HTW

Thomas Gorman ⬤   Rob Goldstone ⬤
Indiana University   Indiana University
tegorman@iu.edu   rgoldsto@indiana.edu

ABSTRACT   In project 1, we applied model-based techniques to quantify and control for the similarity between training and testing experience, which in turn enabled us to account for the difference between varied and constant training via an extended version of a similarity based generalization model. In project 2, we will go a step further, implementing a full process model capable of both 1) producing novel responses and 2) modeling behavior in both the learning and testing stages of the experiment. Project 2 also places a greater emphasis on extrapolation performance following training - as varied training has often been purported to be particularly beneficial in such situations.

KEYWORDS   Learning Generalization; Function Learning; Visuomotor learning; Training Variability

Introduction

In project 1, I applied model-based techniques to quantify and control for the similarity between training and testing experience, which in turn enabled us to account for the difference between varied and constant training via an extended version of a similarity based generalization model. In project 2, we will go a step further, implementing a full process model capable of both 1) producing novel responses and 2) modeling behavior in both the learning and testing stages of the experiment. Project 2 also places a greater emphasis on extrapolation performance following training. Although varied training has often been purported to be particularly beneficial for generalization or transfer, few experiments have compared varied and constant training in contexts with unambiguous extrapolation testing.

Function Learning and Extrapolation

The study of human function learning investigates how people learn relationships between continuous input and output values. Function learning is studied both in tasks where individuals are exposed to a sequence of input/output pairs (DeLosh et al., 1997; McDaniel et al., 2013), or situations where observers are presented with a an incomplete scatterplot or line graph and make predictions about regions of the plot that don't contain data (Ciccione & Dehaene, 2021; Courrieu, 2012; Said & Fischer, 2021; Schulz et al., 2020).

Carroll (1963) conducted the earliest work on function learning. Input stimuli and output responses were both lines of varying length. The correct output response was related to the length of the input line by a linear, quadratic, or random function. Participants in the linear and quadratic performed above chance levels during extrapolation testing, with those in the linear condition performing the best overall. Carroll argued that these results were best explained by a ruled based model wherein learners form an abstract representation of the underlying function. Subsequent work by Brehmer (1974),testing a wider array of functional forms, provided further evidence for superior extrapolation in tasks with linear functions. Brehmer argued that individuals start out with an assumption of a linear function, but given sufficient error will progressively test alternative hypothesis with polynomials of greater degree. Koh & Meyer (1991) employed a visuomotor function learning task, wherein participants were trained on examples from an unknown function relating the length of an input line to the duration of a response (time between keystrokes). In this domain, participants performed best when the relation between line length and response duration was determined by a power, as opposed to linear function. Koh & Meyer developed the log-polynomial adaptive-regression model to account for their results.

The first significant challenge to the rule-based accounts of function learning was put forth by DeLosh et al. (1997) . In their task, participants learned to associate stimulus magnitudes with response magnitudes that were related via either linear, exponential, or quadratic function. Participants approached ceiling performance by the end of training in each function condition, and were able to correctly respond in interpolation testing trials. All three conditions demonstrated some capacity for extrapolation, however participants in the linear condition tended to underestimate the true function, while exponential and quadratic participants reliably overestimated the true function on extrapolation trials. Extrapolation and interpolation performance are depicted in Figure 1.

The authors evaluated both of the rule-based models introduced in earlier research (with some modifications enabling trial-by-trial learning). The polynomial hypothesis testing model (Brehmer, 1974; Carroll, 1963) tended to mimic the true function closely in extrapolation, and thus offered a poor account of the human data. The log-polynomial adaptive regression model (Koh & Meyer, 1991) was able to mimic some of the systematic deviations produced by human subjects, but also predicted overestimation in cases where underestimation occurred.

The authors also introduced two new function-learning models. The Associative Learning Model (ALM) and the extrapolation-association model (EXAM). ALM is a two layer connectionist model adapted from the ALCOVE model in the category learning literature (Kruschke, 1992). ALM belongs to the general class of radial-basis function neural networks, and can be considered a similarity-based model in the sense that the nodes in the input layer of the network are activated as a function of distance. The EXAM model retains the same similarity based activation and associative learning mechanisms as ALM, while being augmented with a linear rule response mechanism. When presented with novel stimuli, EXAM will retrieve the most similar input-output examples encountered during training, and from those examples compute a local slope. ALM was able to provide a good account of participant training and interpolation data in all three function conditions, however it was unable to extrapolate. EXAM, on the other hand, was able to reproduce both the extrapolation underestimation, as well as the quadratic and exponential overestimation patterns exhibited by the human participants. Subsequent research identified some limitations in EXAM's ability to account for cases where human participants learn and extrapolate sinusoidal function Bott & Heit (2004) or to scenarios where different functions apply to different regions of the input space Kalish et al. (2004), though EXAM has been shown to provide a good account of human learning and extrapolation in tasks with bi-linear, V shaped input spaces Mcdaniel et al. (2009).
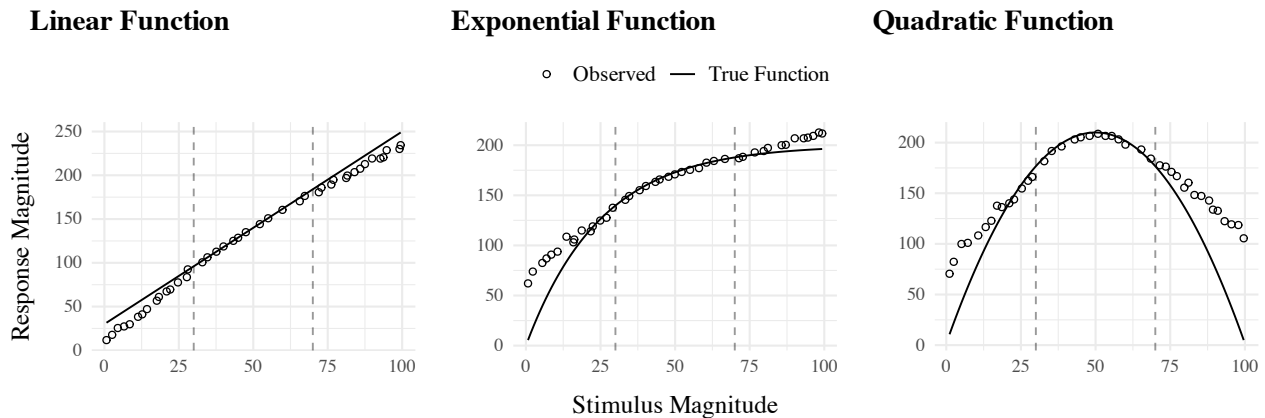
Linear Function   ### Exponential Function   ### Quadratic Function

Figure 1: Generalization reproduced patterns from DeLosh et al. (1997) Figure 3. Stimulii that fall within the dashed lines are interpolations of the training examples.

