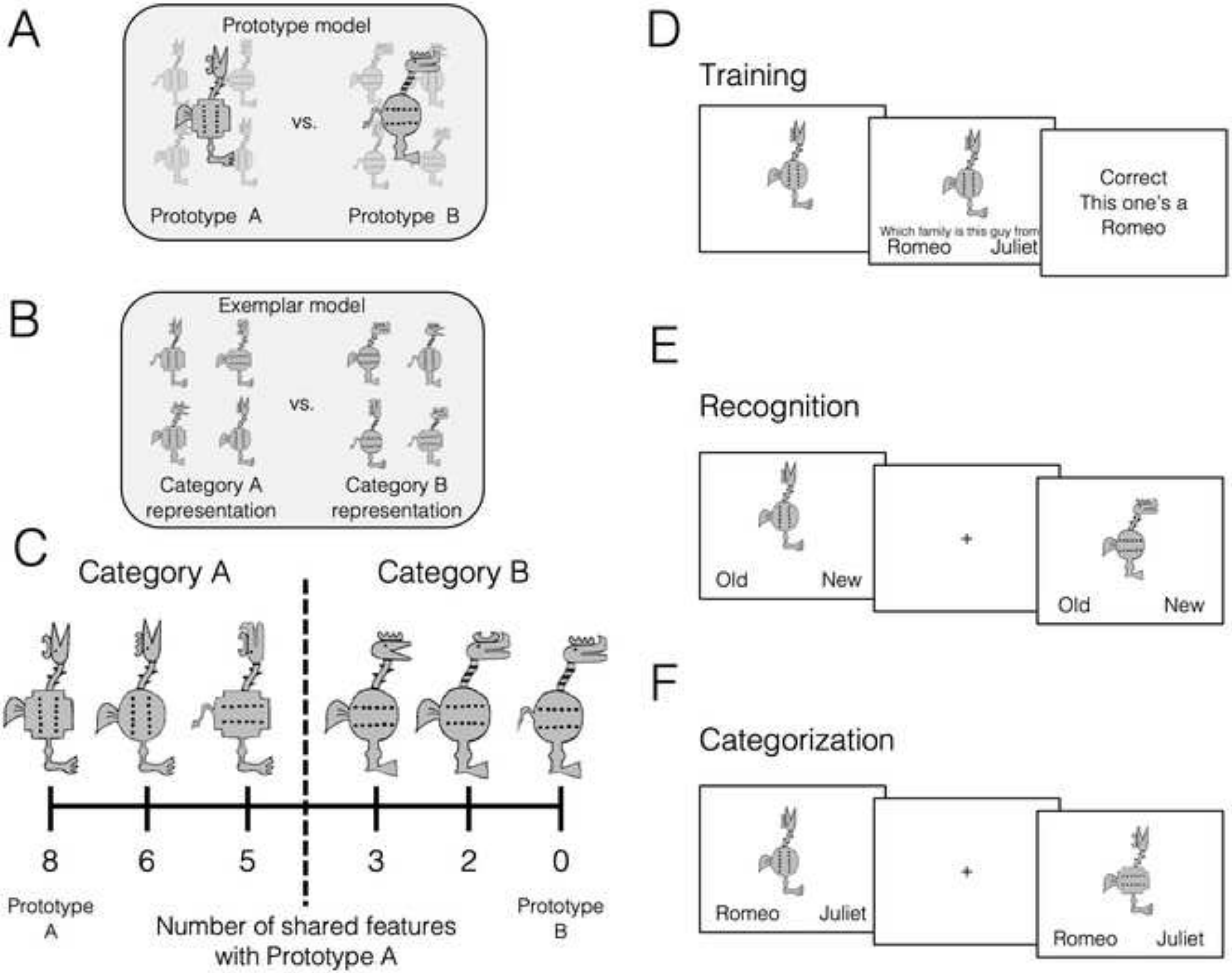
Category	Dimension									
	1	2	3	4	5	6	7	8	9	10
A	0	0	1	1	1	1	1	1	1	1
A	1	1	0	0	1	1	1	1	1	1
A	1	1	1	1	0	0	1	1	1	1
A	1	1	1	1	1	1	0	0	1	1
A	1	1	1	1	1	1	1	1	0	0
A	0	1	1	1	1	1	1	1	0	1
A	1	0	1	0	1	1	1	1	1	1
A	1	1	1	1	0	1	0	1	1	1
A	1	1	1	1	1	0	1	0	1	1
A	1	1	0	1	1	1	1	1	1	0
B	1	1	0	0	0	0	0	0	0	0
B	0	0	1	1	0	0	0	0	0	0
B	0	0	0	0	1	1	0	0	0	0
B	0	0	0	0	0	0	1	1	0	0
B	0	0	0	0	0	0	0	0	1	1
B	1	0	0	0	0	0	0	0	1	0
B	0	1	0	1	0	0	0	0	0	0
B	0	0	0	0	1	0	1	0	0	0

