A classic debate in research on human category learning has involved the contrast between exemplar and prototype models. According to prototype models, people represent a category by abstracting its central tendency from constituent instances, and make categorization judgements based on similarity to the prototype (Homa, Cross, et al., 1973; Reed, 1972; Smith & Minda, 1998). In contrast, according to exemplar models, people represent a category by storing individual constituent exemplars in memory, and base categorization judgements on similarity to the exemplars (Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986).

Brief Review of the Debate in the Context of the Dot-Pattern Classification Paradigm

Many experimental results that initially appeared to strongly favor the prototype model were obtained from the dot-pattern paradigm developed by Posner and Keele (1968). Our focus in this article will involve new results derived from this paradigm. In a typical dot-pattern experiment, prototypes representing different categories are first generated by randomly placing nine dots in a grid; then patterns of various levels of distortion are constructed by displacing the dots of the prototypes according to a statistical-distortion rule. Higher levels of distortion produce dot patterns that are systematically less similar to the originating prototypes. The experiment typically consists of a learning phase, in which subjects are trained to classify a number of distorted patterns, followed by a transfer phase, in which subjects are tested on classifying a variety of patterns including the old training distortions, the prototypes, and various new distortions of the prototypes. Early studies found that in the transfer phase, the classification accuracy of the prototype not presented in the learning phase was higher than the various new distortions, and sometimes even exceeded that of the old distortions themselves. This prototype enhancement effect was cited as evidence for the abstraction of a prototype as a basis for representing the categories (e.g., Homa et al., 1973; Posner & Keele, 1968, 1970). Moreover, classification accuracy tended to decrease for new patterns with higher levels of distortion from the prototypes, producing a systematic “typicality gradient”. The typicality gradient is also consistent with the prototype model as patterns of higher levels of distortion are less similar to the prototype.

However, these classic results were also shown to be compatible with the predictions from exemplar models (e.g., Hintzman, 1986; Nosofsky, 1988; Shin & Nosofsky, 1992), which posit that category evidence is related to the summed similarity of test items to the training exemplars of the categories. The typicality-gradient effect arises because the summed similarity of novel test items to the old exemplars of a category tends to be higher for patterns closer to the center of category. The prototype-enhancement effect arises because the prototype is highly similar to virtually all the old training distortions; by contrast, any given old training distortion may be highly similar only to itself.

Since the classic studies of Posner and Keele (1968, 1970), many prototype theorists have argued that the prototype-abstraction process in the dot-pattern paradigm is more likely to operate when category size (i.e. the number of distinct training exemplars) is large and/or when the transfer phase is delayed (e.g., Homa, Sterling & Trepel, 1981). Specifically, Homa et al. (1981) found that classification accuracy for a new distortion increased as a function of its similarity to specific high-level old training distortions; critically, however, the contribution of this specific new-old similarity effect to classification performance was attenuated as category size increased. It was also found that the old distortions were classified more accurately than the prototypes immediately after the learning phase, but that the reverse pattern was observed after a one-week delay. These effects were once considered strong evidence for a prototype abstraction process. However, formal modeling of the specific new-old similarity × category size interaction and the differential forgetting of old distortions vs. prototypes revealed that both phenomena are qualitatively consistent with the predictions from pure exemplar models (e.g., Busemeyer, Dewey and Medin, 1984, Hintzman and Ludlam, 1980; Hintzman, 1986; Shin & Nosofsky, 1992). In general, the similarity of a new distortion to a specific old distortion makes a smaller relative contribution to overall summed similarity as category size increases: this specific new-old similarity tends to get “swamped” by similarity relations of the new distortion to the many other old training exemplars that compose the large-size categories. In addition, the differential forgetting of the old distortions versus the prototypes with delay of the transfer phase is also well-accounted for by pure exemplar models. The reason is that any given old distortion may be highly similar only to itself; thus, loss of fidelity in the memory representations for the old exemplars will be highly detrimental to old-item classification. By contrast, the prototype has high similarity to numerous old distortions that compose its category; this redundancy in the category representation protects the prototype from suffering major performance loss due to delayed testing.

More recently, Smith (2002) focused attention on the pattern of results observed in a particular version of the dot-pattern classification pattern introduced by Knowlton and Squire (1993). In this version, observers are exposed to 40 high distortions of a single prototype during an incidental training phase. Following the training, participants are tested on the prototype; new-low and new-high distortions of the prototype; and on random patterns not generated from the prototype. The participants are instructed to judge whether or not each test item is a member of the category that they experienced during the training phase. Smith (2002) argued that even though both exemplar and prototype models predict correctly the ordering of classification endorsements of the different pattern types, the steepness of the typicality gradient and the magnitude of the prototype enhancement effect observed in this paradigm falsify exemplar models.

