Technical Report:

Formal Modeling of the Effects of Training Variability

on Classification Learning and Generalization

Mingjia Hu, Robert M. Nosofsky

Method

Subjects

The subjects were 214 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 used in this experiment were dot patterns generated using Posner-Keele (1968) statistical-distortion algorithm. 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-level, medium-level 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.

Procedure

A standard learning-transfer paradigm was used in this experiment. In the learning phase, subjects were trained to classify a set of training patterns into three categories. 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, the corrective feedback appeared for 2s below the presented pattern. A different set of training patterns were presented in each of the 10 training blocks. The learning phase was followed 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 subject, prototypes for three different categories were randomly generated. Subjects were randomly assigned to one of the four training conditions that differ in terms of the variability of the training patterns. 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 by various levels 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 mixture (equal number) of the three distortion levels, respectively. The test patterns consisted of 27 old distortions that were presented in the training phase (9 per category), 3 prototypes (1 per category), 9 new low-level distortions (3 per category), 18 new medium-level distortions (6 per category), 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 trials.

Results

Learning. Figure 1 shows the average proportion of correct classification responses over the training blocks for each of the four training conditions. Across the training conditions, the classification accuracy gradually improves over the course of training. Particularly, the low-distortion training condition exhibits the fastest rate of improvement, and the medium- and mixed-distortion conditions showed intermediate rate of learning and classification accuracy at the end of training phase, while the high distortion condition has the lowest terminal 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 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. The mean proportion of correct responses for the final training blocks is higher in the low condition (M = 0.905) than in the medium condition (M = 0.695), t(132.3) = 8.05\*\*, p < .001, and higher in the medium condition than in the high condition (M = 0.502), t(151.0) = 6.33, p < .001.

Transfer. Figure 2 shows the mean proportion of correct responses for various types of test patterns. The general trend is that, across training conditions, the classification accuracy is the highest for the prototypes, and gradually decreases in the order of low-, medium- and high-level distortion test patterns. Moreover, 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 training condition than in the low-distortion condition.

To confirm these observations, we 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 learning 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. 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 = .036, and even in 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 = .036.

In our subsequent modeling analysis, we decided to conduct separate analyses on all subjects and those subjects with adequate overall accuracy during the transfer phase. As can be seen in figure 1, there are discernible variations in both the overall classification accuracy and the subject-level distribution across training conditions. For each condition, we computed individual test accuracies by averaging the proportions of correct responses over all pattern types, and decided to retain for our subsequent analyses the proportion of subjects with the highest 90% of individual test accuracies in each condition (rather than setting separate training criterion for each condition). As a result, there remained 70 subjects in the low-distortion condition, 71 in the medium-distortion condition, 68 in the high-distortion condition and 67 in the mixed-distortion condition. Even after removing the lower performing subjects from all conditions, mean proportion correct for novel high-distortion patterns in the high-distortion condition (M = 0.538) remained significantly lower than in the medium-distortion condition (M=0.661), t(135.0)=4.345, p< .001, the mixed-distortion condition (M=0.626), t(132.7) = 3.045, p = .032, and the low-distortion condition (M=0.657),t(119.4) = 4.660, p< .001. For the novel medium-distortion patterns, the mean proportion correct in the medium-distortion condition (M = 0.732) also remained significantly lower than the low-distortion condition (M = .809), t(125.4) = 3.210, p = .008. As shown in figure X, the overall patterns of transfer results stayed the same as in the analysis in which all subjects were included.

\* The Greenhouse-Geisser correction was applied for violation of the sphericity assumption.

\*\* The Welch t-test was conducted which assumes unequal population variances.

Model-based accounts of the Classification Transfer Data

Since the psychological dimensions composing the dot patterns are unknown, it is currently not feasible to develop a rigorous quantitative model of classification responses based on individual patterns, until further research is conducted to uncover the psychological representations of the dot patterns. The current modeling analysis aims to show that the exemplar model can provide a viable account of the qualitative patterns observed in the overall subject performance during the transfer phase. Following Hintzman’s (1986) influential style of modeling, we decided to simulate the psychological structure of dot-pattern stimuli and categories in an analogous manner to Posner-Keele statistical-distortion procedure.

In our simulations, we represent the dot patterns as points in a six-dimensional psychological space (psychological space is defined to have six dimensions as MDS studies by Nosofsky & Zaki [1992] reported that six-dimensional solutions can adequately account for the similarity relations among dot patterns). 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 freely estimated parameter *between* controls the degree to which different category prototypes are similar to each other. In general, larger values of *between* will result in larger psychological distances among category prototypes, hence less similarity between categories.

For each category, the statistical distortions were then generated by sampling *z* scores from a standard normal distribution, and adding scaled values of *z* to the dimension values of the corresponding category prototype. Different scaling factors were used to represent different levels of distortions, as in equation 1.

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

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

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

In the equation above, *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*. *within* is a free parameter that primarily determines the degree of within-category dissimilarity, or how dissimilar the patterns are from one another in the same category. The parameters *low*, *medium* and *high* specify the magnitude of low, medium and high distortion levels. Without loss of generality, we set the between-category scaling parameter *between* to be fixed at 2, while allowing the parameters *within* and *c* to be freely estimated.

We fitted two versions of the exemplar models: the baseline version where the distortion-level parameters were constrained to take the same values as the average dot-distance movements according to the Posner-Keele statistical-distortion algorithm (*low* = 1.20, *medium* = 2.80, and *high*= 4.60), and the free-distance version where the distortion-level parameters were freely estimated. In the free-distance version, the parameter *within* is omitted since it cannot be estimated separately from the distortion-level parameters. Of course, to create a new distortion, a new random z score is sampled along each individual dimension.

Formal Models of Categorization

Once the patterns are created for each individual simulation, the standard equations of exemplar and prototype models (xxxx) are used to generate the classification predictions in the transfer phase.

According to the exemplar model (i.e. GCM), the probability that a pattern *i* is classified into category A is found by summing its similarity to all the training examples *a* that belong to category A, 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 grows larger in magnitude, the observer responds more deterministically with the category that yields the largest summed similarity.

The similarity between test pattern *i* and training example *j* (*sij*) was defined as an exponential-decay function of the distance between the two patterns in the psychological space:

(3)

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.

