High-Variability Training Does Not Enhance Generalization

in the Prototype-Distortion Paradigm

Mingjia Hu and Robert M. Nosofsky

Indiana University Bloomington

Robert Nosofsky

Psychological and Brain Sciences

1101 E. Tenth Street

Indiana University

Bloomington, IN 47405

[nosofsky@indiana.edu](mailto:nosofsky@indiana.edu)

Running Head: Category Generalization

An issue of fundamental theoretical and practical significance in the domain of categorization is how to arrange learning conditions to promote successful generalization to novel members of the learned categories. There is a vast literature that examines the influence of multiple variables on varieties of category generalization. One of the major such variables is training-instance variability (e.g., Bowman & Zeithamova, 2023; Cohen & Nosofsky, 2001; Hahn, Bailey, & Elvin, 2005; Hintzman, 1986; Homa & Cultice, 1984; Homa & Vosburgh, 1976; Minda & Smith, 2001; Peterson, Meagher, Herschel, & Gillie, 1973; Posner & Keele, 1968; Stewart & Chater, 2002). In the present article our focus is on the influence of training-instance variability in promoting successful generalization in cases involving prototype-based categories. These are categories defined around central prototypes in which training instances are constructed by distorting the prototypes using statistical-distortion procedures. The question is whether it is better to learn such categories by being trained on low-variability instances that closely resemble the prototypes or on higher-variability instances that are not highly similar to the prototypes. Here, we are restricting consideration to cases in which all the to-be-learned categories share the same amount of training-instance variability.

Early classic articles in the human categorization literature suggested that although initial learning is fostered by training with low-variability instances, subsequent generalization to novel high-distortion category members is fostered if observers are trained with higher-variability instances. Much of this evidence came from the seminal “dot-pattern prototype-distortion” paradigm developed by Posner and Keele (1968, 1970) and pursued in comprehensive fashion by Homa and his colleagues (e.g., Homa, Blair, et al., 2019; Homa & Cultice, 1984; Homa, Sterling, & Trepel, 1981; Homa & Vosburgh, 1976). In typical versions of this paradigm, category prototypes are defined by placing dots in random locations of a grid (and often connecting the dots sequentially with lines to form polygons). Training instances are then constructed by using a statistical-distortion algorithm on those prototypes. Low distortions that are highly similar to the prototype are constructed by displacing the individual-dot locations by small magnitudes in random directions; higher-level distortions are constructed by displacing the dot locations by larger magnitudes. Posner and Keele (1968) found that generalization to novel high-distortion instances was fostered if participants were trained using high-distortion instances and they suggested that participants learned about both the prototypes and the variability of the categories. Homa and Vosburgh (1976) jointly manipulated the variability of the instances and the size of the to-be-learned categories (i.e, the number of training instances in each category) and found that generalization to novel instances of the larger-size categories was fostered with higher-variability training.

The general take-home message from these classic studies was summarized by Homa, Sterling, and Trepel (1981, p. 420) who wrote: “…when the within-category stimulus variance is increased for categories defined by numerous exemplars (Homa & Vosburgh, 1976), generalization to new instances is improved”. This general take-home message has survived to the present day (e.g., Doyle & Hourihan, 2016; Raviv, Lupyan, and Green; 2022). For example, Raviv et al. (2022) presented a comprehensive review of how different forms of variability influence wide varieties of learning and generalization. Included in their review was how variability of training instances influences generalization in the prototype-distortion paradigm. One of their primary lead-off examples (see Raviv et al., 2022, Figure 1) was the finding from Posner and Keele (1968) that, although initial learning was more difficult, generalization was enhanced when participants were trained with high-variability instances.

However, in influential work that introduced a form of exemplar-based modeling of category learning and representation, Hintzman (1984, 1986) pointed to an important limitation of these early studies: participants had been trained to a performance criterion prior to their generalization performance being tested in a transfer phase. The training phase tended to take longer for participants trained on higher-variability instances. Thus, the length of the training phase tended to be confounded with the manipulation of the factor of training-instance variability. Hintzman (1984, 1986) noted that his exemplar-based MINERVA-2 model predicted that if the amount of training were to be equated, low-variability training sets should lead to better generalization than high-variability sets in the prototype-distortion paradigm.

Bowman and Zeithamova (2020, 2023) have pursued this issue in recent studies using an alternative version of the prototype-distortion paradigm (see also Minda & Smith, 2001). Rather than using dot-pattern stimuli, Bowman and Zeithamova used cartoon-animal stimuli defined along binary-valued dimensions. The prototype of Category A might have logical-value 1 along each of its dimensions, whereas the prototype of Category B might have logical-value 2. Low- and high-distortions of the prototypes are constructed by varying the number of dimensions in which the prototype value is switched. In accord with the type of prediction derived from Hintzman’s (1984, 1986) modeling, under conditions in which the amount of training was equated across conditions, Bowman and Zeithamova (2020, 2023) found that generalization to novel transfer items was enhanced in training conditions using low distortions (low variability training conditions) compared to high-variability ones.

Although Bowman and Zeithamova’s (2020, 2023) evidence is intriguing, questions arise regarding the basis and generalizability of their findings. For example, in creating low- vs. high-variability instances, their paradigm simultaneously changed the diagnosticity of the component dimensions of the training stimuli. In their low-variability condition, a dimension value of 1 (2) might point to Category A (B) 80% of the time, whereas in the high-variability condition the dimension values might be only 60% diagnostic. Thus, it is unclear whether the patterns of generalization that they observed are due to the manipulation of variability per se or to changing levels of individual-dimension diagnosticity. More generally, it is unknown the extent to which Bowman and Zeithamova’s findings involving objects varying along binary-valued dimensions might generalize to objects varying in continuous-dimension spaces.

Thus, in the present work, we were motivated to revisit the seminal dot-pattern prototype-distortion paradigm to further investigate these issues. Our idea was to conduct manipulations similar to those reported in the classic studies of Posner and Keele (1968) and Homa and Vosburgh (1976) except, following Bowman and Zeithamova (2020, 2023), train participants for a fixed number of training trials across low-variability and high-variability conditions.