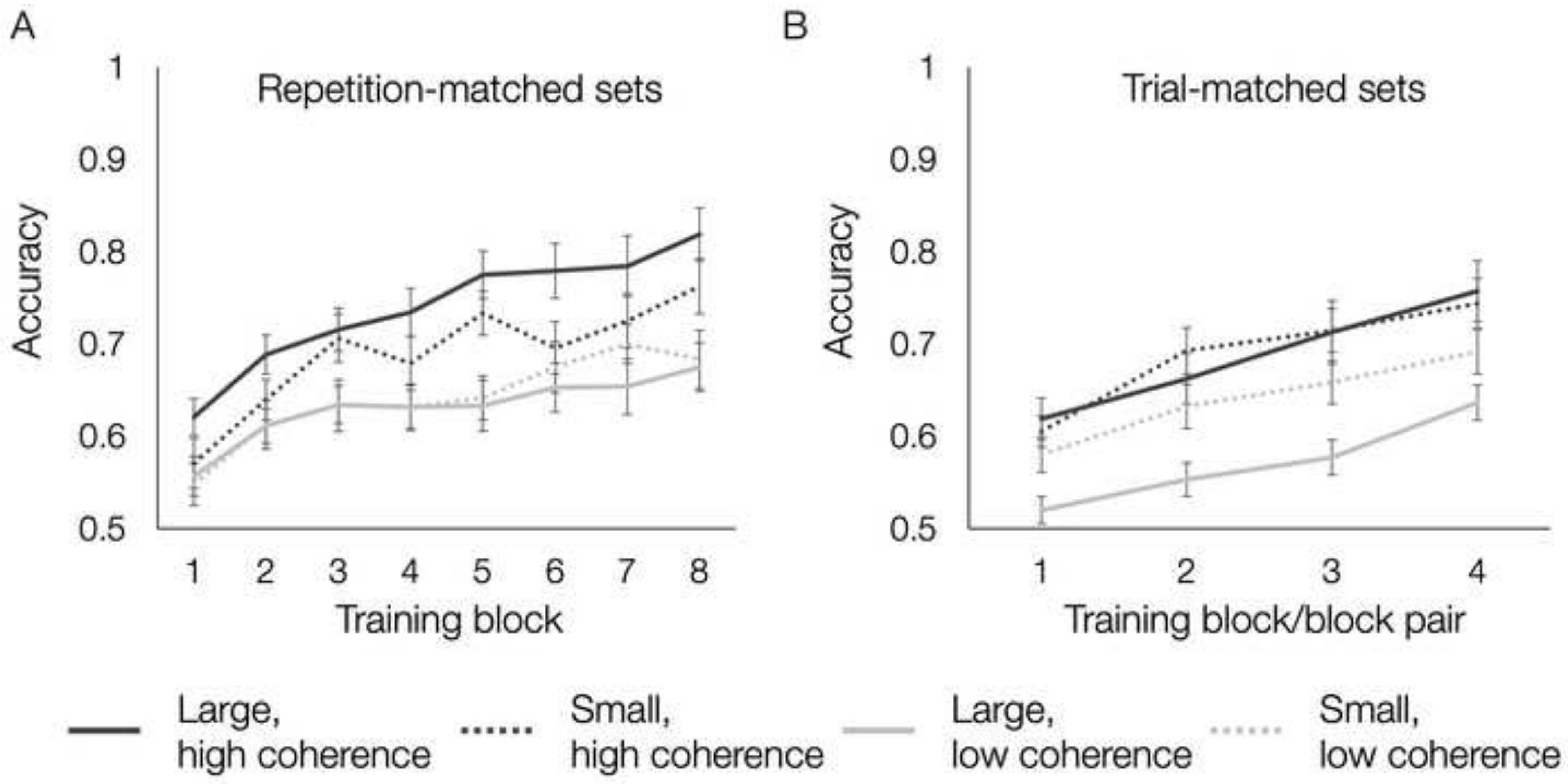
B	0	0	0	0	0	1	0	1	0	0
B	0	0	1	0	0	0	0	0	0	1