The standard Euclidean distance formula is used to compute the distance between test pattern *i* and training example *j*,

(4)

where and denotes the values of the patterns i and j on dimension *m*, respectively. In sum, there are three free parameters (*within*, *c*, *γ*) in the baseline exemplar model and five free parameters in the free-distance exemplar model (*low*, *medium*, *high*, *c*, *γ*).

According to the prototype model, the probability that pattern i is classified into category A is given by

(5)

where is the similarity between the test pattern i to the prototype of category A. Note that the response-scaling parameter γ cannot be estimated separately from the sensitivity parameter c (as defined in eq. 3) in the prototype model, so γ is fixed at 1 for the prototype model.

Due to the random nature of the dot-distortion algorithm, the prototype used to generate the training examples in the same category may not be precisely centrally located among them, especially for med- and high-distortion training conditions. Therefore, the prototype representations are computed by averaging across the dimension values for each of the training exemplars of the corresponding category. Otherwise, the similarity and psychological distances between the test patterns and the prototypes are computed in an analogous way as in eqs. 3 and 4.

In sum, there are two free parameters (*within*, *c*) in the baseline prototype model.

For each of the exemplar and prototype models, the best-fitting parameters are estimated by minimizing the sum-of-squared deviations between the predicted and observed probabilities that the test patterns are correctly classified for each pattern type across all conditions. The observed probabilities are computed by averaging across subjects the proportion of correct responses. The predicted probabilities are calculated by averaging across the results of 10000 simulations in predicting the probabilities of making correct classifications. We used the Hook and Jeeves (1961) algorithm for parameter search.

Model-fitting Results

The right panel of Figure 3 shows the classification accuracies across test pattern types and conditions predicted by the best-fitting baseline exemplar model. Consistent with the observed data (Fig. 3, left panel), the model predicts that the novel high distortions are classified with the lowest accuracy in the high training condition compared to the other conditions, and that the novel medium distortions are classified with lower accuracy in the medium training condition than in the low training condition. It is also worth noting that the classification accuracies of novel high distortions are predicted to be very close in the low and the medium training conditions, indicating that increasing the distortion level of training patterns will have little, if not negative, effect on the generalization performance on highly distorted novel patterns. The best-fitting parameters as well as the sum-of-squared deviations for the two versions of exemplar models are reported in table 1. Despite the slight improvement in the quantitative model fit of the free-distance model, the more complex model predicts the same qualitative patterns as the baseline model (Fig. 4).

The right panel of Figure 5 illustrates the predictions yielded by the best-fitting baseline prototype model. Apparently, the model predicts virtually no difference in the classification performance for the novel test patterns across all four training conditions. In other words, according the prototype model, the level of distortions of the training patterns has no effect on the generalization performance. The best-fitting parameters and the sum-of-squared deviations for the baseline exemplar model are reported in table 1.

Multidimensional Scaling

Although the node activations extracted from neural network may encode the psychological representations of the dot patterns to some extent, the representations in a neural network are distributed across many nodes. Therefore, the activation values of individual nodes per se contain little information about the underlying psychological dimensions, and further analysis is needed to reduce the activation values to psychological feature vectors with lower dimensionality.

For this purpose, we applied a metric-scaling model to the between-pattern distances derived from the deep-learning activation values. First, we computed the standard Euclidean distances between the vectors of activation values associated with every pair of test pattern (as in Eq. 1). Second, we used mdscale function from MATLAB to conduct the metric scaling analysis[[1]](#footnote-1). The MDS program searches for the locations of the points each representing a test pattern in a multidimensional space so as to approximate a linear relation between the inter-point distance in the MDS space and the pairwise distances computed from the activation values. Thus, patterns that seem more similar tend to be located closer together in the space. The departure from a perfect linear relation is known as stress (Kruskal & Wish, 1978). As one increases the number of dimensions, one can reduce the stress, but at the expense of requiring a greater number of free coordinate parameters to achieve this fit.

The number of dimensions was varied from 1 through 10. Figure [5](https://link.springer.com/article/10.3758/s13428-017-0884-8#Fig4) shows a plot of stress against the number of dimensions assumed in the analysis. As can be seen, there is a drastic decrease in stress with increases in dimensionality from 1 to 2, and very little to no decreases in stress thereafter. Based on the value of stress alone, the minimum number of dimensions need to provide a good fit is 2. However, we decided to choose the three-dimensional MDS solution in order to derive more interpretable dimensions, which will be explained shortly.

After obtaining the MDS solution, the next step is to test if the dimensions of the MDS configuration have natural interpretations that correspond to important characteristics of the dot patterns. It is important to note that the MDS modeling analyses, with the inter-point distance conforming to a Euclidean metric, has the property of rotation-invariance, namely, any rigid rotation of the scaling solution will yield the same inter-point distances in the space. Therefore, the orientation of the MDS solution is arbitrary, so additional analyses are needed to address the issue of the interpretability of the derived dimensions.

To address the interpretability problem, we first need to establish a number of quantitative measures characterizing some of the most salient emergent properties of the dot patterns that subjects can rely on for classification decisions. Based on preliminary inspections of two-dimensional projections on the three-dimensional MDS solution, we hypothesized three objective dimensions of dot pattern features whose values can be computed from the x-y coordinates of the composing dots. The first dimension is the extent to which the overall size of the pattern is wider than it is taller (fig. 6, left panel, x axis). Intuitively, it can be formally measured as the width-to-height ratio:

(7)

where xi and yidenote the x and y coordinate of the ith out of the 9 dots, respectively

The second dimension represents the extent to which the pattern seems split in half horizontally (fig. 6, left panel, y axis). To measure it objectively, we first divided the pattern into two clusters by bisecting the range of the x-coordinates of all 9 dots and grouping the subset of dots in the same section into one cluster. Then, the degree of splitness is measured as the ratio of the average inter-point distance for dots within the same clusters and that for any pair of dots between the clusters. Formally,

(8a)

(8b)

(8c)

where xli denotes the ith dot in the left cluster and xrj denotes the jth dot in the right cluster. pair(i,j), pair(i1,i2), pair(j1,j2) represents the number of all possible pairs of dots between the two clusters, and within the left and right clusters, respectively.