To our surprise, it was only after completing our experiment that we discovered that Homa and Cultice (1984) had already conducted such a study, although their findings are rarely discussed in relation to the present issues. (As will be seen, our own experiment differs substantially from the one reported by Homa and Cultice in a number of fundamental ways, so still provides a significant contribution to new knowledge in the field.) Homa and Cultice’s main interest was in comparing category learning under feedback versus no-feedback conditions in the dot-pattern prototype-distortion paradigm. (In feedback tasks, participants are provided with knowledge of the correct category following each of their category guesses during the training phase; no such feedback is provided in the no-feedback condition.) Homa and Cultice examined this issue across conditions in which both variability of training instances and category size were manipulated. Participants learned to classify prototype-distorted dot patterns into three categories. Category size was manipulated in within-subject fashion, with the individual categories represented by either 3, 6 or 9 training instances. Category variability was manipulated in between-subjects fashion, with participants trained on either low-, medium-, high-, or mixed-level distortions of the prototypes. Crucially, the number of training trials was held fixed across the different variability conditions. Following training, the participants’ generalization performance to novel members of the categories was tested. Homa and Cultice did not report analyses of the effects of the variability manipulation on generalization performance in the feedback condition, focusing instead on overall comparisons of performance across the feedback and no-feedback conditions, which was their main interest in this study. However, inspection of their Figure 2 (top panel) makes clear that – in contrast to the pattern of results reported by Posner and Keele (1968) and Homa and Vosburgh (1976) -- overall generalization performance was best in the low-variability training condition, intermediate in the medium condition, and worst in the high condition (with performance in the mixed-variability training condition being intermediate). Generalization to the novel high distortions was roughly the same across the conditions: there was certainly no advantage in classifying the novel high-distortion patterns for participants in the high- compared to the low-distortion training condition.

Our own experiment is similar to the feedback condition reported by Homa and Cultice (1984) but differs in several significant respects. Overall, our new experiment advances the field by lending a great deal more generality to the investigation of whether low- versus high-variability training enhances generalization in the prototype-distortion paradigm in situations in which the amount of training is equated across the conditions.

In our experiment, participants learned to classify prototype-distorted dot patterns into three prototype-based categories and their generalization performance to novel dot patterns of varying degrees of distortion was then tested. We manipulated across conditions whether participants were trained using low-, medium-, high-, or mixed-level distortions of the prototypes. As in Bowman and Zeithamova (2020, 2022) and Homa and Cultice (1984), the different variability conditions had the same number of training trials.

There were numerous differences in the design of our experiment compared to the related one used by Homa and Cultice (1984). Here we highlight three differences that we believe are highly relevant. First, in Homa and Cultice, the same three prototype patterns defined the alternative categories for all participants (see Design section of Homa & Cultice, 1984, p. 86). By comparison, in our design, the three category prototypes were generated randomly anew for each and every participant. In our view, if one is seeking generality and robustness in identifying the phenomena of interest, then our random-generation procedure seems preferable. Homa and Cultice’s design limits the inquiry to a restricted subset of the population of materials. If there were any idiosyncratic properties of the particular prototypes that were used, then the pattern of results that they observed may be limited in generality. We discuss some possibilities along these lines in subsequent sections of our article.

Second, recall that Homa and Cultice (1984) used category training-set sizes of 3, 6 and 9 in which the same instances were presented repeatedly across the different blocks of training. By contrast, in our design, following recent studies reported by Homa, Blair et al. (2019) and Hu and Nosofsky (2021), no single training instance was ever repeated. Instead, novel distortions of the prototypes were displayed on each and every training trial. As argued by Homa, Blair et al. (see also Ashby & Maddox, 1993), when people learn numerous types of categories in the real world, such as birds, trees, faces, and so forth, it appears that only a tiny proportion of the training instances are ever exactly repeated. Hence, the non-repeating training paradigm seemed like an important one to test in the present situation. Furthermore, given Homa and Vosburgh’s (1976) finding that the benefits of high-variability training occurred mainly for large-size categories, the use of non-repeating sets of training exemplars would seem to be friendly to the high-variability hypothesis. Indeed, there is an important theoretical basis for testing the non-repeating paradigm. As described in more detail in the Formal Modeling section of our article, according to exemplar models, people classify novel test patterns on the basis of their similarity to training instances of the alternative categories (e.g., Medin & Schaffer, 1978; Hintzman, 1986; Nosofsky, 1986). If only a small set of high-distortion training patterns are presented during study, then they may occupy relatively sparse regions of the high-dimensional similarity spaces in which the present types of dot patterns are presumably embedded. Hence, in testing generalization to novel high-distortion test patterns, the use of high-distortion training may not be beneficial because the novel high-distortion test items may not be very similar to the old training instances. By using large sets of non-repeating training instances, we may increase the chances of having the high-distortion training instances “cover” the training space in a denser fashion, thereby potentially promoting successful generalization under high-variability training conditions.

A third important difference between Homa and Cultice’s (1984) paradigm and ours is that Homa and Cultice (1984) allowed participants to use a “None” response during the transfer tests: If participants believed that a test pattern did not belong to any of the three trained categories, they could respond “None”. (In addition, Homa and Cultice included foil patterns in their test phase that were generated from two new prototypes patterns that had not been used to generate instances during the training phase.) Although the use of “None” responses is a potentially important vehicle for investigating the detailed nature of people’s category representations (e.g., see Hahn et al., 2005), we decided not to follow that procedure here, instead requiring participants to classify all test patterns into one of the three candidate categories. Our concern was that making use of the “None” response might lead to under-estimates of the participants’ category knowledge across the different experimental conditions, especially in testing the ability of participants to correctly classify high-distortion test patterns under low-distortion training conditions. In particular, when presented with a high-distortion test pattern, a participant may realize that it far less similar to the target category representation than are the low distortions with which they were trained. On this basis, they might respond “None” if that option were available. However, such participants might also realize that, relatively speaking, the high distortion is more similar to the category representation of its target category than to the contrast categories. Such information would allow them to correctly classify the novel high distortion. To avoid under-estimation of this form of category knowledge, we decided not to allow “None” responses in the present experiment.

Before proceeding, we should re-emphasize that the present design deliberately introduces a type of confound into the testing situation (see also Bowman & Zeithamova, 2020, 2022). Because we are equating the total number of training trials across conditions, it will almost certainly be the case that, in the training phase itself, terminal learning performance will be worse in the high-variability condition than in the low-variability one. As reviewed by Raviv et al. (2022), such a pattern is observed ubiquitously across numerous learning paradigms that manipulate training-instance variability. The question is, despite this type of confound, might generalization to novel high-distortion instances *still* be enhanced in the high-variability training conditions (Raviv et al., 2022). To repeat, for numerous previous paradigms in the category-learning literature that addressed the variability question, it was the reverse confound that was involved: the amount of training in the high variability condition exceeded the amount of training in the low-variability one. It may be nearly impossible to design an instance-variability-training experiment in which both confounds are not present.