Methods

Participants

Data was collected from 647 participants (after exclusions). The results shown below consider data from subjects in our initial experiment, which consisted of 196 participants (106 constant, 90 varied). The follow-up experiments entailed minor manipulations: 1) reversing the velocity bands that were trained on vs. novel during testing; 2) providing ordinal rather than numerical feedback during training (e.g. correct, too low, too high). The data from these subsequent experiments are largely consistently with our initial results shown below.

Task

We developed a novel visuomotor extrapolation task, termed the Hit The Wall task, wherein participants learned to launch a projectile such that it hit a rectangle at the far end of the screen with an appropriate amount of force. Although the projectile had both x and y velocity components, only the x-dimension was relevant for the task. Link to task demo

Procedure

The HTW task involved launching projectiles to hit a target displayed on the computer screen. Participants completed a total of 90 trials during the training stage. In the varied training condition, participants encountered three velocity bands (800-1000, 1000-1200, and 1200-1400). In contrast, participants in the constant training condition encountered only one velocity band (800-1000).

During the training stage, participants in both conditions also completed "no feedback" trials, where they received no information about their performance. These trials were randomly interleaved with the regular training trials.

Following the training stage, participants proceeded to the testing stage, which consisted of three phases. In the first phase, participants completed "no-feedback" testing from three novel extrapolation bands (100-300, 350-550, and 600-800), with each band consisting of 15 trials.

In the second phase of testing, participants completed "no-feedback" testing from the three velocity bands used during the training stage (800-1000, 1000-1200, and 1200-1400). In the constant training condition, two of these bands were novel, while in the varied training condition, all three bands were encountered during training.

The third and final phase of testing involved "feedback" testing for each of the three extrapolation bands (100-300, 350-550, and 600-800), with each band consisting of 10 trials. Participants received feedback on their performance during this phase.

Throughout the experiment, participants' performance was measured by calculating the distance between the produced x-velocity of the projectiles and the closest edge of the current velocity band. Lower distances indicated better performance.

After completing the experiment, participants were debriefed and provided with an opportunity to ask questions about the study.
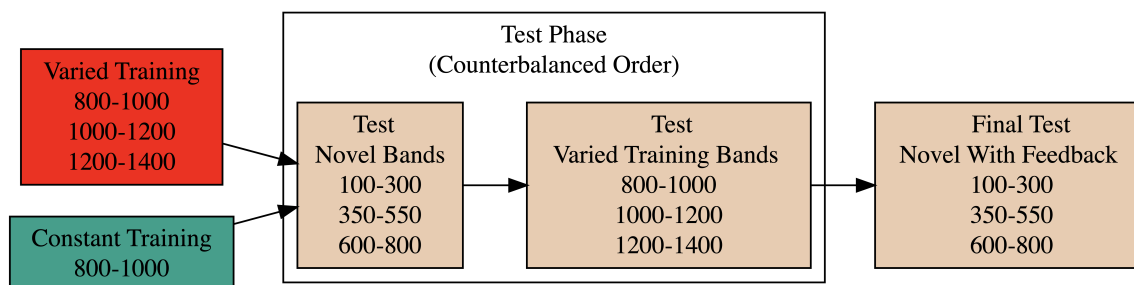

Figure 2: Experiment 1 Design. Constant and Varied participants complete different training conditions.

Analyses Strategy

All data processing and statistical analyses were performed in R version 4.31 Team (2020). To assess differences between groups, we used Bayesian Mixed Effects Regression. Model fitting was performed with the brms package in R Bürkner (2017), and descriptive stats and tables were extracted with the BayestestR package (?). Mixed effects regression enables us to take advantage of partial pooling, simultaneously estimating parameters at the individual and group level. Our use of Bayesian, rather than frequentist methods allows us to directly quantify the uncertainty in our parameter estimates, as well as circumventing convergence issues common to the frequentist analogues of our mixed models. For each model, we report the median values of the posterior distribution, and 95% credible intervals.

Each model was set to run with 4 chains, 5000 iterations per chain, with the first 2500 of which were discarded as warmup chains. Rhat values were generally within an acceptable range, with values <=1.02 (see appendix for diagnostic plots). We used uninformative priors for the fixed effects of the model (condition and velocity band), and weakly informative Student T distributions for for the random effects.

We compared varied and constant performance across two measures, deviation and discrimination. Deviation was quantified as the absolute deviation from the nearest boundary of the velocity band, or set to 0 if the throw velocity fell anywhere inside the target band. Thus, when the target band was 600-800, throws of 400, 650, and 1100 would result in deviation values of 200, 0, and 300, respectively. Discrimination was measured by fitting a linear model to the testing throws of each subjects, with the lower end of the target velocity band as the predicted variable, and the x velocity produced by the participants as the predictor variable. Participants who reliably discriminated between velocity bands tended to have positive slopes with values ~1, while participants who made throws irrespective of the current target band would have slopes ~0.

Table 1: Testing Deviation - Empirical Summary

(a) Full datasets

Band	Band Type	Mean	Median	Sd
100-300	Extrapolation	254	148	298
350-550	Extrapolation	191	110	229
600-800	Extrapolation	150	84	184
800-1000	Trained	184	106	242
1000-1200	Extrapolation	233	157	282
1200-1400	Extrapolation	287	214	290

(b) Intersection of samples with all labels available

Band	Band Type	Mean	Median	Sd
100-300	Extrapolation	386	233	426
350-550	Extrapolation	285	149	340
600-800	Extrapolation	234	144	270
800-1000	Trained	221	149	248
1000-1200	Trained	208	142	226
1200-1400	Trained	242	182	235

Results

Testing Phase - No feedback.

In the first part of the testing phase, participants are tested from each of the velocity bands, and receive no feedback after each throw.