The third dimension measures the extent to which there are two distinctive dots lying in the center of a pattern (fig. 6, right panel, y axis). The centrality of individual dots are measured as the dot distances to the centroid, the ideal central location found by averaging over the coordinates of all nine dots. With the notion of centrality quantified, the dimension value is then defined as the ratio of the average distance of all nine dots to the centroid and that of the closest two dots to the centroid. Formally,

(9)

where denotes the mean of the x coordinates of all 9 dots, and denotes the mean of the y coordinates of all 9 dots.

Having defined the physical dimensions, we then ran a FORTRAN program which rotates, translates, and scales the derived MDS solution in an attempt to bring it into correspondence with the normative solutions consisting of individual-pattern measurements along the three explicitly defined feature dimensions. The “target” MDS solution (produced by rotation, translation and scaling) was defined to be the one that minimized the sum of squared deviations (SSD) between the x im values and the corresponding z im values across all test patterns and the three hypothesized dimensions, formalized as

(10)

Where xim and zim denote the coordinate value of the pattern i on dimension m, in the MDS solution and the normative solution respectively.

The correlations between the three dimensions of the rotated MDS solution and those of the normative solution were shown in Table 3. As can be seen, the second dimension of the MDS solution highly correlates with the horizontal splitness measures, but the first and third dimensions are less strongly correlated with the measures of width-to-height and the dual centrality. Notably, these correlations were computed with respect to a single rigid rotation of all three axes of the original MDS solution. However, even though rotating the solution separately for individual dimensions slightly increases the correlations, the general trend still stays the same for individual-dimension fits.

Exemplar model fitting

Using the rotated MDS solution as the feature representations, we first fitted two versions of the GCM to the response data in the test phase. As can be seen from the rotated MDS solution (fig. 7), some dimensions are more diagnostic of the category membership of the test patterns than others. Therefore, it stands to reason that assigning freely-estimated attention weights to different dimensions in the MDS-derived feature representations could largely improve the model fit to the observed responses. As such, we fitted a baseline version of GCM assuming equal attention weights to the three dimensions and another version with freely estimated dimensional weights. The other aspects of GCM models were formulated in the same way as for the GCM model with raw activation representations (as specified in equations 2-4). The scatter plots in figure 8 compare the observed category response probabilities of individual test patterns with the corresponding predictions yielded by the two GCM models. Again, the solid dots indicate correct classification responses.

Comparison of the summary fits (table 1, exemplar models) reveals that applying MDS techniques and dimension weighting to CNN activation patterns substantially improves the fit to the classification data. Although the unweighted MDS exemplar model performs slightly more poorly than does the average-similarity exemplar model, the weighted version yield noticeably better summary fits to the data than does the average-similarity model.

Prototype model fitting

We also fitted four versions of the prototype models using the same MDS-derived feature representations. As can be seen from the rotated MDS solution (fig. 7, left panel), the category prototypes tend to lie at extreme, peripheral regions of the category distributions, which is in contrast to the general intuition that the category prototypes should be centrally located among the training exemplars of the same category. In other words, the prototypes as represented by the MDS coordinates are more like exaggerated, ideal characterization of the training exemplars in the same category than an average representation thereof. To explore the effects of the alternative notions of prototypes on the model fitting performance, we defined the feature-space representation of the prototypes in two methods to capture both notions. We called the first method ideal-point version where the prototype representations are defined to be their extreme, ideal-point MDS coordinates, and the second were labeled central-tendency version where the prototype representations are computed by averaging across the MDS coordinates for each of the training exemplars of the corresponding category. For each of the two methods of the feature-space representations, we fitted two versions of the prototype models: one with equal dimension weights, and another with freely estimated dimension weights. The scatter plots in figure 9 show the observed and predicted probabilities of category responses for individual test patterns. In addition, the summary-fit statistics from the four prototype models are reported in the bottom rows of table 1, and the best-fitting parameters for the two models are reported in table 4.

The summary fits of prototype models reported in table 1 show similar patterns as the ones of exemplar model. First, both central-tendency prototype models are chosen over the respective ideal-point prototype models using the AIC and BIC statistics, indicating that the central tendency is the more valid prototype representation in terms of prototype model fits. As with the exemplar model fits, the weighted MDS central-prototype model outperforms the unweighted counterpart in fitting the data. Moreover, the weighted MDS prototype model even yields better summary fits than do both the average-similarity prototype and exemplar models. It is also noteworthy that there is little difference in the summary fits of the weighted MDS exemplar model and the weighted MDS central-prototype model. The reason may be that in the low-distortion training condition, the similarity of a test pattern to the prototype of each category is roughly the same as the similarity of a test pattern to the low-distortion training examples.

Table 1. Summary Fits of Models to the Classification Test Data

**Exemplar Models**

Model -lnL AIC BIC P

Average Similarity 1159.7 2331.4 2370.8 6

Dot Coordinates 1227.5 2459.0 2471.1 2

CNN Activation 1221.7 2447.4 2460.5 2

CNN\_MDS\_unweighted 1178.8 2361.6 2374.7 2

CNN\_MDS\_weighted 1119.3 2246.6 2272.8 4

**Prototype Models**

Model -lnL AIC BIC P

Average Similarity 1159.9 2329.8 2354.0 4

Dot Coordinates 1195.3 2392.6 2399.2 1

CNN Activation 1272.0 2546.0 2552.6 1

CNN\_MDS\_unweighted\_extreme 1272.5 2547.0 2553.6 1

CNN\_MDS\_weighted\_extreme 1186.4 2378.8 2398.5 3

CNN\_MDS\_unweighted\_central 1179.3 2360.6 2367.2 1

CNN\_MDS\_weighted\_central 1123.5 2253.0 2272.7 3

Table 2. Best-Fitting Free Parameters of Models with Alternative Methods of Stimulus Representation.