Finally, although our central focus in the present research is empirical in nature, we pursued an important formal-modeling goal as well. In particular, we evaluated the extent to which extant formal models of human category learning could capture the observed results. As we discuss in detail in the Formal Modeling section, this evaluation should advance the further development of such models.

Method

Subjects

The study was approved by the Indiana University Institutional Review Board. The subjects were 304 students from Indiana University who participated in partial fulfillment of an undergraduate psychology course requirement. Subjects were randomly assigned to four training conditions. There were 77 subjects in the low-distortion condition, 78 in the medium-distortion condition, 75 in the high-distortion condition and 74 in the mixed-distortion condition. All subjects had normal or corrected-to-normal vision.

Stimuli and apparatus

The stimuli were dot patterns generated using Posner et al.’s (1967) procedure (i.e., the algorithm used by Posner & Keele, 1968). For each individual subject, prototypes for three different categories were generated by placing 9 dots at random grid positions in the central 30 × 30 area of a 50 × 50 grid. Different training and transfer patterns of each category were generated using the statistical-distortion procedure of Posner et al. (1968). Each pattern was constructed from the prototype of its category by displacing each dot by a random direction and distance in accordance with the Posner et al. procedure. Low-, medium-, and high-level distortions were generated by moving the individual dots, on average, 4, 6 and 7.7 Posner-levels away from their prototype. Each individual subject was presented with a unique set of randomly generated prototypes and training and transfer patterns.

We used Dell computers to display the stimuli and control the experiment. The patterns were white in color and displayed at the center of a gray computer screen.

Procedure

A standard learning-transfer paradigm was used. In the learning phase, subjects were trained to classify a set of training patterns into three categories, A, B and C. On each trial, a dot pattern was presented at the center of the computer screen and remained visible until a subject responded with a key press. After the response, corrective feedback appeared for 2s below the presented pattern. In all conditions, the learning phase consisted of 10 training blocks, each of which had 27 trials (270 trials total). A different set of training patterns was randomly generated in each of the 10 training blocks. No individual training item was ever repeated. The learning phase was followed shortly after by a transfer phase where subjects classified selected novel patterns as well as a subset of training patterns into the same three categories. No corrective feedback was given on any test trial.

For each individual participant, three random prototypes were generated to define each of the three categories. Participants were randomly assigned to one of the four training conditions described above. In each condition, 90 training patterns (9 per block) were randomly generated around each of the three category prototypes (270 patterns in total). The category prototypes were distorted using the Posner-Keele (1968) statistical-distortion algorithm to generate the training patterns for the four training conditions: all low-distortions, all medium-distortions, all high-distortions, and a mixture of the three distortion levels. In all conditions, an equal number of training patterns from each category was presented in each block. In the mixed condition, there was an equal number of low, medium and high distortions from each category in each block. In all conditions, order of presentation of the generated patterns was fully randomized within each block.

The test patterns consisted of 27 old patterns that were presented in the training phase (9 per category, with at least 2 of the 27 patterns from each of the 10 training blocks), 3 prototypes (1 per category), 9 new low-level distortions (3 per category), 18 new medium-level distortions (6 per category), and 27 new high-level distortions (9 per category). Each pattern was presented once in a random order for each subject for a total of 84 test trials. We decided to use fewer test trials for prototypes and low distortions than for medium and high distortions for two reasons. First, because all low distortions within a category are highly similar to one another (and to their prototype), there is the possibility that observers can learn the category structure at time of test if there are too many presentations of the prototypes and low distortions (e.g., Palmeri & Flanery, 1999; Zaki & Nosofsky, 2007). Second, in the case of larger-size categories, Homa and Vosburgh (1976, Figure 1) found that the main benefits of higher-distortion training on subsequent generalization were for the medium- and high-distortion test patterns, so those seemed the more important patterns to test.

Results

Learning. Figure 1 shows the average proportion of correct classification responses over the training blocks for each of the four training conditions. Across all the training conditions, classification accuracy gradually improves over the course of training. The low-distortion training condition shows the highest accuracy, the medium- and mixed-distortion conditions show intermediate levels of accuracy, while the high distortion condition shows the lowest accuracy. To confirm these observations, we conducted a 4x10 mixed-model ANOVA using training conditions (low, med, high, mixed) and blocks as factors. The analysis revealed a significant main effect of blocks, F(6.35, 1905.07) = 84.44, p < .001, η2 = .220 (the Greenhouse-Geisser correction was applied for violation of the sphericity assumption). The main effect of training conditions was also significant, F(3,300) = 82.85 , p < .001, η2 = .453, as was the interaction effect between learning condition and blocks, F(19.05, 1905.07) = 2.865, p < .001, η2 = .028. To confirm the differences in classification performance across training conditions near the end of the training phase, we compared the mean proportion of correct responses for the last three training blocks. The analysis showed that the mean proportion of correct responses is higher in the low condition (M = 0.885) than in the medium condition (M = 0.691), t(141.2) = 7.55, p < .001, and higher in the medium condition than in the high condition (M = 0.499), t(149.4) = 6.92, p < .001.

Transfer. The left panel of Figure 2 shows the mean proportion of correct responses for the different types of test patterns as a function of training conditions. The general trend is that, in all conditions, classification accuracy is the highest for the prototypes, and decreases in the order of low-, medium- and high-level distortion test patterns. In addition, the general trend is that overall transfer performance is best in the low-distortion training condition, intermediate in the medium- and mixed-distortion training conditions, and worst in the high-distortion training condition. Among our key questions was whether higher-distortion training might specifically benefit generalization performance to higher-distortion transfer items, as has been reported in cases in which participants were trained to a common learning criterion across conditions (Homa & Vosburgh, 1976; Posner & Keele, 1968). Inspection of Figure 2, left panel reveals clearly that such a result was not observed in the present paradigm. Instead, the novel high distortions were classified with notably *lower* accuracy in the high-distortion training condition than in the three other conditions. In addition, the novel medium distortions were also classified with *lower* accuracy in the medium-distortion and mixed-distortion training condition than in the low-distortion condition. Another question of interest, which is potentially relevant to evaluating predictions from formal models, concerns transfer performance on the old-distortion training patterns compared to novel patterns of the same level of distortion. Although most dot-pattern category-learning paradigms that use small category sizes show patterns of results in which there is an old-item advantage, such does not appear to be the case in the present paradigm in which individual training instances never repeated and category size is extremely large (see Table 1).