Deviation From Target Band Descriptive summaries testing deviation data are provided in Table 1 and Figure 3. To model differences in accuracy between groups, we used Bayesian mixed effects regression models to the trial level data from the testing phase. The primary model predicted the absolute deviation from the target velocity band (dist) as a function of training condition (condit), target velocity band (band), and their interaction, with random intercepts and slopes for each participant (id).

$$dist_{ij} = \beta_0 + \beta_1 \cdot condit_{ij} + \beta_2 \cdot band_{ij} + \beta_3 \cdot condit_{ij} \cdot band_{ij} + b_{0i} + b_{1i} \cdot band_{ij} + \epsilon_{ij} \tag{1}$$
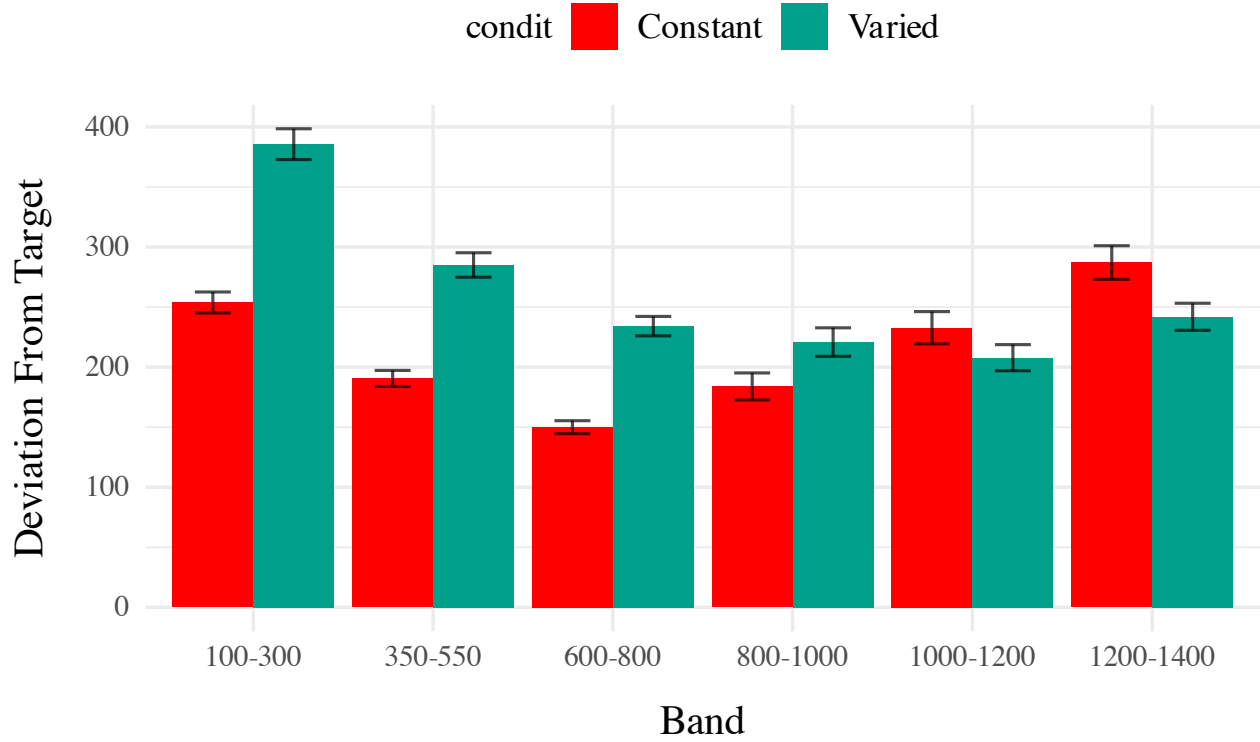
Figure 3: E1. Deviations from target band during testing without feedback stage.

Table 2: Experiment 1. Bayesian Mixed Model predicting absolute deviation as a function of condition (Constant vs. Varied) and Velocity Band

(a) Constant Testing1 - Deviation

Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	205.09	136.86	274.06	1.00
conditVaried	157.44	60.53	254.90	1.00
Band	0.01	-0.07	0.08	0.57
condit*Band	-0.16	-0.26	-0.06	1.00

(b) Varied Testing - Deviation

contrast	Band	value	lower	upper	pd
Constant - Varied	100	-141.49	-229.2	-53.83	1.00
Constant - Varied	350	-101.79	-165.6	-36.32	1.00
Constant - Varied	600	-62.02	-106.2	-14.77	1.00
Constant - Varied	800	-30.11	-65.1	6.98	0.94
Constant - Varied	1000	2.05	-33.5	38.41	0.54
Constant - Varied	1200	33.96	-11.9	81.01	0.92

The model predicting absolute deviation (dist) showed clear effects of both training condition and target velocity band (Table X). Overall, the varied training group showed a larger deviation relative to the constant training group (β = 157.44, 95% CI [60.53, 254.9]). Deviation also depended on target velocity band, with lower bands showing less deviation. See Table 2 for full model output.

Discrimination between bands In addition to accuracy/deviation, we also assessed the ability of participants to reliably discriminate between the velocity bands (i.e. responding differently when prompted for band 600-800 than when

prompted for band 150-350). Table 3 shows descriptive statistics of this measure, and Figure 1 visualizes the full distributions of throws for each combination of condition and velocity band. To quantify discrimination, we again fit Bayesian Mixed Models as above, but this time the dependent variable was the raw x velocity generated by participants on each testing trial.

$$vx_{ij} = \beta_0 + \beta_1 \cdot condit_{ij} + \beta_2 \cdot bandInt_{ij} + \beta_3 \cdot condit_{ij} \cdot bandInt_{ij} + b_{0i} + b_{1i} \cdot bandInt_{ij} + \epsilon_{ij} \tag{2}$$