However, Zaki and Nosofsky (2004, 2007) provided clear evidence that the magnitude of the typicality gradient observed in this paradigm did not arise from the abstraction of a prototype from the training instances; instead, it was an artifact of the structure of the test phase used in this paradigm and a result of continued learning that took place during the test phase (for closely related findings and criticisms of Smith’s 2002 interpretations, see Palmeri & Flanery, XXXX, XXXX). Specifically, in the Knowlton-Squire (2003) paradigm, participants are “flooded” with numerous presentations of the prototype and its low distortions during the test phase. As participants continue to experience these high-similarity patterns during the test phase, they continue to build upon the category representation that was developed during the incidental training phase. Zaki and Nosofsky (2004, 2007) provided strong evidence in favor of this view by manipulating the structure of the test phase itself: they found dramatic changes in the shape and steepness of the typicality gradient as a function of these test-phase manipulations. Moreover, in all cases, a simple exemplar model provided excellent quantitative accounts of the shape and steepness of the typicality gradient that was observed across the different test-phase manipulations.

The New Challenge: Classification and Recognition When Exemplars Never Repeat

As briefly reviewed above, the prototype-exemplar debate in the context of the dot-pattern paradigm has a long history. However, the debate was recently renewed in an interesting new study reported by Homa, Blair, McClure, Medema, and Stone (2019), who claim to have reported results that pose substantial problems for exemplar models. The central purpose of the work reported in the present article was to address these new challenges.

In Homa et al.’s (2019) experiments, participants learned to classify dot patterns into three categories, and then engaged in various transfer tests. As in past versions of the paradigm, each individual category was generated around a dot-pattern prototype. Low, medium, and high distortions of each prototype were generated using the Posner-Keele (1968) statistical-distortion algorithm. Foil patterns were also used, which were medium distortions of prototypes that were not trained during category learning.

The key manipulation across the experiments involved the structure of the learning phase.

Two different learning phases were employed in two conditions. In both conditions, the learning phase was organized into a sequence of 15-trial blocks involving the presentation of medium-level distortions of the prototypes. In the *repeating* condition (REP), the same 15 medium-level distortions (5 per category) were presented in each trial of every 15-trial learning block. By contrast, in the *non-repeating* condition (NREP), no individual training instance was every repeated. Instead, 15 different medium-level distortions (5 per category) were presented in each 15-trial learning block. So, for example, in Experiment 1, in which there were 20 learning blocks, participants experienced 15 distinct training instances in the REP condition (each one repeated 20 times), but experienced 300 unique training instances in the NREP condition (each one presented only one time).

Across experiments, participants then engaged in various transfer tests. In Experiment 1, participants were required to classify novel patterns (prototypes, low-, medium-, and high-level distortions) into the trained categories. In Experiments 2 and 3, participants instead engaged in old-new recognition tests, in which they judged whether test patterns had or had not been presented during the training phase. In Experiment 2, the test patterns were old distortions, new-medium distortions, and foils; in Experiment 3, the test patterns were old distortions, new-medium distortions, and prototypes.

Homa et al. (2019) focused on three main patterns of results in their study that they claimed severely challenged exemplar-only models, but that were well accommodated by a model that assumed exemplar-based classification in the REP condition, but prototype-based classification in the NREP condition. The first result was that, across the three experiments, there was no difference in speed of learning across the REP and NREP conditions. The second result was that, in the recognition-transfer tests, participants were unable to discriminate between the old- versus the new-medium distortions in the NREP condition, but showed well-above-chance discrimination of these pattern types in the REP condition. The third result was that participants classified patterns with high accuracy in the transfer phase of the NREP condition, despite the fact that no single training instance was ever repeated during the learning phase.

The second result can be demonstrated by modeling recognition transfer phase. In the experiments 2 and 3, the probability with which a transfer pattern i is recognized as old can be expressed by the equation (1):

(1)

where is the background noise present at the beginning at the learning phase, and is the response-scaling factor (subjects make classification decisions exactly based on the probability matching when , but towards the category with the largest probability more deterministically when ). The criterion parameter takes the value of in the REP condition and of in the NREP condition to reflect different decision criteria. The difference between the old response probabilities of old and new distortions is essentially determined by the difference between the summed similarities of the two item types to all three categories.

To mirror the modeling approach used by Homa et al., summed similarities of an old and new distortion to stored patterns in each category can be approximated in terms of “within-category similarity” between any two medium-level distortions within the same category, and “between-category similarity” between any two medium-level distortions belonging to different categories, as in equations (6) and (7). It is assumed that pattern i belongs to category A.