To confirm these observations, we first conducted a 4x4 mixed-model ANOVA, using condition (low, medium, high, mixed) and novel pattern type (prototype, new-low, new-medium, new-high) as factors. The analysis revealed a significant main effect of pattern type, F(2.62, 779.36) = 128.5, p < .001, η2 = .092; a significant main effect of training condition, F(3,300) = 15.35, p < .001, η2 = .091; and a significant interaction between the two factors, F(7.79, 779.36) = 4.4, p < .001, η2 = .010. In more focused tests of interest, we found that for the novel high-distortion patterns, the mean proportion of correct responses is significantly lower in the high condition (M = .512) than in the medium condition (M = .631), t(150.7) = 4.024, p < .001, the mixed condition (M = .593), t(146.8) = 2.655, p = .009, and the low condition (M = .637), t(135.5) = 4.786, p < .001. For the novel medium-distortion patterns, the mean proportion of correct responses is significantly lower in the medium condition (M = .692) than in the low condition (M = .771), t(146.8) = 2.631, p = .009.

One possibility is that our failure to observe generalization enhancement for high-distortion patterns under higher-variability training may arise because of our random prototype-generation procedure. By chance, for some participants, some of the random prototypes of the alternative categories may have been difficult to discriminate. In such cases, high variability training may have caused the “noise” to overwhelm the “signal” rather than providing generalization enhancements. We do not have a rigorous theory of dot-pattern similarity at the level of individual items to allow us to evaluate this hypothesis. However, in follow-up analyses, we examined overall patterns of performance after deleting fixed proportions of participants from each condition who showed low overall classification accuracy. (Although there are multiple reasons why individual participants may perform poorly in this paradigm, one of the major reasons is that they were tested with difficult-to-discriminate categories.) The overall pattern shown in Figure 2, left panel remained the same when we considered only the top 90% of participants in each condition and only the top 50% of participants in each condition.

Model-Based Accounts of the Classification Transfer Data

In this section of our article we report preliminary model-based accounts of the patterns of observed classification transfer data, with a focus on the types of exemplar and prototype models that are often contrasted in the dot-pattern prototype-distortion paradigm (e.g., Busemeyer, Dewey, & Medin, 1984; Hintzman, 1986; Homa, et al., 1981; Homa, Blair et al., 2019; Hu & Nosofsky, 2021; Palmeri & Flanery, 1991; Shin & Nosofsky, 1992). As argued by Hu and Nosofsky (2021), because the underlying psychological dimensions of the dot-pattern stimuli are unknown, it is currently not advisable to develop rigorous quantitative model-based comparisons in this domain. Nevertheless, it seems reasonable to test the extent to which extant modeling approaches capture the broad qualitative patterns of results. Following Hintzman’s (1984, 1986) influential style of modeling performance in the dot-pattern paradigm, we pursue this goal here by using computer-simulation methods intended to produce category structures analogous to those that are thought to be produced through use of the Posner-Keele prototype-distortion procedure. As a representative from the class of exemplar models, rather than using Hintzman’s (1986) MINVERVA-2 model, we instead use Hu and Nosofsky’s (2021) simulation-based version of Nosofsky’s (1986) *generalized context model* (GCM). As argued by Jamieson et al. (2022), when varieties of formal models from the same general theoretical class (such as exemplar models) yield converging predictions, it reinforces more strongly the general principles that those models are intended to formalize.

General Simulation Procedure

Following Hu and Nosofsky (2021), in our main simulations, we represent the dot patterns as points in a six-dimensional psychological space. This is consistent with multidimensional-scaling (MDS) studies by Shin and Nosofsky (1992) who reported that six-dimensional MDS solutions provided good accounts of similarity-ratings data among prototype-distorted dot-pattern stimuli. For each simulation and for each category, a prototype was generated by randomly choosing a value in the range [0, *between*] along each of the six dimensions. The free parameter *between* controls the degree to which different category prototypes are dissimilar from one another other. In general, relatively larger values of *between* will result in larger distances among category prototypes, hence less similarity between categories.

For each category, the statistical distortions (*x*) were then generated by sampling random *z* scores from a standard normal distribution and adding scaled values of *z* to the dimension values of the corresponding category prototype (*P*). Different scaling factors were used to represent different levels of distortions, as follows:

*xim* = *Pim* + *within*\**low*\**z*, for low distortions

*xim* = *Pim* + *within*\**medium*\**z*, for medium distortions

*xim* = *Pim* + *within*\**high*\**z*, for high distortions (1)

In Equation 1, *Pim* denotes the value of prototype *i* on dimension *m*, and *xim* denotes the value of a statistical distortion generated from prototype *i* on dimension *m*. The parameter *within* is a freely estimated scaling parameter that primarily determines the relative degree of *within*-category dissimilarity (i.e., how dissimilar the statistical distortions of a category tend to be from their generating prototypes as well as from one another). The parameters *low*, *medium* and *high* specify the magnitude of low, medium and high distortion levels. To reduce the number of free parameters, in a baseline version of the model we set these values at the average dot-distance movements produced by the Posner-Keele statistical-distortion algorithm, reported by Homa et al. (2019) to be *low* = 1.20, *medium* = 2.80, and *high* = 4.60.

Exemplar Model of Categorization

Once the patterns are created for each individual simulation, the standard equations of the GCM (Nosofsky, 1986, 2011) are used to generate the classification predictions in the transfer phase. According to the GCM, the probability that pattern *i* is classified into category A is found by summing its similarity to all the training examples *a* that belong to category A (*sia*), and dividing by the summed similarity of *i* to all the training examples of all categories:

(2)

where the parameter γ is a response-scaling parameter. When γ=1, observers respond by probability-matching to the relative summed similarities of each category; as γ grows larger in magnitude, observers respond more deterministically with the category that yields the largest summed similarity (Ashby & Maddox, 1993).

The similarities between patterns are computed as follows. First, for the present types of stimuli, the standard Euclidean distance formula is used to compute the distance between test pattern *i* and training example *j*,

(3)

where and denotes the values of patterns i and j on dimension *m*, respectively. Next, following Shepard (1987), the similarity between test pattern i and training example *j* (*sij*) is assumed to be an exponential-decay function of their distance,

(4)

where *c* is a sensitivity parameter that describes the rate at which similarity declines with distance. The sensitivity parameter provides a measure of overall discriminability among patterns in the feature space.