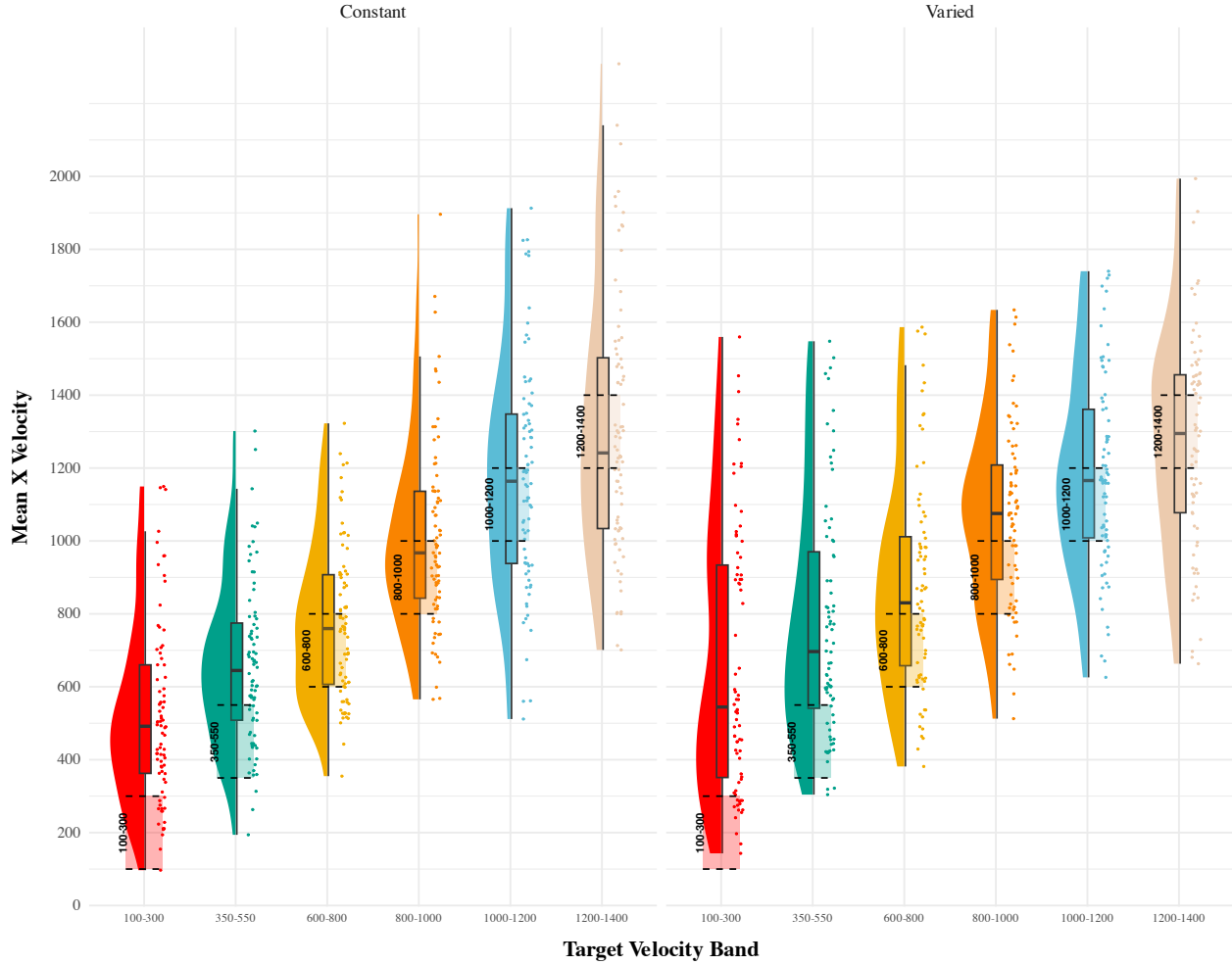**Testing Performance (no-feedback) - X-Velocity Per Band**

Figure 4: E1 testing x velocities. Translucent bands with dash lines indicate the correct range for each velocity band.

Table 3: Testing vx - Empirical Summary

(a) Constant

Band	Band Type	Mean	Median	Sd
100-300	Extrapolation	524	448	327
350-550	Extrapolation	659	624	303
600-800	Extrapolation	770	724	300
800-1000	Trained	1001	940	357
1000-1200	Extrapolation	1167	1104	430
1200-1400	Extrapolation	1283	1225	483

(b) Varied

Band	Band Type	Mean	Median	Sd
100-300	Extrapolation	664	533	448
350-550	Extrapolation	768	677	402
600-800	Extrapolation	876	813	390
800-1000	Trained	1064	1029	370
1000-1200	Trained	1180	1179	372
1200-1400	Trained	1265	1249	412

Table 4: Experiment 1. Bayesian Mixed Model Predicting Vx as a function of condition (Constant vs. Varied) and Velocity Band

(a) Model fit to all 6 bands

Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	408.55	327.00	490.61	1.00
conditVaried	164.05	45.50	278.85	1.00
Band	0.71	0.62	0.80	1.00
condit*Band	-0.14	-0.26	-0.01	0.98

(b) Model fit to 3 extrapolation bands

Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	478.47	404.00	551.45	1.00
conditVaried	142.04	37.17	247.59	1.00
Band	0.50	0.42	0.57	1.00
condit*Band	-0.07	-0.17	0.04	0.89

See Table 4 for the full model results. The estimated coefficient for training condition (B = 164.05, 95% CrI [45.5, 278.85]) suggests that the varied group tends to produce harder throws than the constant group, but is not in and of itself useful for assessing discrimination. Most relevant to the issue of discrimination is the slope on Velocity Band (B = 0.71, 95% CrI [0.62, 0.8]). Although the median slope does fall underneath the ideal of value of 1, the fact that the 95% credible interval does not contain 0 provides strong evidence that participants exhibited some discrimination between bands. The estimate for the interaction between slope and condition (B = -0.14, 95% CrI [-0.26, -0.01]), suggests that the discrimination was somewhat modulated by training condition, with the varied participants showing less senitivity between vands than the constant condition. This difference is depicted visually in Figure 5.@tbl-e1-slope-quartile shows the average slope coefficients for varied and constant participants separately for each quartile. The constant participant participants appear to have larger slopes across quartiles, but the difference between conditions may be less pronounced for the top quartiles of subjects who show the strongest discrimination. Figure Figure 6 shows the distributions of slope values for each participant, and the compares the probability density of slope coefficients between training conditions. Figure 7