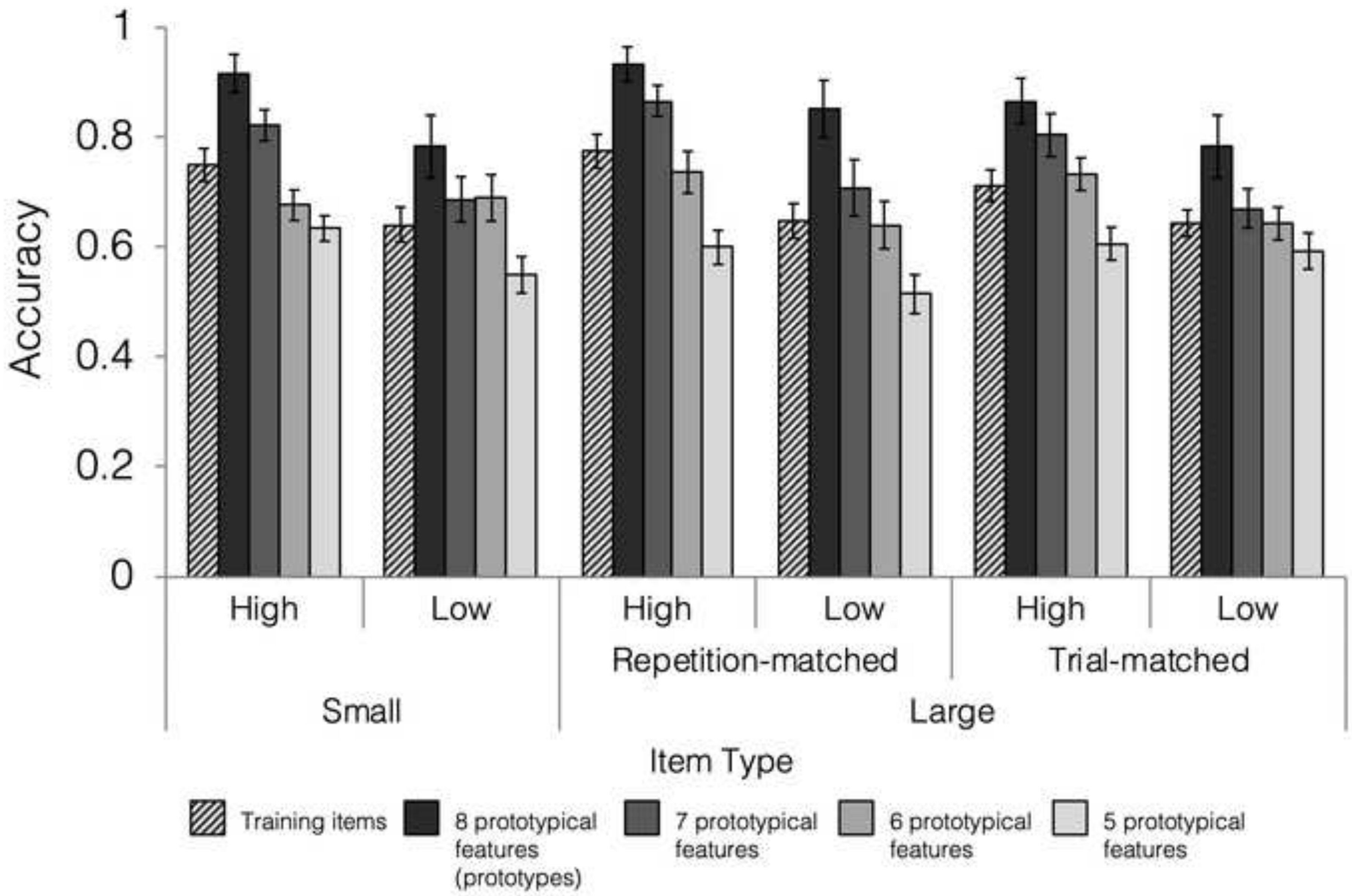
Table 10 - Large, low coherence category structure