(2)  
 (3-1)  
 (3-2)

where the total number of learning blocks is equal to 20 in experiments 2 and 3. Equation 3-1 represents the summed similarity from an old medium-level distortion o to all stored exemplars in category A. Its self-similarity is maximal at 1 and its similarity to 99 other old distortions is defined to be . Equation 3-2 represents the summed similarity from a new medium-level distortion n to all stored exemplars in the category A. Its similarity to all 100 old distortions is defined to be . Intuitively, tend to be much smaller on average than , so the summed similarities from either the old or new distortion to category B and C (equation 2) is negligible compared to the summed similarity to category A (equations 3) when computing the grand total. is greater than by , derived from equation (3-1) minus equation (3-2). However, the difference attributed to the single self-similarity term is swamped by all other terms shared by both item types. In other words, and are almost the same, thus the old response probabilities of the two item types are very close, implying that subjects failed to discriminate between old and new patterns.

By contrast, the old response probability is much higher for old than new distortions in the REP condition. , and stay the same as in the NREP condition. Equation 3-3 represents the summed similarity from an old medium-level distortion to all learned patterns in category A:

(3-3)

Its self-similarity of 1 is multiplied by 20 as it was repeated once in each learning block and its similarity to 80 other old distortions is defined to be . is greater than to category A by . The difference attributed to the self-similarity terms is now magnified to produce a substantial difference between and thus the old response probabilities between the two item types, implying that subjects can easily discriminate between old and new distortions in the REP condition.

Another reason causing the old response probability between the two item types to be smaller in the NREP condition than in the REP condition is different values of in the two conditions. As illustrated in figure 1, with more old distortions densely packed in the NREP condition than the REP condition, a new distortion is more likely to be surrounded by old distortions highly similar to it, so the , or the expected similarity between any two distortions, tend to be higher in the NREP condition than the REP condition. The lower in the NREP condition further diminishes the differences between and . Notably, Homa et al. failed to capture this effect even in their exemplar model, which used derived from the same medium-to-medium distortion distances in both conditions.