In the present formalization, it turns out that the parameters *between*, *within*, and *c* cannot be estimated separately from one another, one can estimate only their relative values. Without loss of generality, here we set *between* = 2 for scaling convenience. Thus, the present baseline version of the exemplar model uses only 3 free parameters: *within* (Equation 1), γ (Equation 2), and *c* (Equation 4).

We fitted the exemplar model to the classification-transfer data of Figure 2, left panel by searching for the values of the free parameters that minimized the sum of squared deviations between the predicted and observed correct-classification probabilities across all the item types across all the training conditions. (The three free parameters were constrained to be held fixed across the different training conditions.) We conducted 10,000 simulations to generate the predictions. For each individual simulated subject, the set of predicted classification probabilities was computed. We then averaged across the predictions from the 10,000 simulated subjects to generate the predictions of the averaged data in Figure 2, left panel. We used the Hook and Jeeves (1961) parameter search to derive the values of the best-fitting parameters and used 10 different starting configurations for the parameter search.

The predictions from the baseline exemplar are model are shown in the middle panel of Figure 2. (For viewing convenience, the observed data are shown again in the left panel.) The best-fitting parameters and summary fit are reported in Table 2. Inspection of Figure 2 reveals that the exemplar model does a good job of capturing the main qualitative trends in the data, predicting correctly the main effects of test-pattern type, training condition, and the form of the interaction between these variables. As can be seen, it also predicts only very small differences in accuracy between the old training patterns and new transfer patterns at the same distortion level, which is the basic pattern seen in the data – for details, see Table 1. (The reason is that the perfect similarity-match of an old test item to its own memory representation is swamped by its similarity to the 89 other training items from its category, such that there is little difference in summed similarity for old and new test patterns at the same level of distortion.) Finally, it predicts correctly that the new high distortions are classified with lower accuracy in the high-distortion training condition than in the low, medium, and mixed training conditions; and that the new medium distortions are classified with lower accuracy in the medium and mixed training conditions than in the low-distortion training condition. Hence, the model predicts correctly the central result of interest observed in our study: when amount of training is equated across conditions, higher-variability training does not lead to enhanced generalization of higher-variability test patterns in this prototype-distortion paradigm.

From a quantitative perspective, the baseline model tends to overestimate overall accuracy in the high-distortion training condition and to underestimate overall accuracy in the low-distortion condition. By allowing the *low*, *medium*, and *high* scaling parameters in Equation 1 to vary freely rather than being held fixed at their default values, these quantitative predictions can be improved, as reported in Table 2. (This extended model uses only 2 additional free parameters rather than 3, because now one can set *within*=1 without loss of generality.) In our view, however, the more important result is that the model appears to capture in natural fashion the main qualitative pattern of results observed in the paradigm.

A question that arises is the extent to which the qualitative pattern of predictions in Figure 2 is robust across different parameter settings. To explore this question, we generated predictions from the baseline model that varied the settings of the *within*-category scaling parameter and the sensitivity parameter *c*. We set both parameters at their best-fitting values, at values 50% below the magnitude of their best-fitting values, and at values 50% greater than their best-fitting values. The predictions for each combination of parameter values are shown in Figure 3. As can be seen, although varying these free parameters results in changing the overall levels of accuracy, the overall pattern of qualitative predictions remains the same.

Prototype Model of Categorization

For completeness, it is also interesting to consider the predictions of baseline versions of the types of prototype models that are often fitted to classification data (e.g., Bowman & Zeithamova, 2023; Minda & Smith, 2001; Nosofsky, 1987; Palmeri & Flanery, 1999; Reed, 1972). The prototype of a category is generally formalized as the central tendency of the category, computed across all the category’s training instances. Within the present modeling framework, the prototype model is the same as the exemplar model, except that instead of summing similarities of test items to the individual training exemplars, one computes their similarity to the prototypes. Thus, according to the prototype model, the probability that pattern *i* is classified into category A is given by

(5)

where is the similarity of test pattern *i* to the prototype of category A. These similarities are computed in analogous fashion as in the exemplar model through use of Equations 3 and 4. Note that within the framework of the prototype model, the response-scaling parameter γ cannot be estimated separately from the sensitivity parameter *c* (as defined in Eq. 4), so γ is fixed at 1 for the prototype model (for extensive discussion, see Nosofsky & Zaki, 2002). Thus, the baseline prototype model uses two free parameters: the *within*-category scaling parameter (Equation 1) and the sensitivity parameter *c* (Equation 4).

We fitted this baseline prototype model to the classification transfer data using the same methods as already described for the exemplar model. For each simulation, we computed *empirical* prototypes of Categories A, B and C by averaging across the simulated training distortions and then applying Equations 5, 3 and 4 to generate the predictions. The best-fitting predictions from the baseline model that we just described are shown in the right panel of Figure 2, with best-fitting parameters and summary fits reported in Table 2. As can be seen, the baseline prototype model captures the typicality gradient within each condition, with the prototypes predicted to be the most accurately classified pattern, followed in order by the low, medium, and high distortions. However, the baseline model fails to predict the effect of training condition on performance, with the predicted performance curves essentially overlapping across the different training conditions. The reason for this prediction should be self-evident: with the present large-size samples of training instances, the computed central tendency of each category is nearly the same regardless of whether participants are trained with low-variability or high-variability training instances. For the same reason, it makes little difference if the baseline model is extended by allowing the parameters *low*, *medium*, and *high* to be freely estimated rather than set at their default values (see Table 2).

Of course, we do not conclude that the present data rule out all members of the class of prototype models. For example, a prototype theorist might argue that our method of creating simulated training instances for the present paradigm is flawed. Although we accept this criticism, we note that numerous other methods of coding the patterns might also yield nearly identical central tendencies across the low-, medium- and high-distortion training conditions. Alternatively, prototype theorists could argue that there are psychological limits on the number of training instances that can be averaged to compute the central tendency of each category. In that case, with limited psychological sample sizes, training with low distortions is more likely to yield empirically computed prototypes that better approximate the objective prototypes used to generate the categories. Still another possibility is that there is variation in the values of the free parameters across the different training conditions. For example, it would be straightforward to fit the present data with a prototype model simply by allowing the sensitivity parameter *c* to take on different values across the conditions. (In our view, rather than providing an explanation of the pattern of results, such an approach would simply re-describe them.) The crucial point is that these mechanisms have not been part of past theorizing involving prototype models. Thus, the present results may encourage further development of this class of models. In the meantime, we note that an exceedingly simple and natural application of the exemplar model does succeed in capturing the main qualitative patterns of results.