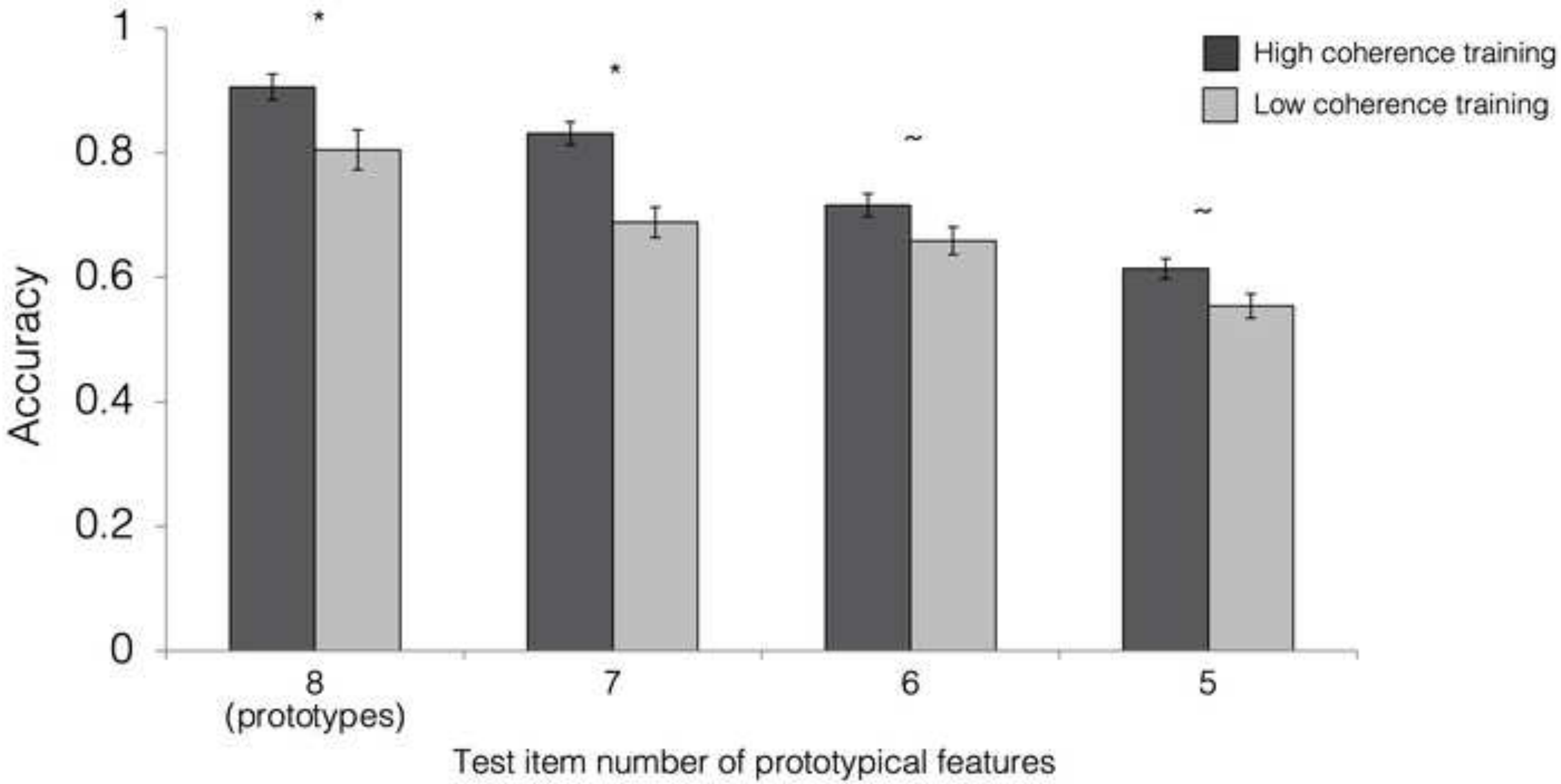
Category	Dimension									
	1	2	3	4	5	6	7	8	9	10
A	0	0	0	0	1	1	1	1	1	1
A	1	1	0	0	0	0	1	1	1	1
A	1	1	1	1	0	0	0	0	1	1
A	1	1	1	1	1	1	0	0	0	0
A	0	0	1	1	1	1	1	1	0	0
A	0	1	0	1	0	1	0	1	1	1
A	1	0	1	0	1	0	1	0	1	1
A	0	1	1	0	1	1	1	0	1	0
A	1	1	0	1	0	1	0	1	0	1
A	1	0	1	1	1	0	1	1	0	0
B	1	1	1	1	0	0	0	0	0	0
B	0	0	1	1	1	1	0	0	0	0
B	0	0	0	0	1	1	1	1	0	0
B	0	0	0	0	0	0	1	1	1	1
B	1	1	0	0	0	0	0	0	1	1
B	1	0	1	0	1	0	1	0	0	0
B	0	1	0	1	0	1	0	1	0	0
B	1	0	0	1	0	0	0	1	0	1
B	0	0	1	0	1	0	1	0	1	0
B	0	1	0	0	0	1	0	0	1	1