**Exemplar models**

Parameter Average Similarity Dot Coordinates CNN Activation

*Sp*  1.000 -- --

*Sl* 0.523 -- --

*Sm* 0.370 -- --

*Sh* 0.277 -- --

*Sc* 0.204 -- --

*c* -- 0.170 0.212

*γ* 2.906 1.000 1.000

**Prototype models**

Parameter Average Similarity Dot Coordinates CNN Activation

*Sp*  1.000\* -- --

*Sl* 0.208 -- --

*Sm* 0.062 -- --

*Sh* 0.026 -- --

*Sc* 0.011 -- --

*c* -- 0.166 0.172

*γ* 1.000 1.000\*\*  1.000\*\*

\* Sp is set to 1 for self-match similarity

\*\* γ is fixed at 1 as it cannot be estimated separately from c

Table 3. correlation coefficients between the corresponding dimensions of the rotated MDS solution and the normative solution.

|  |  |
| --- | --- |
| Dimension | Correlation |
| 1.height-to-width ratio | 0.373 |
| 2.horizontal splitness | 0.802 |
| 3.dual centrality | 0.600 |

Table 4. Best-Fitting Free Parameters of Models based on the Rotated CNN-MDS Solution

**Exemplar Models**

Parameter GCM\_unweighted GCM\_weighted

*w1*  -- 0.084

*w2* -- 0.669

*\*w3* -- 0.247

*c* 0.221 0.421

*γ* 1.000 1.000

**Prototype Models**

Parameter PM\_central PM\_extreme PM\_central PM\_extreme

unweighted unweighted weighted weighted

*w1*  -- -- 0.091 0.000

*w2* -- -- 0.652 0.636

*\*w3* -- -- 0.257 0.364

*c* 0.213 0.173 0.402 0.341

*γ* 1.000 1.000 1.000 1.000

*\*w3 = 1 - w1 - w2*

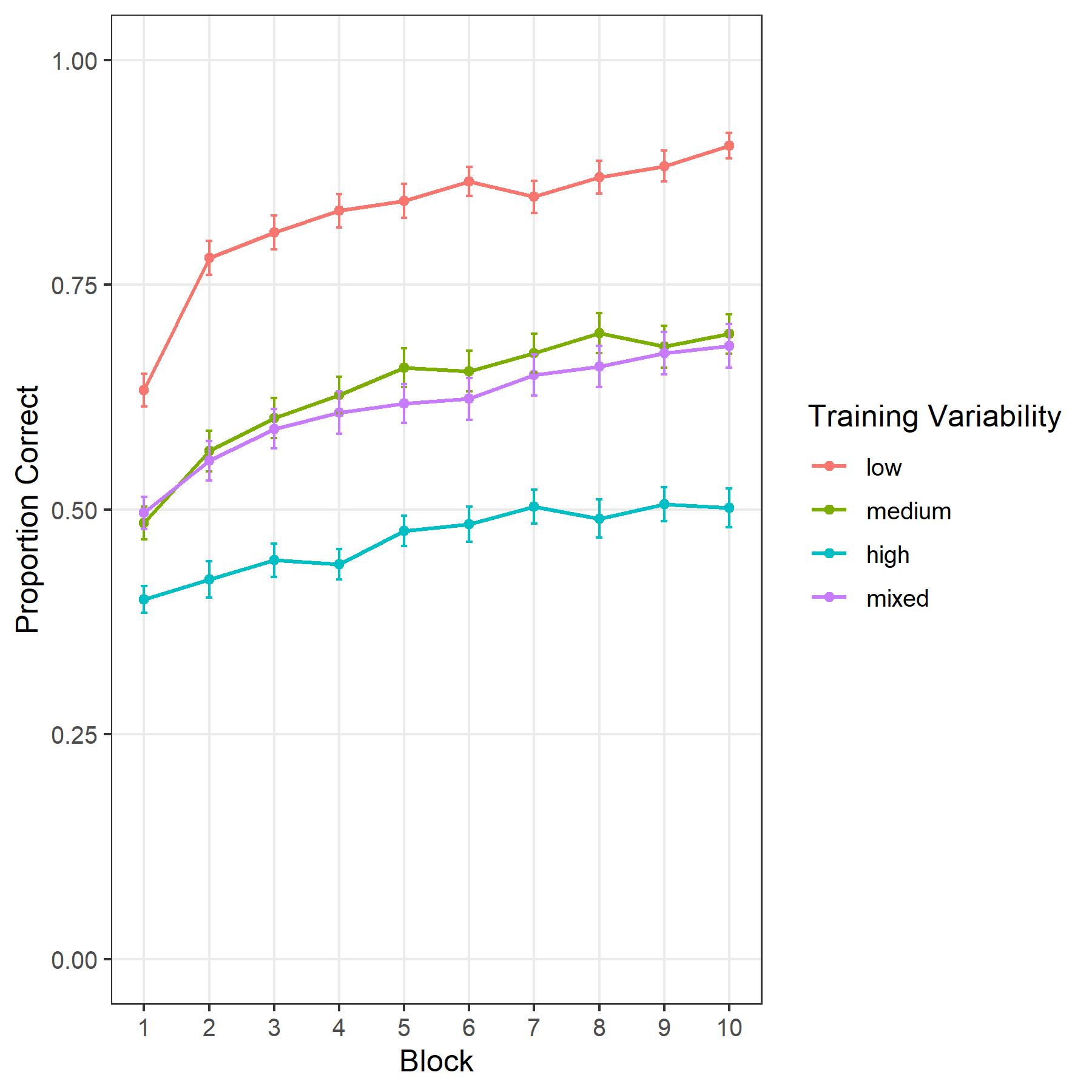
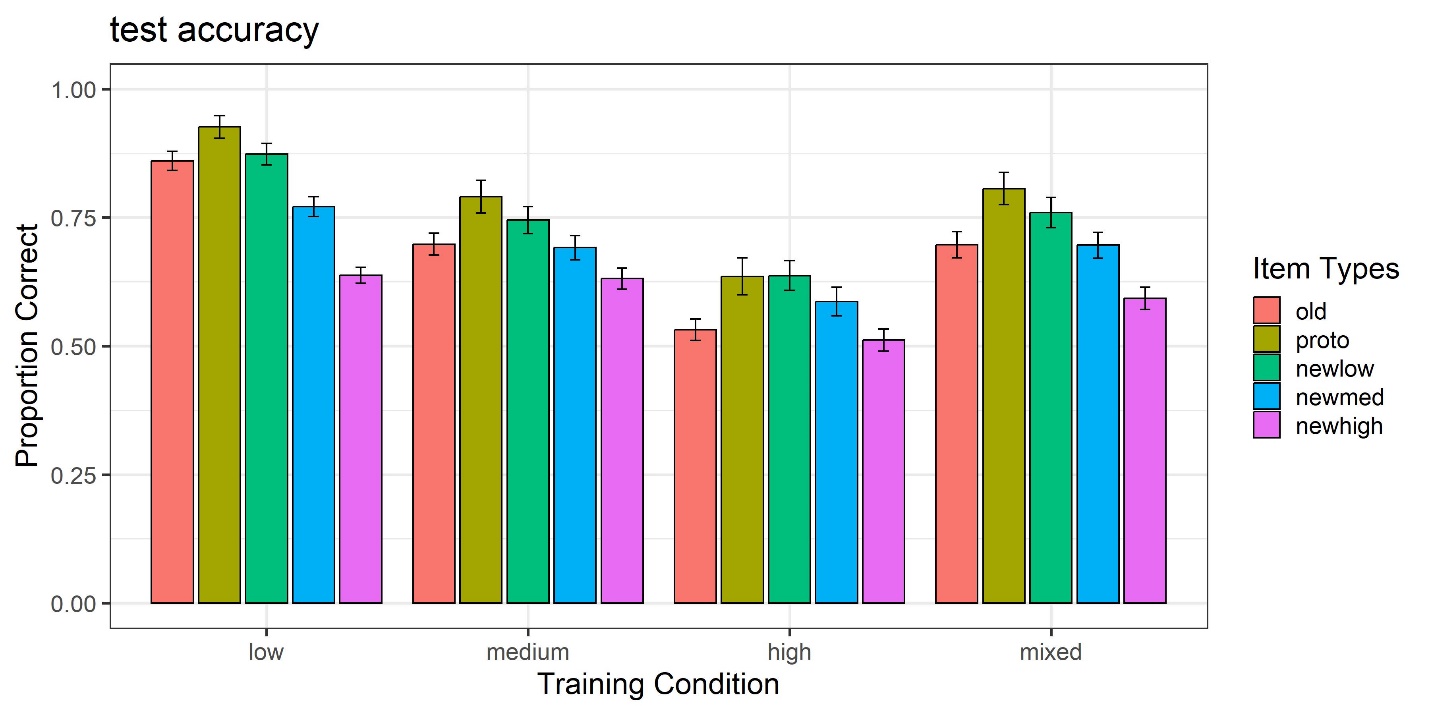
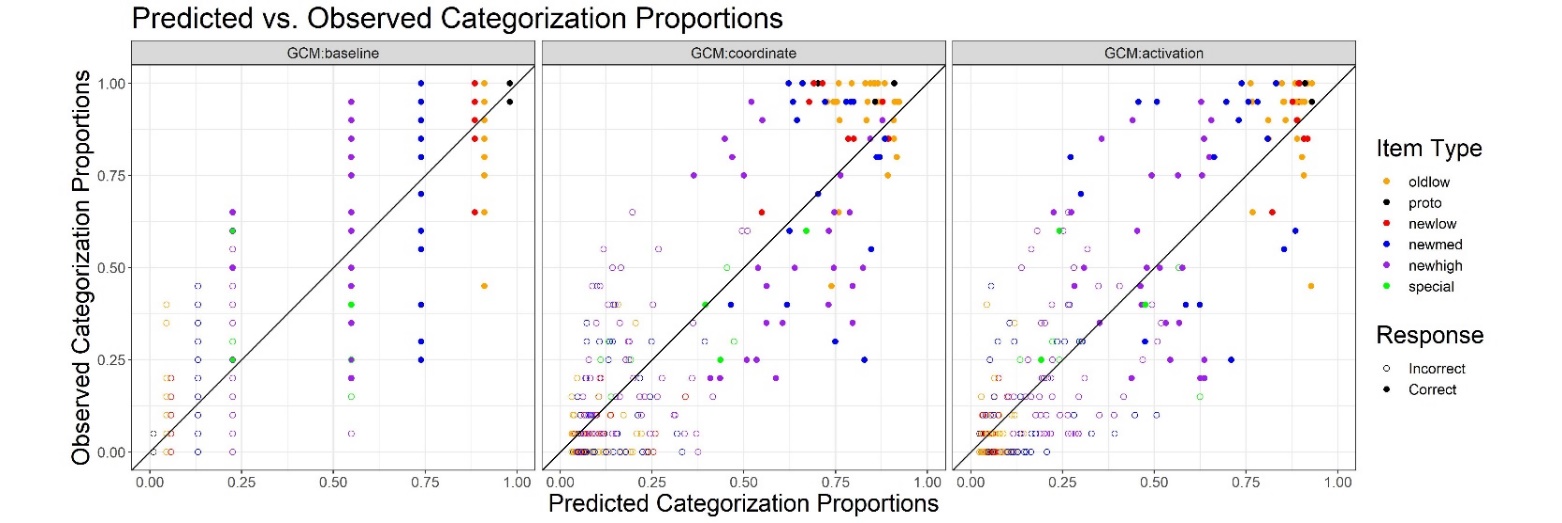