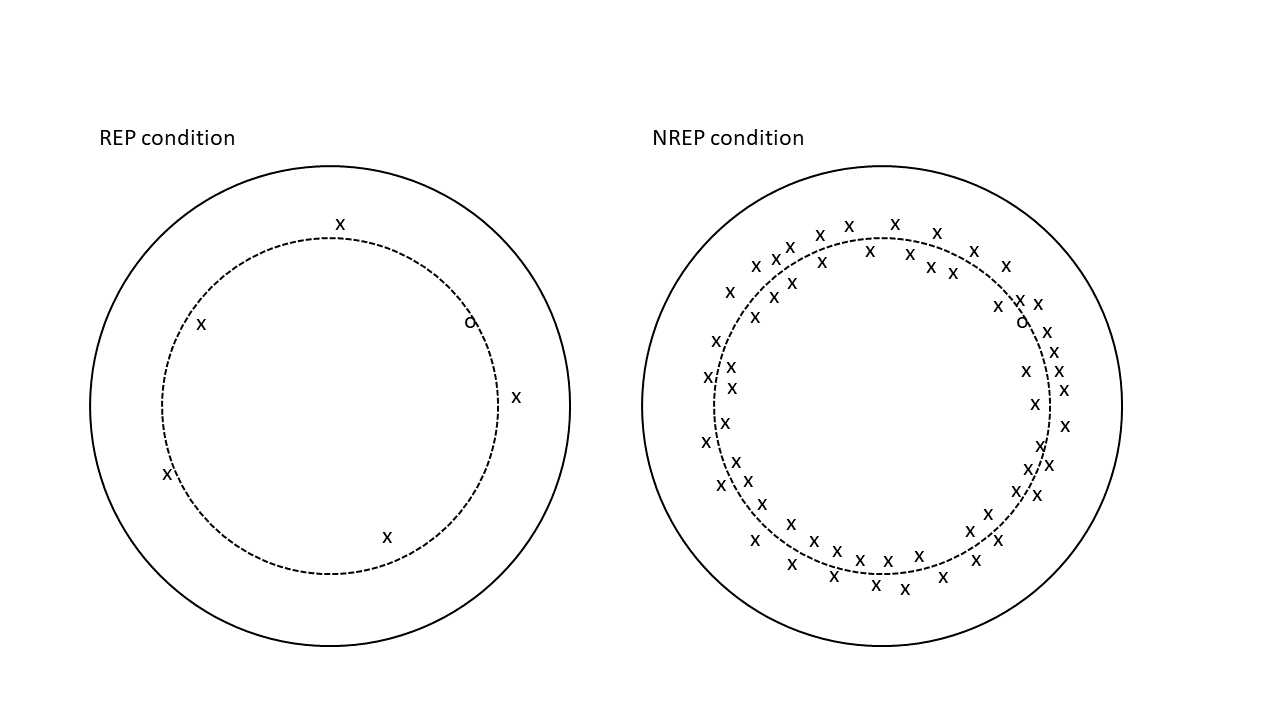


Figure 1 schematic illustration of similarity structure of stored exemplars in the REP and NREP conditions. A solid circle represents a category and a dashed circle represents the average medium-level distortion distances from its prototype. The symbol o denotes old medium-level distortions and the symbol x denotes new medium-level distortions. Smaller spatial distance between any two symbols indicates higher similarity between the two.

The third result can be demonstrated by modeling classification transfer phase. In the experiment 1, the probability with which a transfer pattern i is correctly classified as belonging to category A can be expressed by the equation (4):

In both REP and NREP conditions, the summed similarities from any novel pattern to each category can be estimated by equations (2) and (3-2). The classification accuracy is primarily determined by the proportion of the summed similarity to all three categories constituted by category A. As shown before, regardless of the item types, and are much smaller than , so the classification accuracy is generally very high. Although the summed similarity to category A can be expressed by the same equation (3-2) in both conditions, is higher in the NREP condition than the REP condition, so thus the classification accuracy is higher in the NREP condition.

To capture the subtle effect of category density on , we represent dot patterns as points in a six-dimensional feature space and compute pairwise similarities from Euclidean distances between the points. According to an MDS analysis of the similarity ratings of dot patterns (Shin & Nosofsky, 1992), six psychological dimensions can well account for variability in the perceived similarity among dot patterns. The six-coordinates of the prototypes, various distortions and foils are randomly generated in a way analogous to the statistical-distortion procedure used to generate the dot patterns. Two freely varying parameters are adopted in generating dot patterns. A “within” parameter is multiplied by distortion distances to generate various distortions from the prototype, and a “between” parameter defines the maximum distance between any two prototypes on each dimension. The psychological distance between two dot patterns i and j can be expressed by equation (5).

(5)

where represents the coordinate of pattern i on dimension m in the six-dimensional psychological space. The similarity measure is an exponential decay function of the distance, as in equation (6)

(6)

where the sensitivity parameter c reflects subjects’ overall discriminability in the psychological space.

The transfer phases of the three experiments are simulated 10000 times to obtain a reliable prediction of the mean proportions across trials and subjects. In each simulation, a single point is generated to represent each item type of the transfer patterns\*, and its similarities to old distortions are derived from equations (5) and (6). For each category, ( = 15 in the experiment 1, = 20 in the experiments 2 and 3) different training patterns were generated in the REP condition, and 5 different training patterns, each to be repeatedly counted in each training block, were generated for each category in the NREP condition. The probabilities of correct response and old recognition are computed from individual similarities by equations (4) and (1) respectively. After 10000 iterations, the predicted probabilities in each experiment by learning conditions and item types were averaged across iterations.

\* For simplicity, the transfer patterns always belong to the first category as defined in each iteration. The old distortions in the two learning conditions were separately defined as the first exemplar in the first category of the respective condition.

As can be seen from the figure 2, the exemplar model perfectly predicts the ordering of classification accuracies and old response probabilities for the different item types in each condition. The quantitative fit was also reasonably well (SSE = .009). Specifically, there was little difference between the old response probabilities of old and new distortions in the NREP condition. The classification accuracies in the experiment 1 also tend to quite high, with the accuracy of medium-level distortion higher in the NREP condition than the REP condition.