General Discussion

The present results lend considerable generality to findings reported by Bowman and Zeithamova (2020, 2022) and Homa and Cultice (1984): Under conditions in which amount of training across conditions is held constant in prototype-distortion paradigms, high-variability training does not enhance generalization to novel high-distortion test patterns. If anything, the opposite pattern appears to obtain, with low-variability training yielding an advantage. These results place in focused perspective the commonly held impression that high-variability training is beneficial to category generalization in the prototype-distortion paradigm. This impression appears to be derived from studies in which participants in high-variability training conditions were given more total trials of training than were participants in the low-variability training conditions. Future work needs to carefully specify the conditions under which a high-variability training advantage may emerge.

Our present modeling investigations converge with earlier results reported by Hintzman (1986) in suggesting that exemplar-based models naturally predict the pattern of results reported here. This finding forces a re-adjustment of our own intuitions regarding how such models would behave. Prior to conducting the modeling, our intuition was that exemplar models should predict enhanced generalization to novel high-distortion test items if training occurs with high-distortion training instances, especially in situations in which the number of training instances from each category is large, such as was the case in the present experiment. Our thinking was that under such circumstances, the training instances should “cover” the region of the similarity space in which the high-distortion test patterns are embedded, thereby allowing for more successful generalization to those patterns. Instead, our adjusted intuition is now as follows. Because the category structures in the prototype-distortion paradigm are *generated* by distorting the prototypes, an ideal observer would classify novel test patterns in terms of their similarity to the prototypes. Low-distortion training examples are far more similar to the generating prototypes than are high-distortion training examples. Hence, an exemplar-based representation consisting of low-distortion training exemplars would better approximate the ideal-observer prototype representation than would an exemplar-based representation composed of high-distortion training exemplars.

If this adjusted intuition has merit, then it leads us to a new prediction. Here, we limited consideration to cases in which categories were generated through statistical distortion of prototypes and in which all categories had the same overall levels of variability. This is an extremely special case, however, of the possible types of category structures. If observers need to learn categories that are not structured in this manner, then very different forms of category training may be optimal. If categories are not structured in a simple manner around their central tendencies, then training with only low distortions of the central tendencies is almost certain to be detrimental to successful generalization. Higher-variability training may turn out to be highly beneficial in cases involving more complex category structures.

Indeed, work reported by Wahlheim, Finn, and Jacoby (2012) involving the learning of real-world bird categories; and by Nosofsky, Sanders, Zhu and McDaniel (2019) involving real-world rock categories, is suggestive along these lines. Rather than manipulating variability per se, these researchers compared conditions involving manipulations of category size. In small-size/high-repetition conditions, observers were trained on small numbers of distinct instances each repeated numerous times; whereas in large-size/low-repetition conditions, observers were trained on large numbers of distinct instances each repeated fewer times. Total number of training trials was equated across conditions. Although performance on the original training exemplars was better in the small-size/high-repetition condition, generalization to novel transfer items was better in the large-size/low-repetition condition. Because increasing category size is likely to increase category “coverage” in a manner analogous to increasing variability of training instances, these results provide suggestive support in favor of the hypothesis advanced above. Nevertheless, future research involving the learning of real-world categories needs to manipulate the factor of training-instance variability more precisely to test whether high-variability training may truly be beneficial in such domains.

References

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37(3), 372-400

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), 1442

Bowman, C. R., & Zeithamova, D. (2023). Coherent category training enhances generalization in prototype-based categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*

Busemeyer, J. R., Dewey, G. I., & Medin, D. L. (1984). Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 638

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. Memory & Cognition, 29(8), 1165-1175.

Doyle, M. E., & Hourihan, K. L. (2016). Metacognitive monitoring during category learning: How success affects future behaviour. *Memory*, *24*(9), 1197-1207

Hahn, U., Bailey, T. M., & Elvin, L. B. (2005). Effects of category diversity on learning, memory, and generalization. *Memory & cognition*, *33*(2), 289-302

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96-101

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological review*, *93*(4), 411

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

Homa, D., & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(3), 322

Homa, D., Blair, M., McClure, S. M., Medema, J., & Stone, G. (2019). Learning concepts when instances never repeat. *Memory & cognition*, *47*, 395-41

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(6), 418

Hook, R., & Jeeves, T. A. (1961). Direct search Solution of numerical and statistical problem Journal of The Association of computing Machinery

Hu, M., & Nosofsky, R. M. (2021). Exemplar-model account of categorization and recognition when training instances never repeat. *Journal of Experimental Psychology: Learning, Memory, and Cognition*

Jamieson, R. K., Johns, B. T., Vokey, J. R., & Jones, M. N. (2022). Instance theory as a domain-general framework for cognitive psychology. Nature Reviews Psychology, 1(3), 174-183

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39

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

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

Palmeri, T. J., & Flanery, M. A. (1999). Learning about categories in the absence of training: Profound amnesia and the relationship between perceptual categorization and recognition memory. *Psychological Science*, *10*(6), 526-530

Peterson, M. J., Meagher Jr, R. B., Chait, H., & Gillie, S. (1973). The abstraction and generalization of dot patterns. Cognitive Psychology, 4(3), 378-398

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of experimental psychology*, *77*(3p1), 353

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental psychology*, *83*(2p1), 304

Posner, M. I., Goldsmith, R., & Welton Jr, K. E. (1967). Perceived distance and the classification of distorted patterns. Journal of experimental psychology, 73(1), 28.

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in cognitive sciences*, *26*(6), 462-483

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive psychology*, *3*(3), 382-407

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323

Shin, H. J., & Nosofsky, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General*, *121*(3), 278

Stewart, N., & Chater, N. (2002). The effect of category variability in perceptual categorization. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(5), 893.

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

Zaki, S. R., & Nosofsky, R. M. (2007). A high-distortion enhancement effect in the prototype-learning paradigm: Dramatic effects of category learning during test. *Memory & Cognition*, *35*, 2088-2096

Table 1. Mean proportion of correct responses for old and new test items at the same level of distortion from the prototype as a function of training condition.