Figure 1

Figure 2

Figure 3

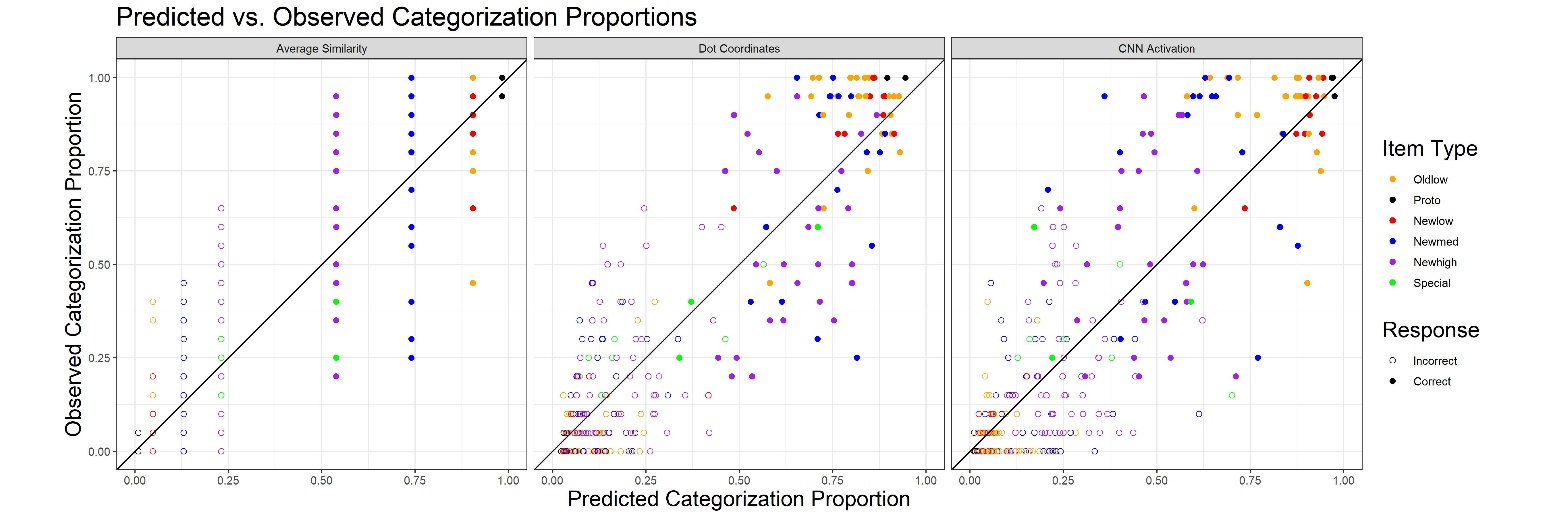


Figure 4

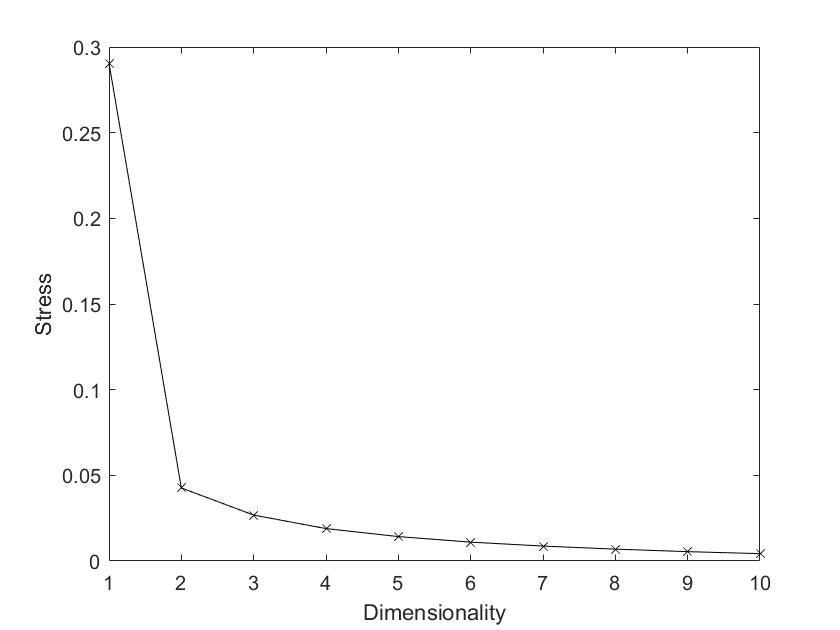


Figure 5

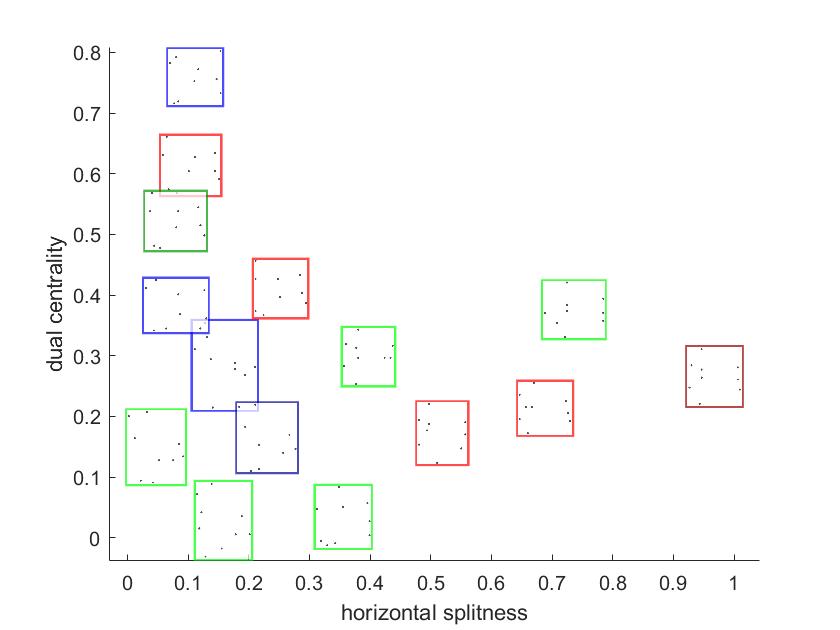
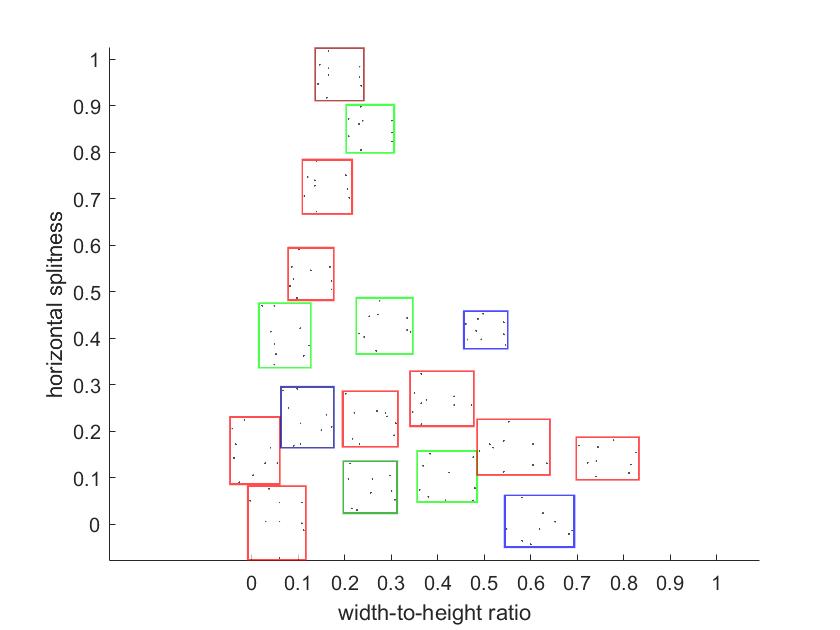