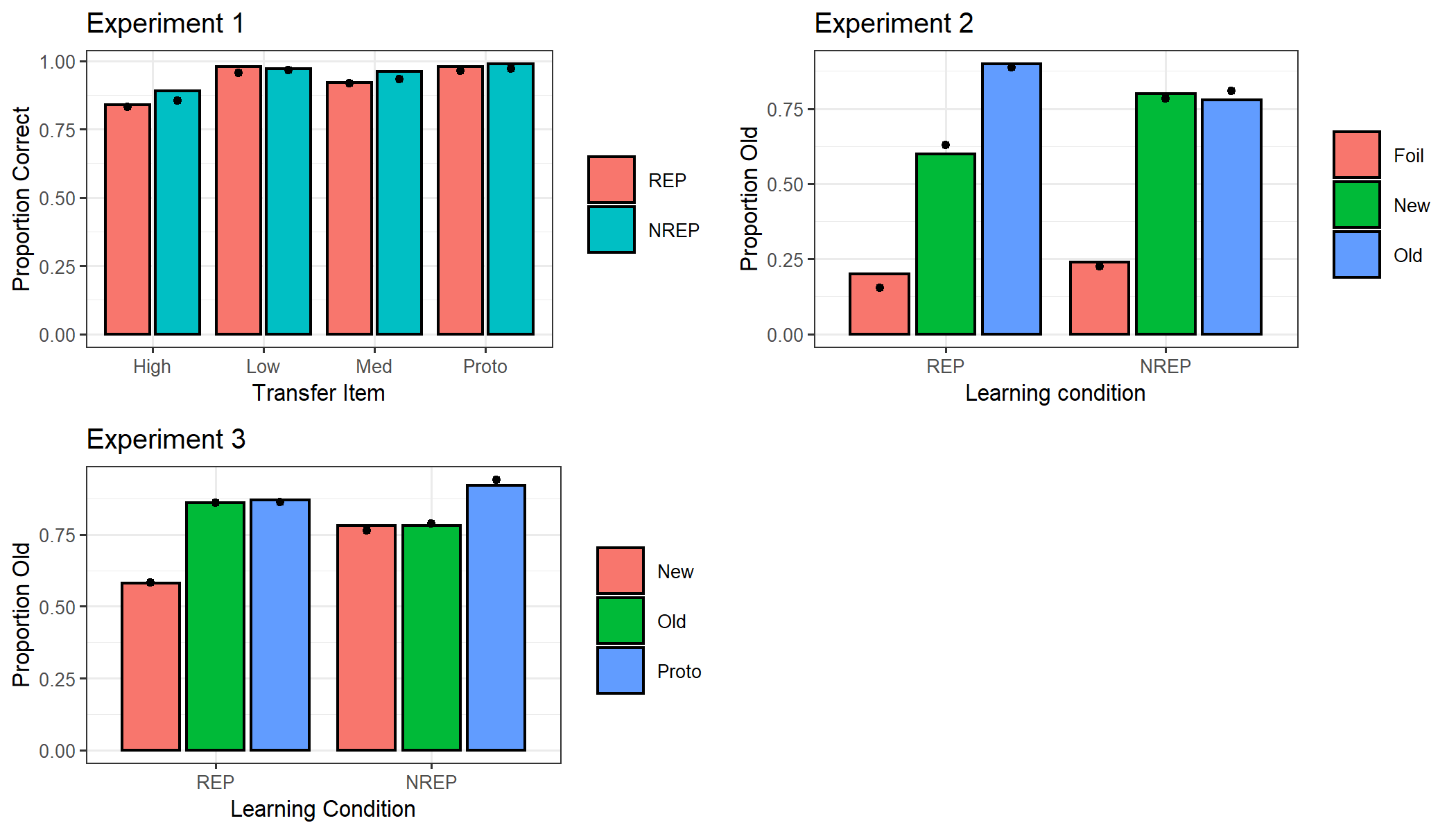


Figure 2 observed and predicted probabilities of correct classification in experiment 1 and of old responses in experiments 2 and 3 for each item types, shown for the REP and NREP conditions separately. The bars represent observed data and the dots on each bar represent predicted data.

Lastly, we acknowledge that the first result severely challenges a pure exemplar model. Indeed, Homa et al. showed that the speed of learning was predicted by a pure exemplar model to be faster in the REP condition than the NREP condition. The learning rate was demonstrated by the rate at which classification accuracy increases across learning blocks. At the beginning of the learning phase, the classification accuracies between the two conditions were very similar as they was dominated by the same background noise ; as the learning progresses across blocks, the classification accuracies became increasingly different in the two conditions as caused by the differential magnitudes of relative to (equation 8). To summarize their argument, in each learning block, the main difference between the classification accuracy for the two learning conditions lies in the difference between s in the two conditions. s for the REP and NREP conditions can be estimated by equations (11-1) and (11-2) respectively.

(11-1)

(11-2)

where B indexes the learning block and begins at B = 2, and o represents an old distortion used in the block B of the learning phase.

As the learning progresses, increased at the rate of in the REP condition as the same pattern o was presented one more time in the previous block, while increased at the rate of in the NREP condition as the pattern o were different from all five patterns presented in the previous block. Therefore, the learning rate was greater in the REP condition than the NREP condition by per block.

We intend to replicate Homa et al.’s study to confirm whether or not the speed of learning is different in the REP and NREP conditions, and whether or not there is any above-chance discrimination of old vs. new medium distortions in the NREP condition.