**Item Type**

**Condition** Old New

Low 0.860 (0.851) 0.873 (0.845)

Medium 0.698 (0.689) 0.692 (0.676)

Mixed 0.697 (0.700) 0.683 (0.683)

High 0.532 (0.593) 0.512 (0.562)

Note. Values in parentheses are predictions from a baseline version of an exemplar model described in the Formal Models section. For the mixed condition, the results are averaged across the patterns with different levels of distortion.

Table 2. Model-fitting results for exemplar and prototype models

Free- Free-

Baseline Distance Baseline Distance

Parameter Exemplar Exemplar Prototype Prototype

*between*  (2.000) (2.000) (2.000) (2.000)

*within* 0.210 (1.000) 0.134 (1.000)

*low* (1.200) 0.109 (1.200) 0.001

*medium* (2.800) 0.415 (2.800) 0.253

*high* (4.600) 0.852 (4.600) 0.621

*c* 0.475 0.371 1.285 1.145

*γ* 5.000 5.000 (1.000) (1.000)

*# Parms* 3 5 2 4

*SSD* 0.016 0.010 0.108 0.095

Note. Parameter values in parentheses are held fixed at constrained settings. SSD = sum of squared deviations between observed and predicted classification probabilities. # Parms = Number of free parameters used by the model.