The second model, which focused solely on extrapolation bands, revealed similar patterns. The Velocity Band term ($B = 0.5$, 95% CrI [0.42, 0.57]) still demonstrates a high degree of discrimination ability. However, the posterior distribution for interaction term ($B = -0.07$, 95% CrI [-0.17, 0.04]) does across over 0, suggesting that the evidence for decreased discrimination ability for the varied participants is not as strong when considering only the three extrapolation bands.
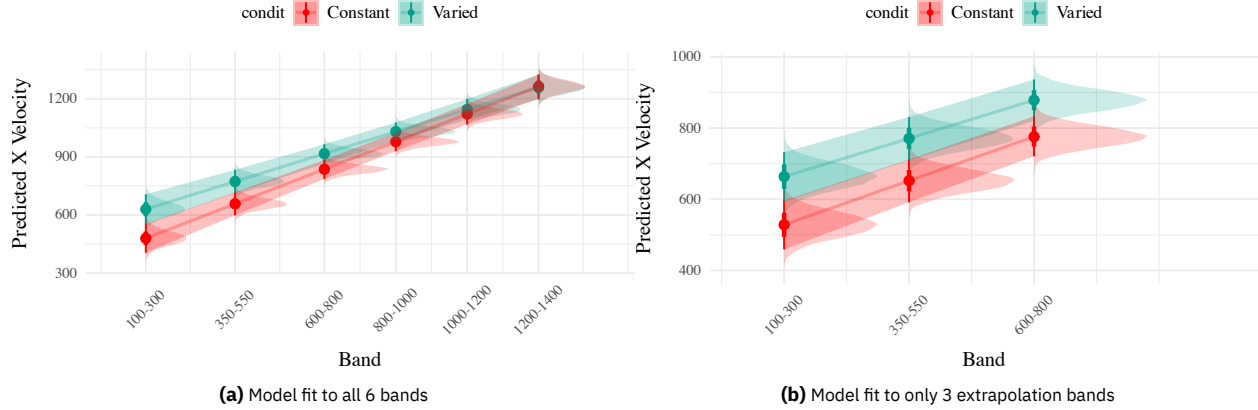

(a) Model fit to all 6 bands

(b) Model fit to only 3 extrapolation bands

Figure 5: Conditional effect of training condition and Band. Ribbons indicate 95% HDI. The steepness of the lines serves as an indicator of how well participants discriminated between velocity bands.

Table 5: Slope coefficients by quartile, per condition

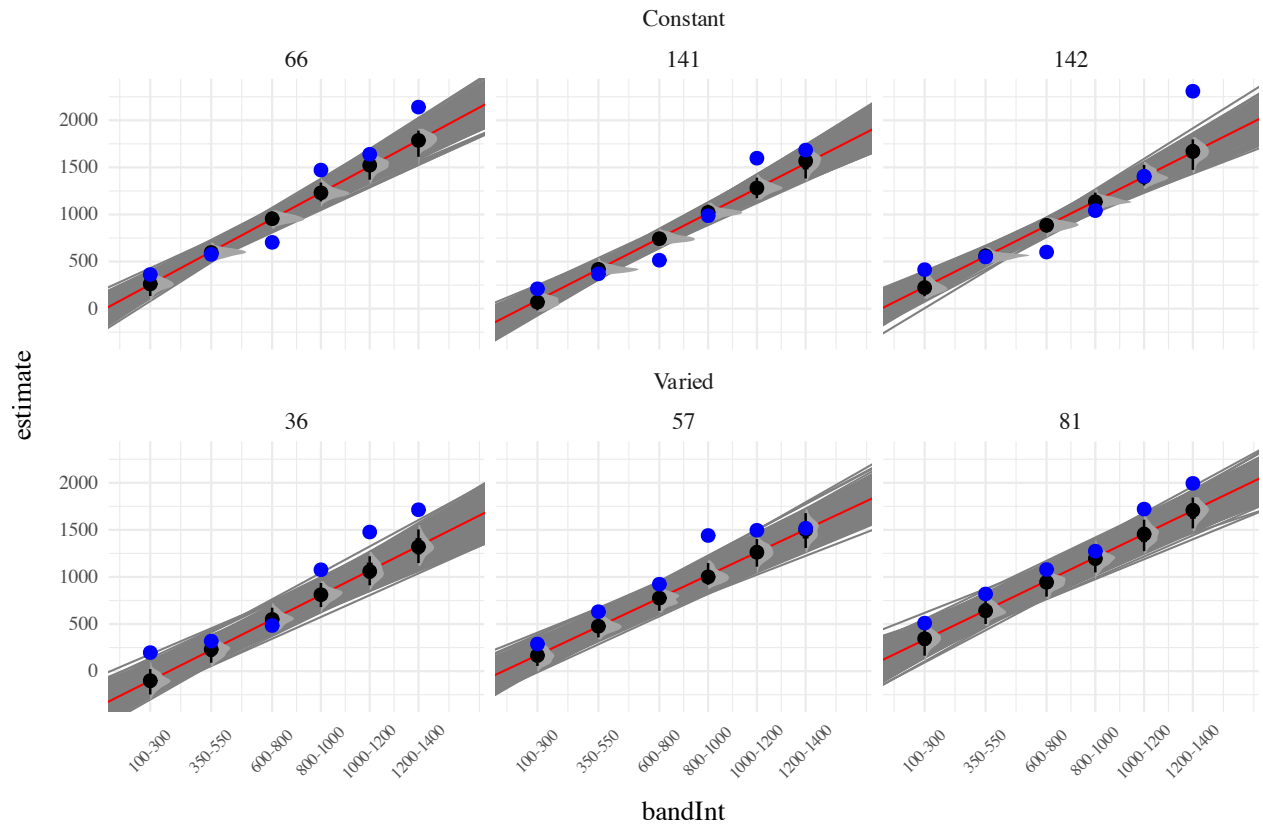| Condition	Q_0%_mean	Q_25%_mean	Q_50%_mean	Q_75%_mean	Q_100%_mean
Constant	-0.111	0.476	0.689	0.931	1.39
Varied	-0.205	0.266	0.587	0.896	1.30
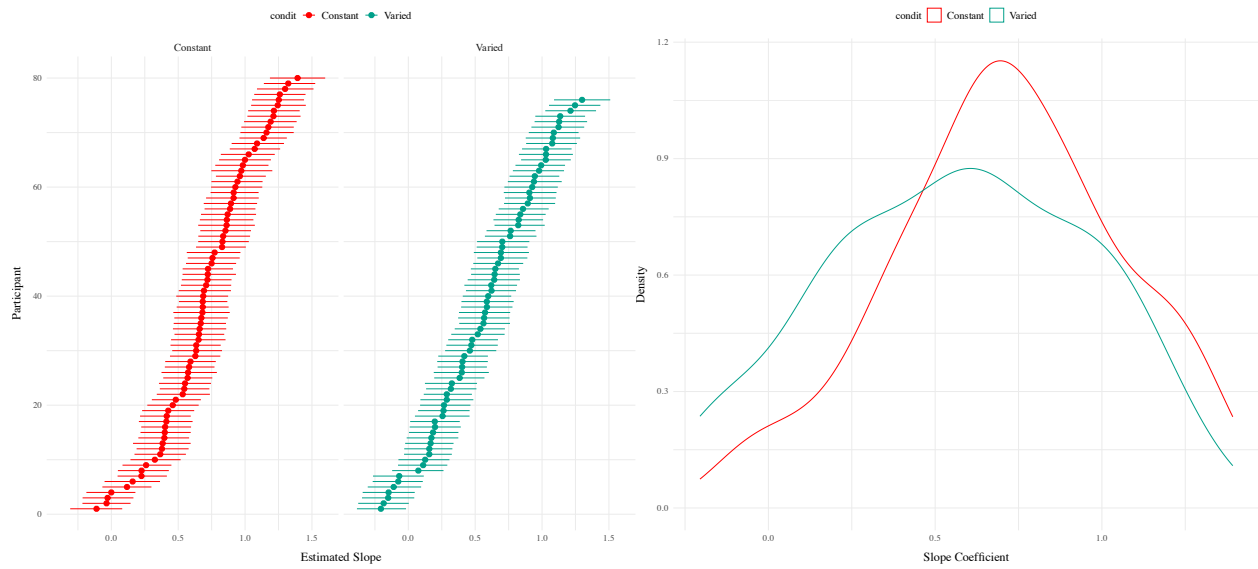
(a) Slope estimates by participant - ordered from lowest to highest within each condition.