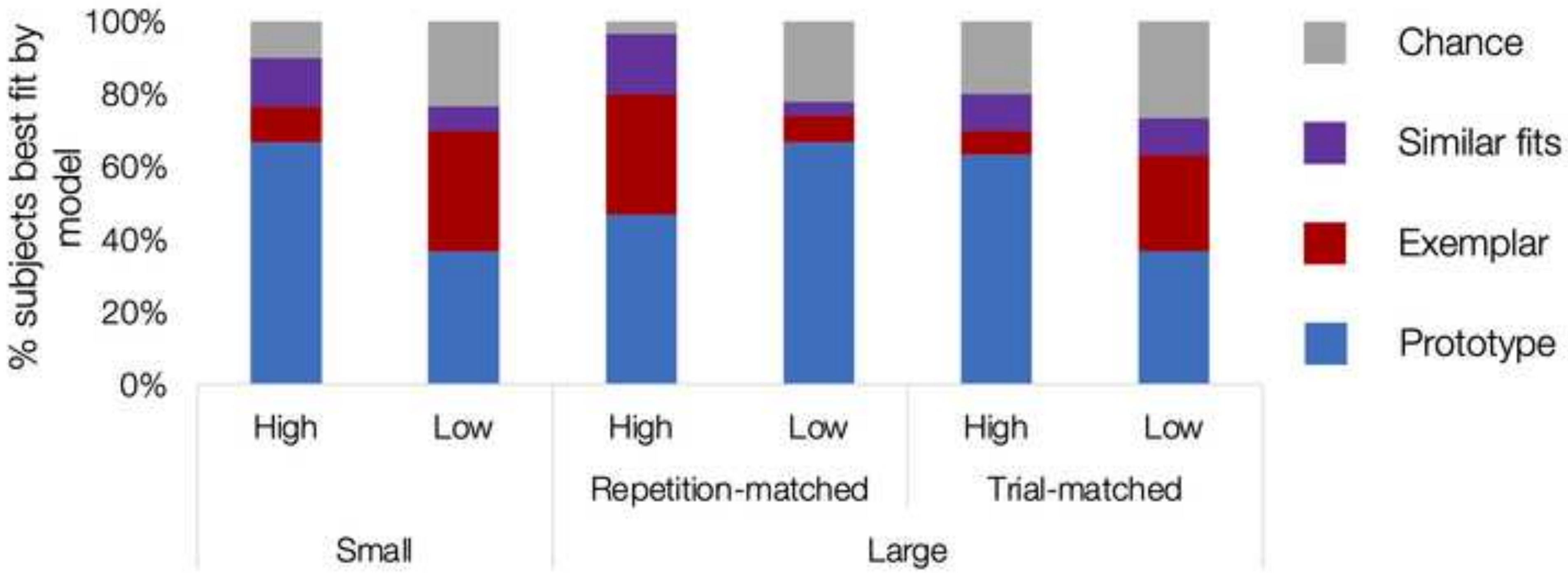
Figure1

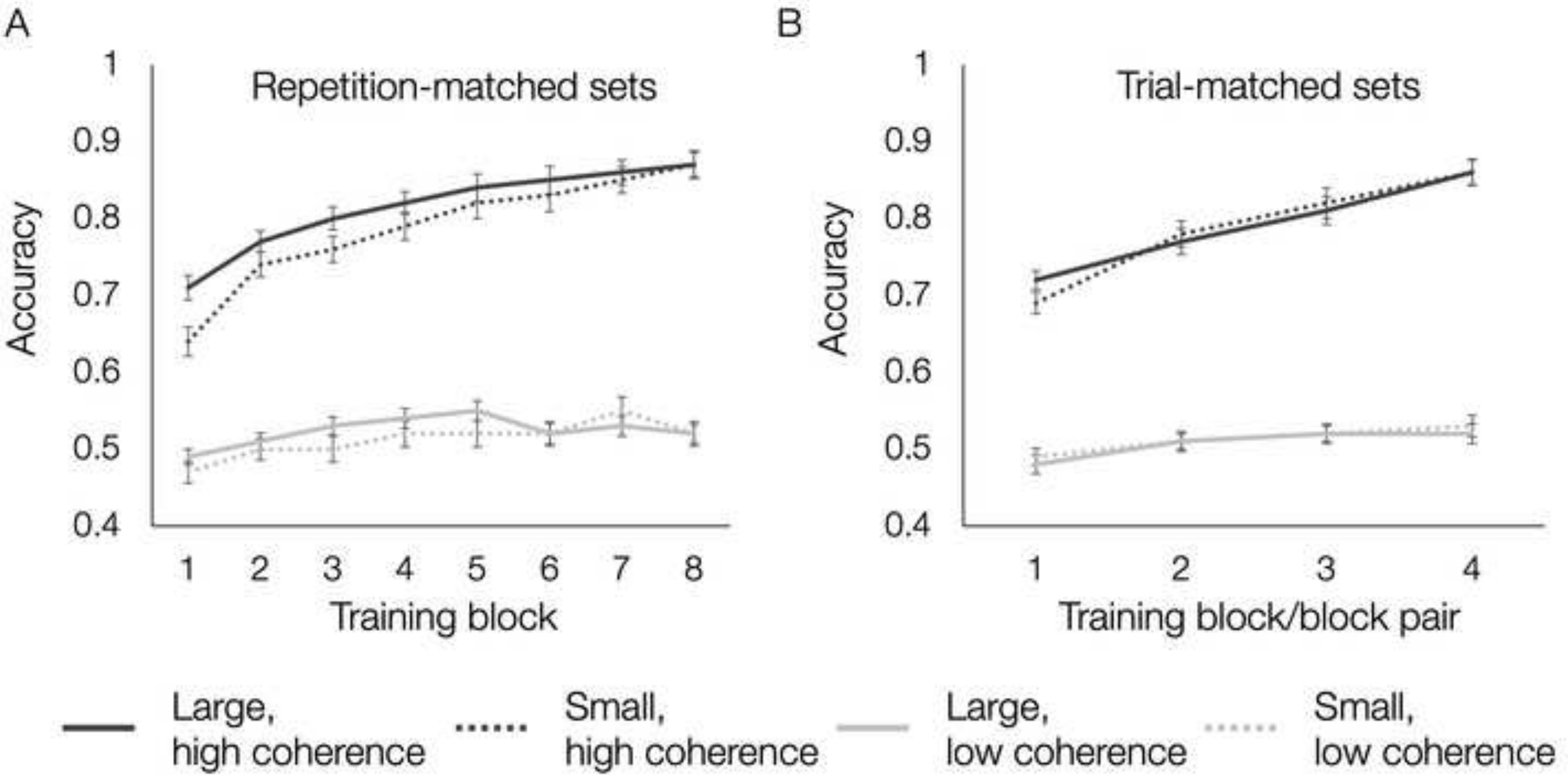

A
Prototype model
vs.
Prototype A
Prototype B
B
Exemplar model
vs.
Category A representation
Category B representation
C
Category A
Category B
8
6
5
3
2
0
Prototype A
Number of shared features with Prototype A
Prototype B
D
Training
Which family is this guy from
Romeo
Juliet
Correct
This one's a
Romeo
E
Recognition
Old
New
+
Old
New
F
Categorization
Romeo
Juliet
+
Romeo
Juliet