Figure 6

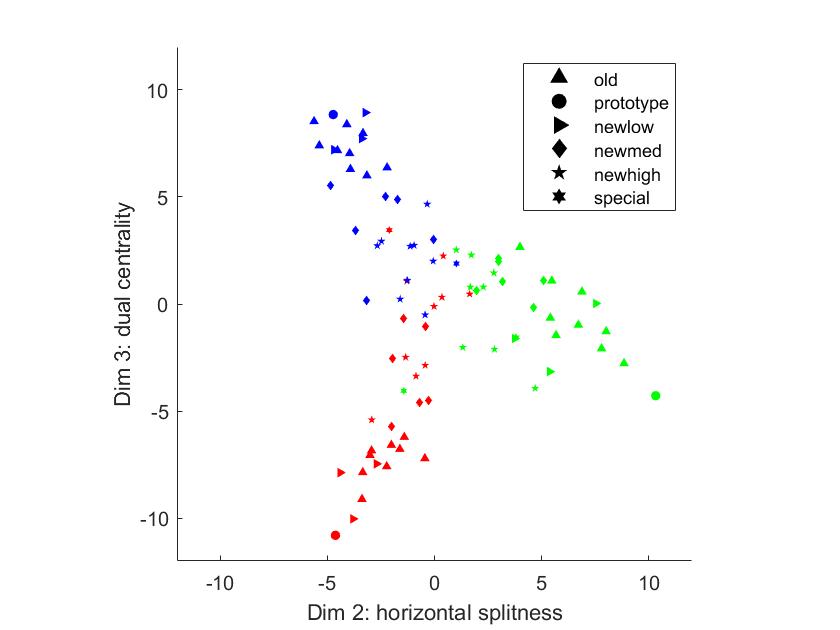
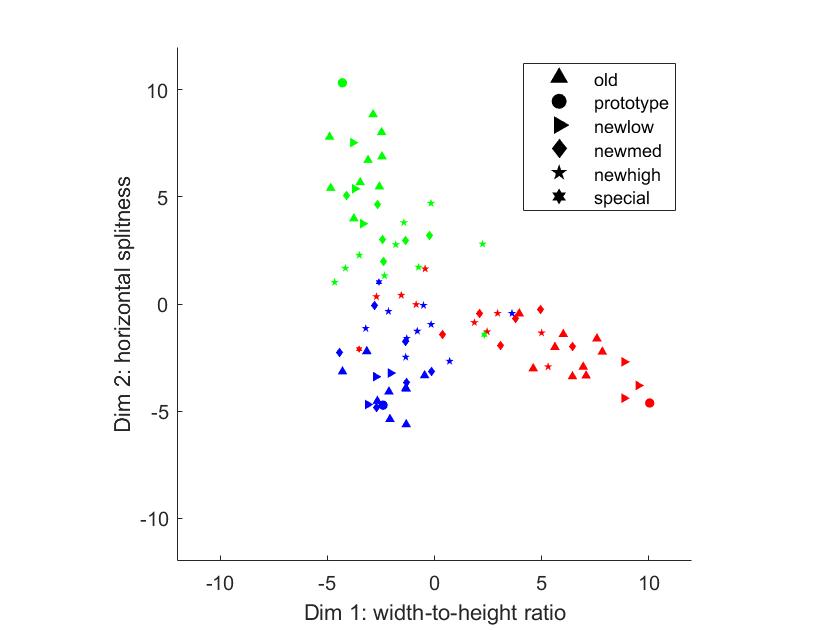


Figure 7. MDS configurations for the test patterns. The different colors indicate the category membership of individual test patterns. The different shapes represent different item types of the patterns. The left panel shows the coordinate values of test patterns on dimensions 1 and 2, and the right panel shows the values on dimensions 2 and 3.