(b) Destiny of slope coefficients by training group

Figure 6: Slope distributions between condition

test
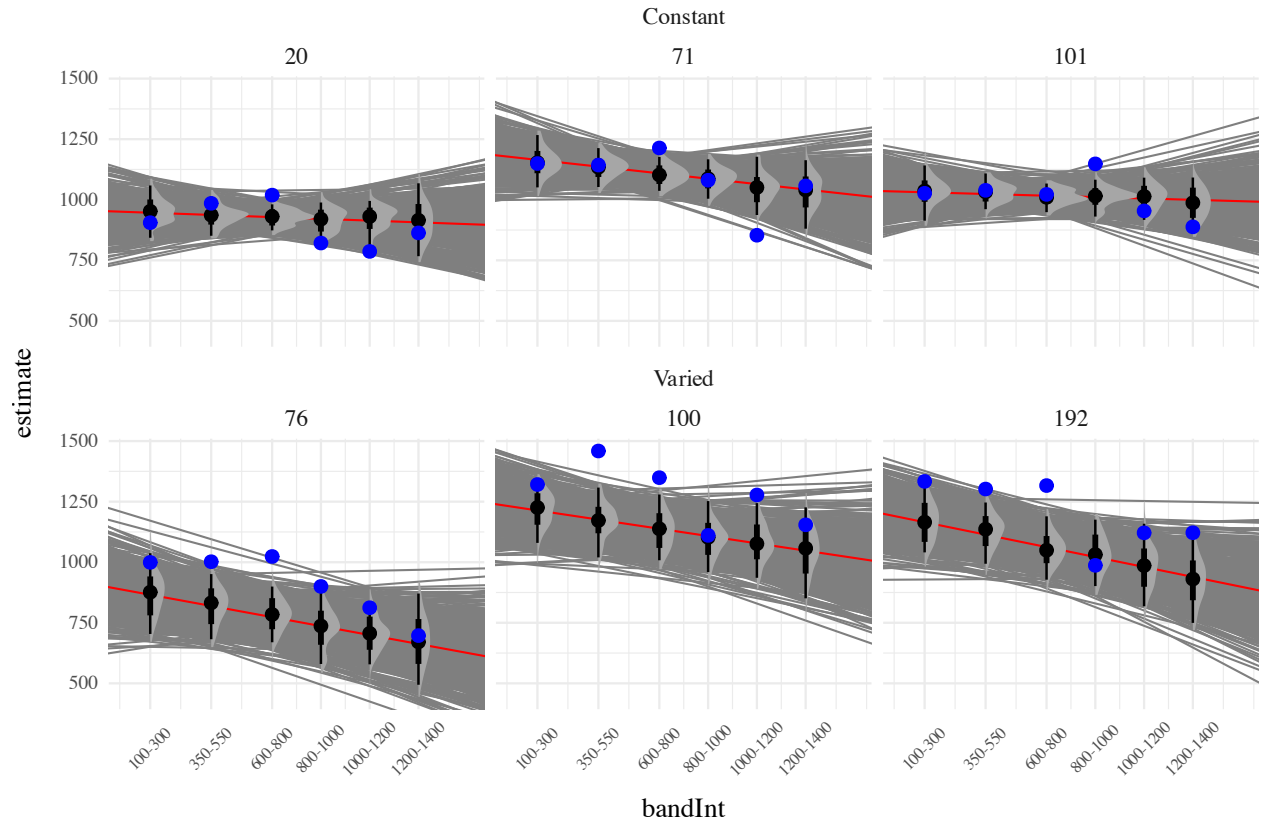
Figure 7: Subset of Varied and Constant Participants with the smallest and largest estimated slope values. Red lines represent the best fitting line for each participant, gray lines are 200 random samples from the posterior distribution. Colored points and intervals at each band represent the empirical median and 95% HDI.

Experiment 2

Figure 8 illustrates the design of Experiment 2. The stages of the experiment (i.e. training, testing no-feedback, test with feedback), are identical to that of Experiment 1. The only change is that Experiment 2 participants train, and then test, on bands in the reverse order of Experiment 1 (i.e. training on the softer bands; and testing on the harder bands).
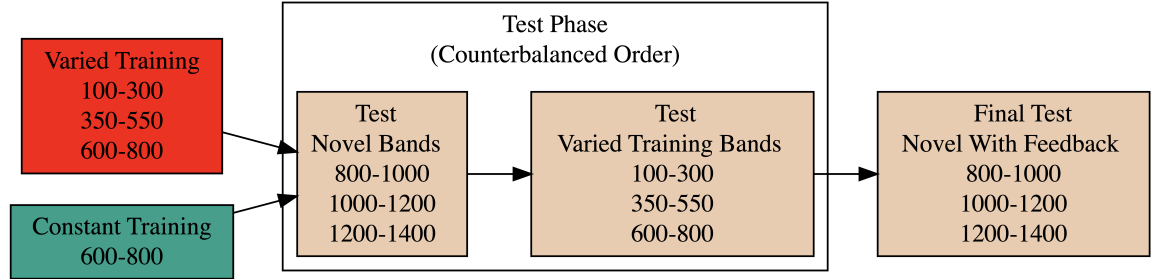

Figure 8: Experiment 2 Design. Constant and Varied participants complete different training conditions. The training and testing bands are the reverse of Experiment

E2 Results

Testing Phase - No feedback.
In the first part of the testing phase, participants are tested from each of the velocity bands, and receive no feedback after each throw.