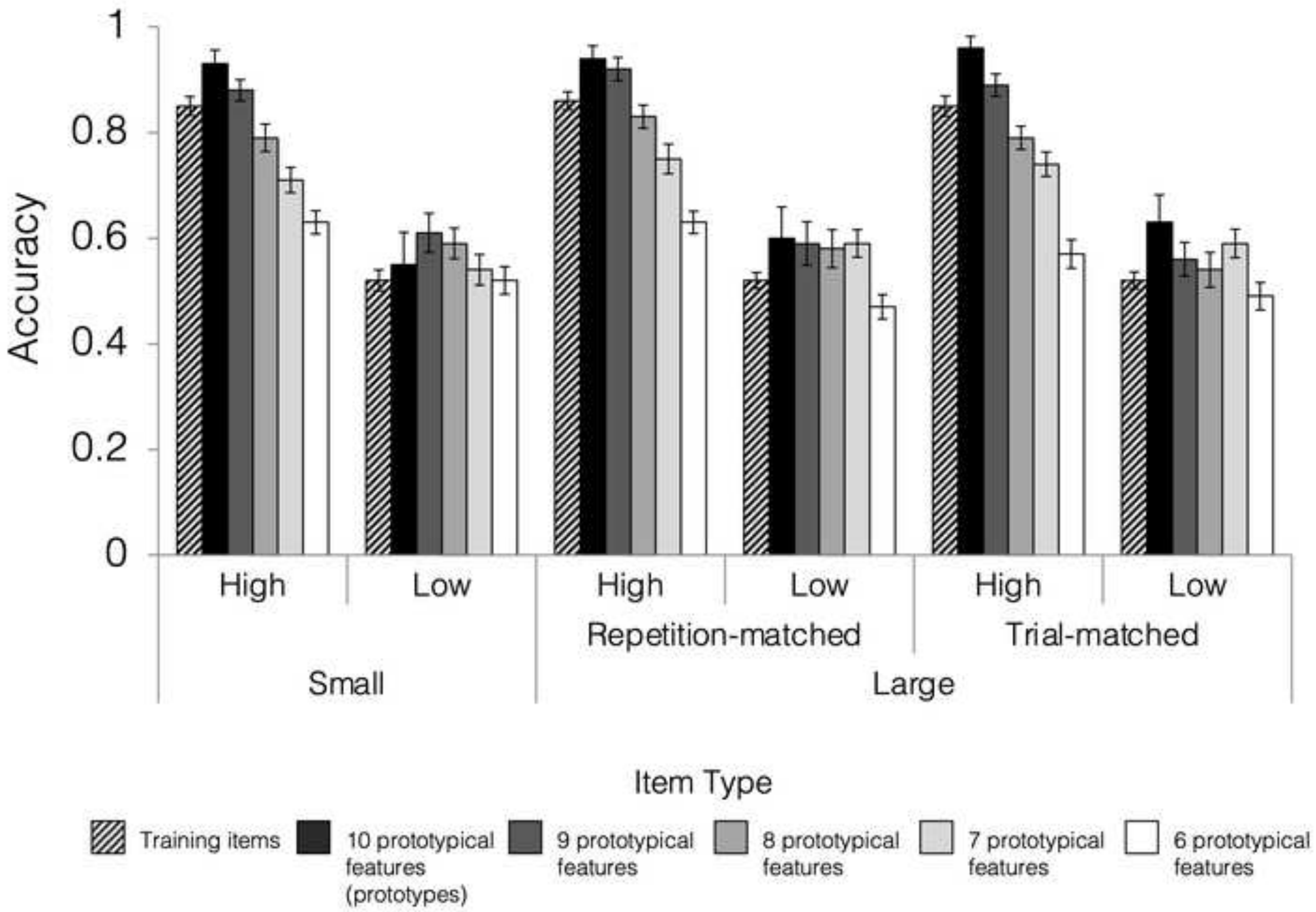

A
Repetition-matched sets
Accuracy
1
0.9
0.8
0.7
0.6
0.5
1
2
3
4
5
6
7
8
Training block
B
Trial-matched sets
Accuracy
1
0.9
0.8
0.7
0.6
0.5
1
2
3
4
Training block/block pair
Large, high coherence
Small, high coherence
Large, low coherence