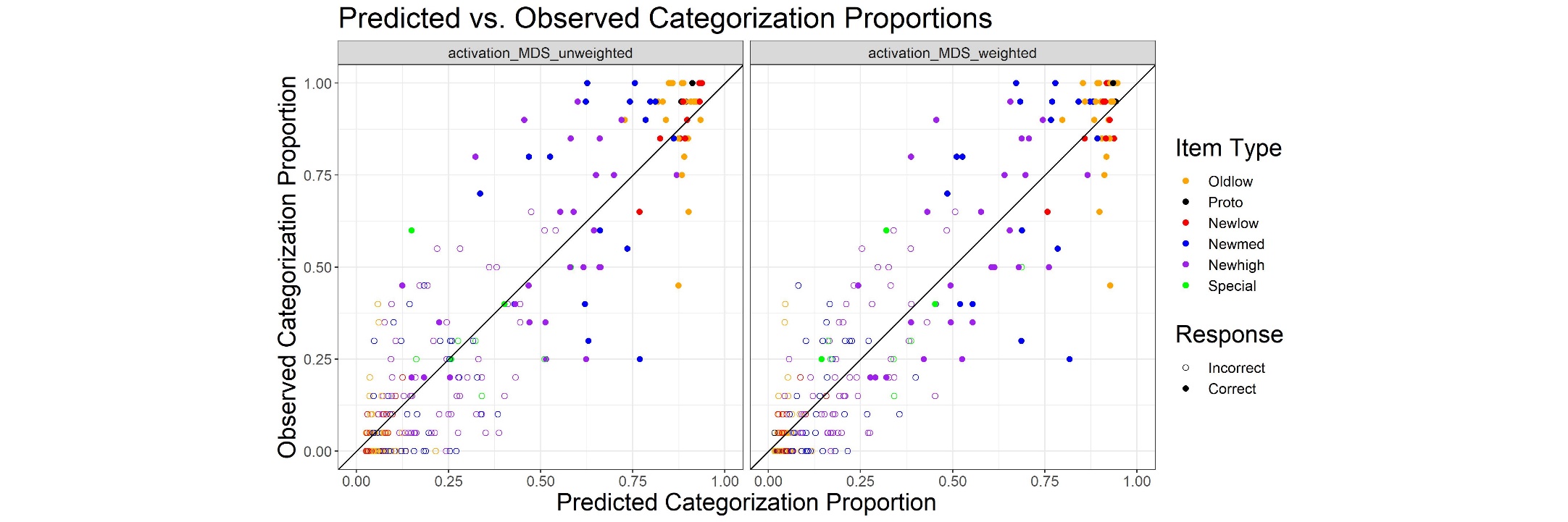


Figure 8

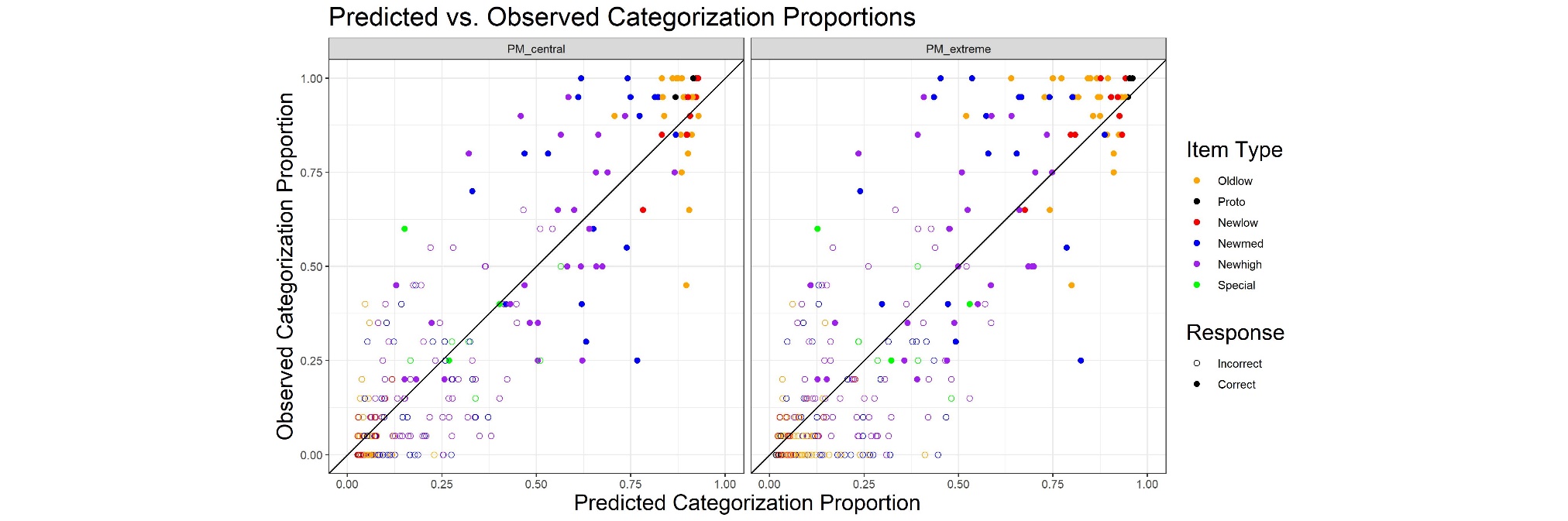


Figure 9 (unweighted)

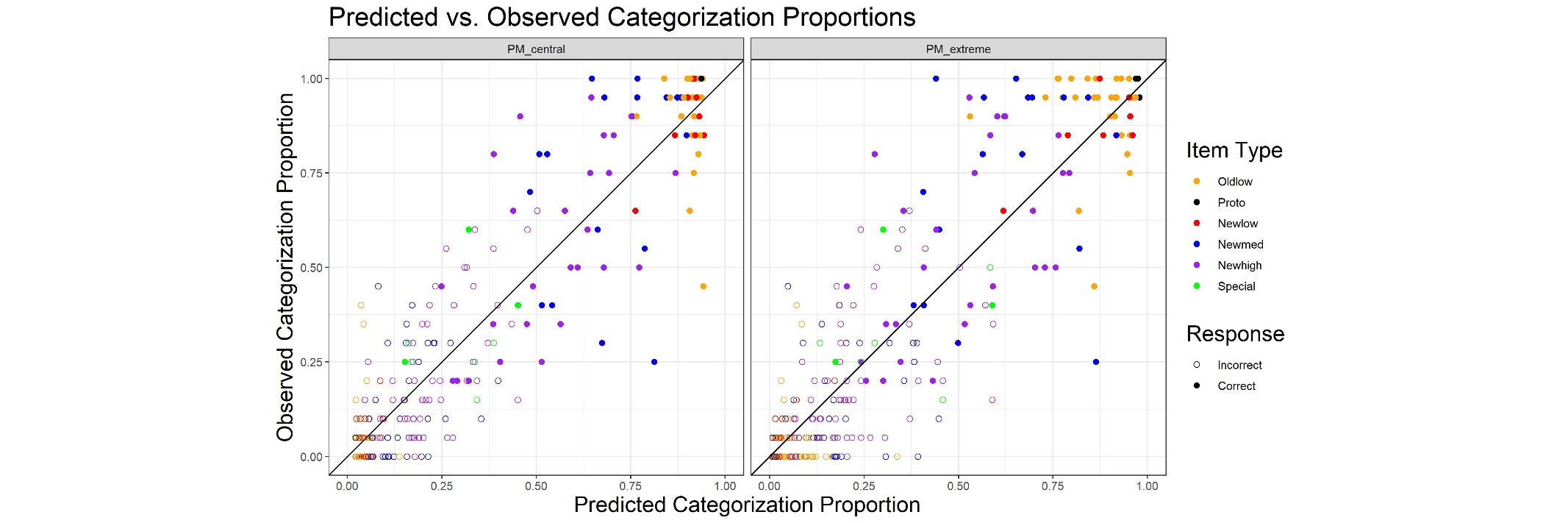


Figure 10 (weighted)

1. Metric scaling is used because we believe the Euclidean distances between deep-learning activation vectors are interval-scaled. In other words, we intend for the MDS solution to capture the numerical differences between the dissimilarities of dot patterns in addition to the mere ordering, which could be achieved by non-metric scaling. Nevertheless, the resulting MDS solution turned out to be virtually the same for both scaling options. [↑](#footnote-ref-1)