Deviation From Target Band   Descriptive summaries testing deviation data are provided in Table 6 and Figure 9. To model differences in accuracy between groups, we used Bayesian mixed effects regression models to the trial level data from the testing phase. The primary model predicted the absolute deviation from the target velocity band (dist) as a function of training condition (condit), target velocity band (band), and their interaction, with random intercepts and slopes for each participant (id).

$$dist_{ij} = \beta_0 + \beta_1 \cdot condit_{ij} + \beta_2 \cdot band_{ij} + \beta_3 \cdot condit_{ij} \cdot band_{ij} + b_{0i} + b_{1i} \cdot band_{ij} + \epsilon_{ij} \tag{3}$$

Table 6: Testing Deviation - Empirical Summary

(a) Constant Testing - Deviation

Band	Band Type	Mean	Median	Sd
100-300	Extrapolation	206	48	317
350-550	Extrapolation	194	86	268
600-800	Trained	182	112	240
800-1000	Extrapolation	200	129	233
1000-1200	Extrapolation	238	190	234
1200-1400	Extrapolation	311	254	288

(b) Varied Testing - Deviation

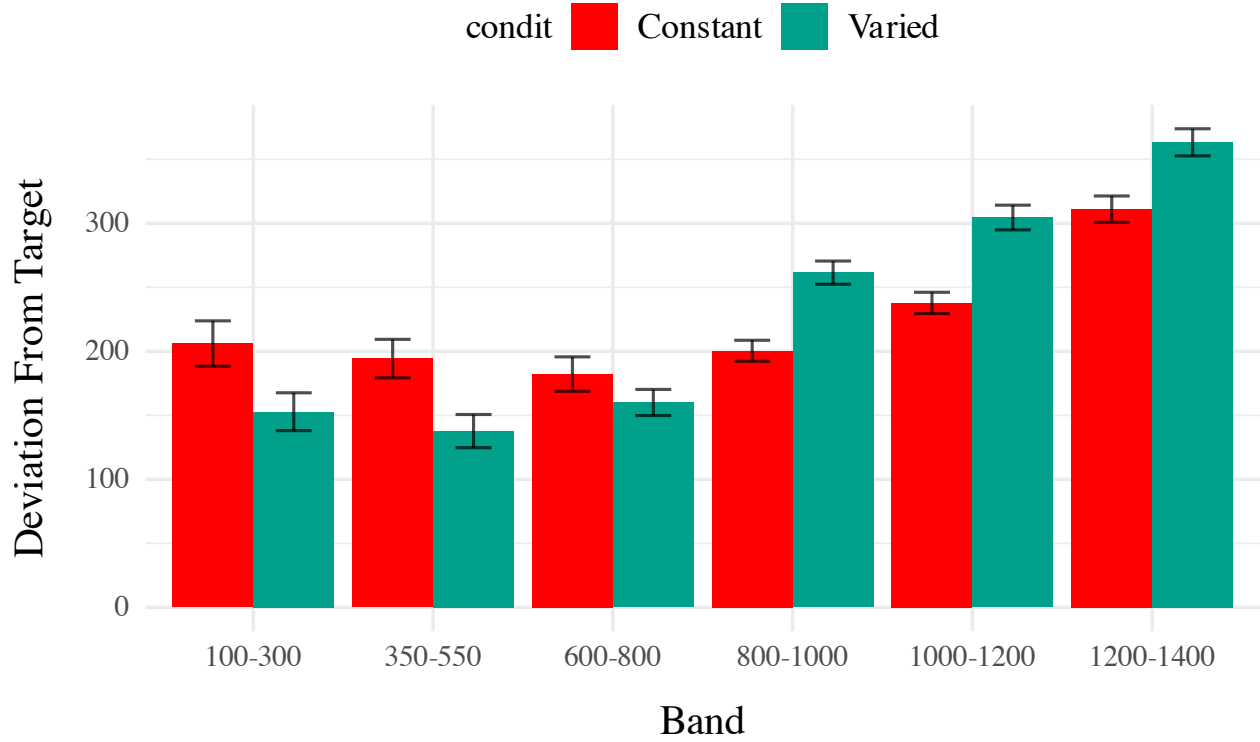| Band	Band Type	Mean	Median	Sd
100-300	Trained	153	25	266
350-550	Trained	138	53	233
600-800	Trained	160	120	183
800-1000	Extrapolation	261	207	257
1000-1200	Extrapolation	305	258	273
1200-1400	Extrapolation	363	314	297

Figure 9: E2. Deviations from target band during testing without feedback stage.

Table 7: Experiment 2. Bayesian Mixed Model predicting absolute deviation as a function of condition (Constant vs. Varied) and Velocity Band

(a) Model fits

Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	151.71	90.51	215.86	1.00
conditVaried	-70.33	-156.87	16.66	0.94
Band	0.10	0.02	0.18	1.00
condit*Band	0.12	0.02	0.23	0.99

(b) Contrasts

contrast	Band	value	lower	upper	pd
Constant - Varied	100	57.6	-20.5	135.32	0.93
Constant - Varied	350	26.6	-30.9	83.84	0.83
Constant - Varied	600	-4.3	-46.7	38.52	0.58
Constant - Varied	800	-29.3	-69.4	11.29	0.92
Constant - Varied	1000	-54.6	-101.1	-5.32	0.98
Constant - Varied	1200	-79.6	-139.5	-15.45	0.99

The model predicting absolute deviation showed a modest tendency for the varied training group to have lower deviation compared to the constant training group (β = -70.33, 95% CI [-156.87, 16.66]),with 94% of the posterior distribution being less than 0. This suggests a potential benefit of training with variation, though the evidence is not definitive.