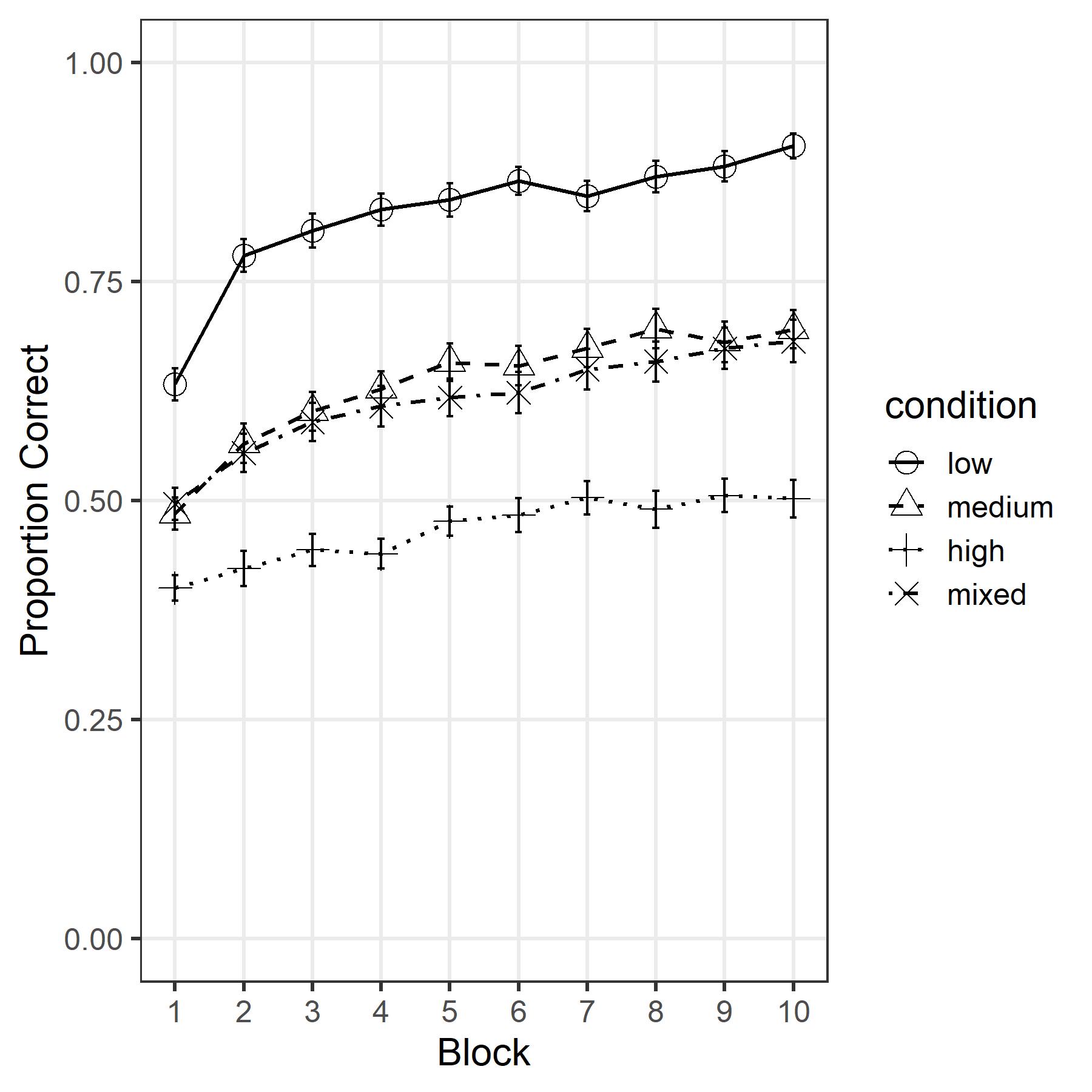


Figure 1

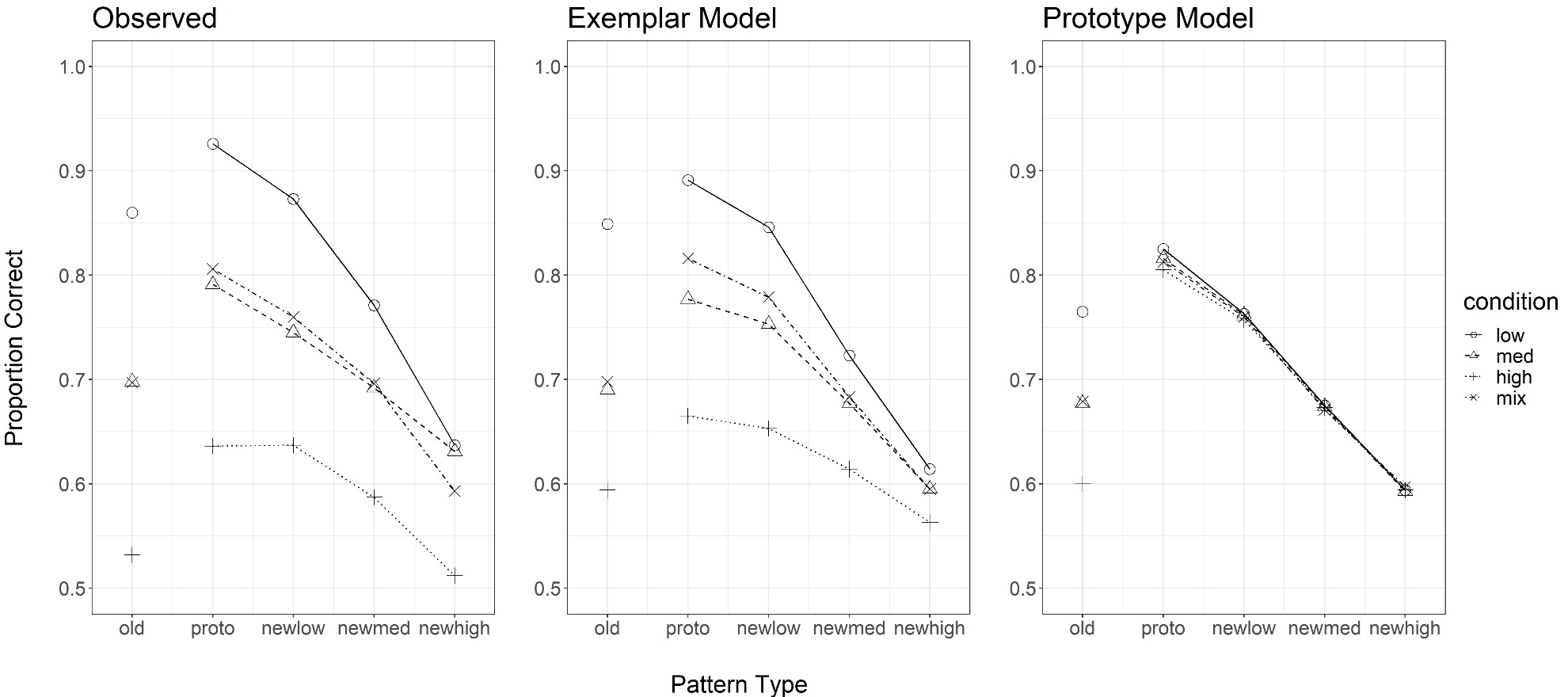


Figure 2

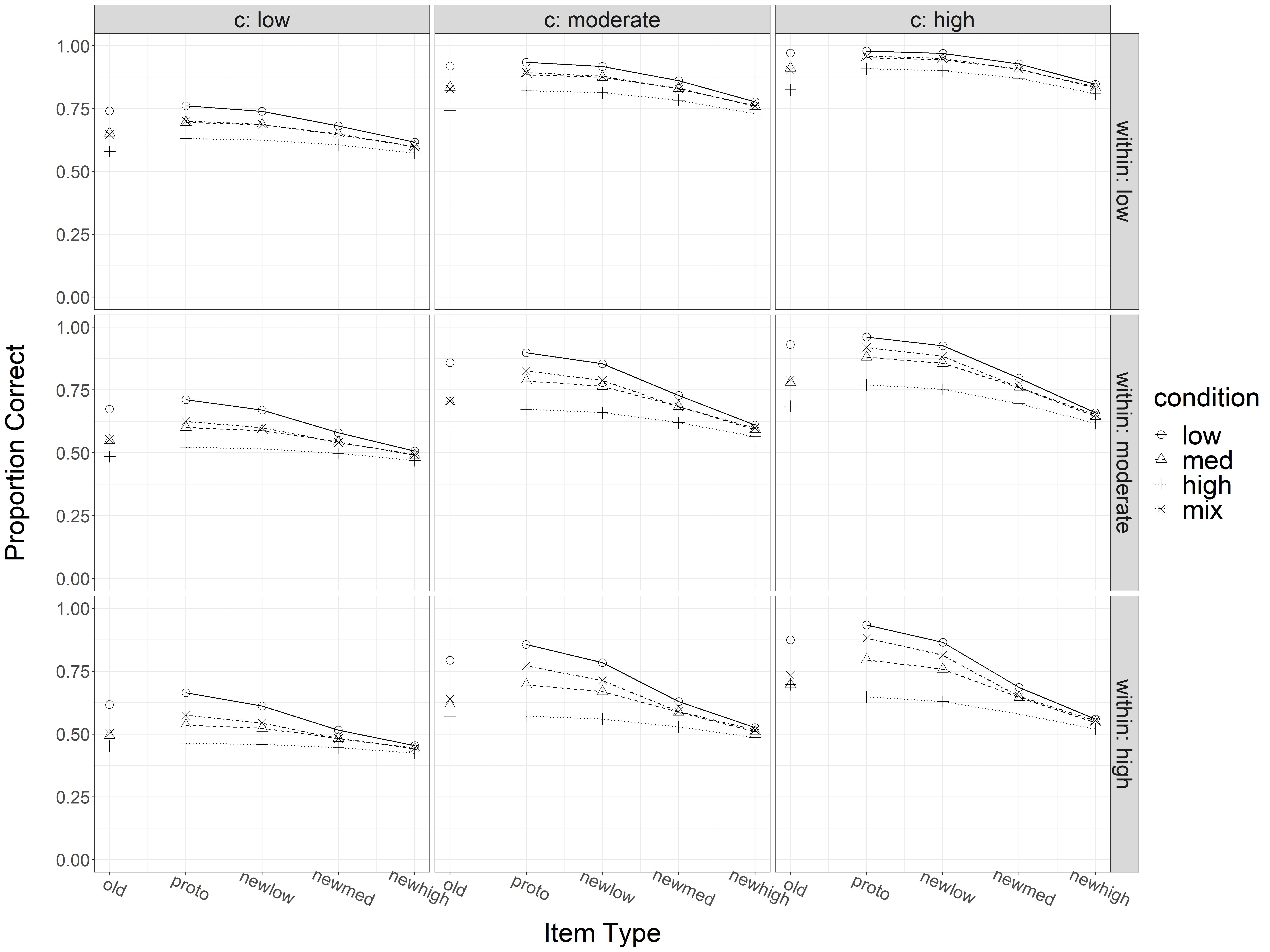


Figure 3

Mean proportion correct on the old-training instances and novel transfer patterns of the same level of distortion are reported for each condition in Table 1. Averaged across the conditions, mean proportion correct is only very slightly higher for the old compared to the new items. A two-way ANOVA using conditions and item type as factors revealed a main effect of training conditions, F(3, 300) = 45.46, MSE = .066, p < .001; but no main effect of item type, F(1, 300) = 1.04, MSE = .0072, p=0.31; and no interaction. F(3, 300) = 1.11, MSE = .0072, p=.35.