Small, low coherence

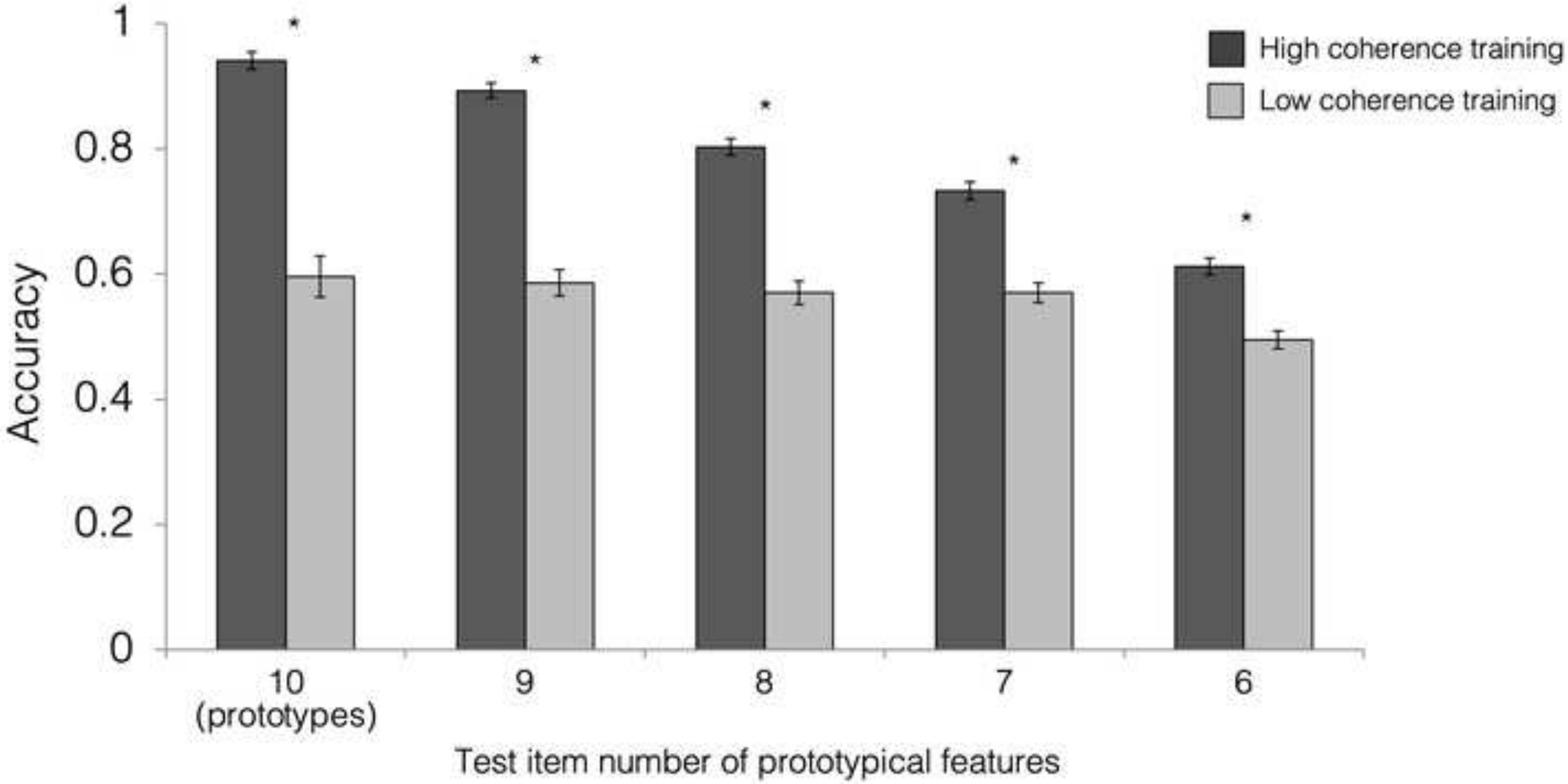

1
0.8
0.6
0.4
0.2
0
Accuracy
High
Low
High
Low
High
Low
Repetition-matched
Trial-matched
Small
Large
Item Type
▨ Training items
8 prototypical features (prototypes)
7 prototypical features
6 prototypical features
5 prototypical features

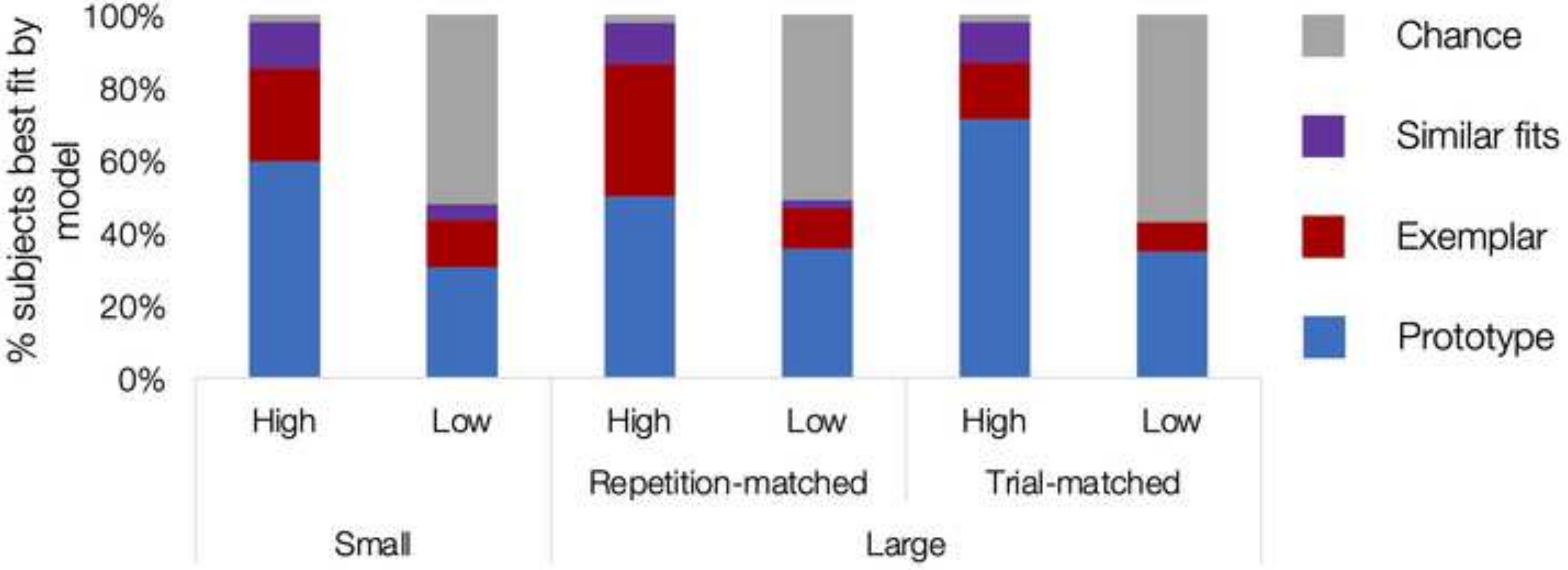

1
0.8
0.6
0.4
0.2
0
Accuracy
*
*
~
~
■ High coherence training
□ Low coherence training
8
(prototypes)
7
6
5
Test item number of prototypical features


% subjects best fit by model
100%
80%
60%
40%
20%
0%
High
Low
High
Low
High
Low
Repetition-matched
Trial-matched
Small
Large
Chance
Similar fits
Exemplar
Prototype


A
Repetition-matched sets
Accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
1 2 3 4 5 6 7 8
Training block
B
Trial-matched sets
Accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
1 2 3 4
Training block/block pair
Large, high coherence
Small, high coherence
Large, low coherence
Small, low coherence


1
0.8
0.6
0.4
0.2
0
Accuracy
High
Low
High
Low
High
Low
Repetition-matched
Trial-matched
Small
Large
Item Type
Training items
10 prototypical features (prototypes)
9 prototypical features
8 prototypical features
7 prototypical features
6 prototypical features


1
0.8
0.6
0.4
0.2
0
Accuracy
*
*
*
*
*
High coherence training
Low coherence training
10
(prototypes)
9
8
7
6
Test item number of prototypical features


% subjects best fit by model
100%
80%
60%
40%
20%
0%
High
Low
High
Low
High
Low
Repetition-matched
Trial-matched
Small
Large
Chance
Similar fits
Exemplar
Prototype