Discrimination between Velocity Bands   In addition to accuracy/deviation. We also assessed the ability of participants to reliably discriminate between the velocity bands (i.e. responding differently when prompted for band 600-800 than

when prompted for band 150-350). Table 8 shows descriptive statistics of this measure, and Figure 1 visualizes the full distributions of throws for each combination of condition and velocity band. To quantify discrimination, we again fit Bayesian Mixed Models as above, but this time the dependent variable was the raw x velocity generated by participants.

$$vx_{ij} = \beta_0 + \beta_1 \cdot condit_{ij} + \beta_2 \cdot bandInt_{ij} + \beta_3 \cdot condit_{ij} \cdot bandInt_{ij} + b_{0i} + b_{1i} \cdot bandInt_{ij} + \epsilon_{ij} \qquad (4)$$

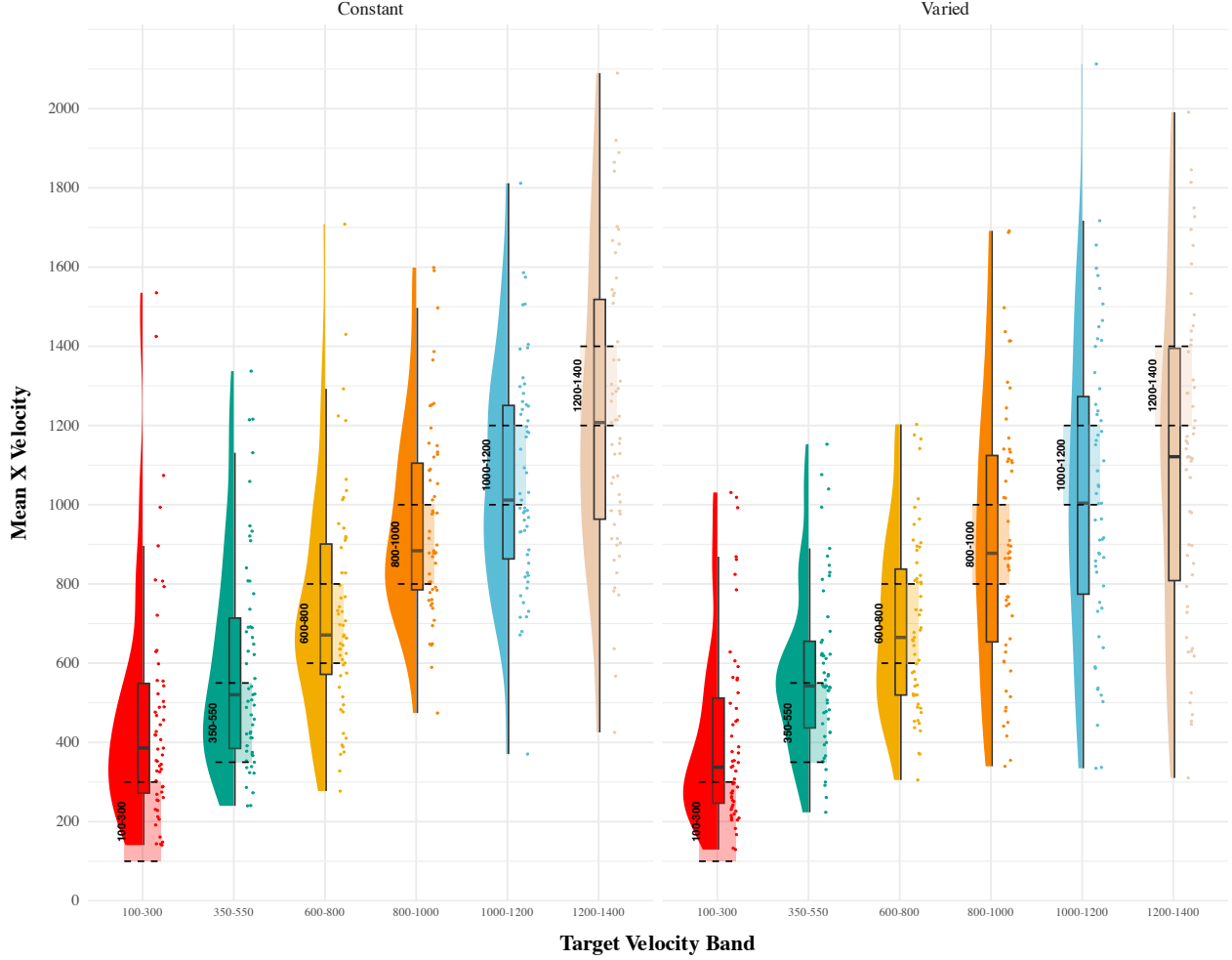### Testing Performance (no-feedback) - X-Velocity Per Band

Figure 10: E2 testing x velocities. Translucent bands with dash lines indicate the correct range for each velocity band.

Table 8: Testing vx - Empirical Summary

(a) Constant Testing - vx

Band	Band Type	Mean	Median	Sd
100-300	Extrapolation	457	346	354
350-550	Extrapolation	597	485	368
600-800	Trained	728	673	367
800-1000	Extrapolation	953	913	375
1000-1200	Extrapolation	1064	1012	408
1200-1400	Extrapolation	1213	1139	493

(b) Varied Testing - vx

Band	Band Type	Mean	Median	Sd
100-300	Trained	410	323	297
350-550	Trained	582	530	303
600-800	Trained	696	641	316
800-1000	Extrapolation	910	848	443
1000-1200	Extrapolation	1028	962	482
1200-1400	Extrapolation	1095	1051	510

Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	362.64	274.85	450.02	1.00
conditVaried	-8.56	-133.97	113.98	0.55
Band	0.71	0.58	0.84	1.00
condit*Band	-0.06	-0.24	0.13	0.73

See Table 9 for the full model results.

When examining discrimination ability using the model predicting raw x-velocity, the results were less clear than those of the absolute deviation analysis. The slope on Velocity Band (β = 0.71, 95% CrI [0.58, 0.84]) indicates that participants showed good discrimination between bands overall. However, the interaction term suggested this effect was not modulated by training condition (β = -0.06, 95% CrI [-0.24, 0.13]) Thus, while varied training may provide some advantage for accuracy, both training conditions seem to have similar abilities to discriminate between velocity bands.
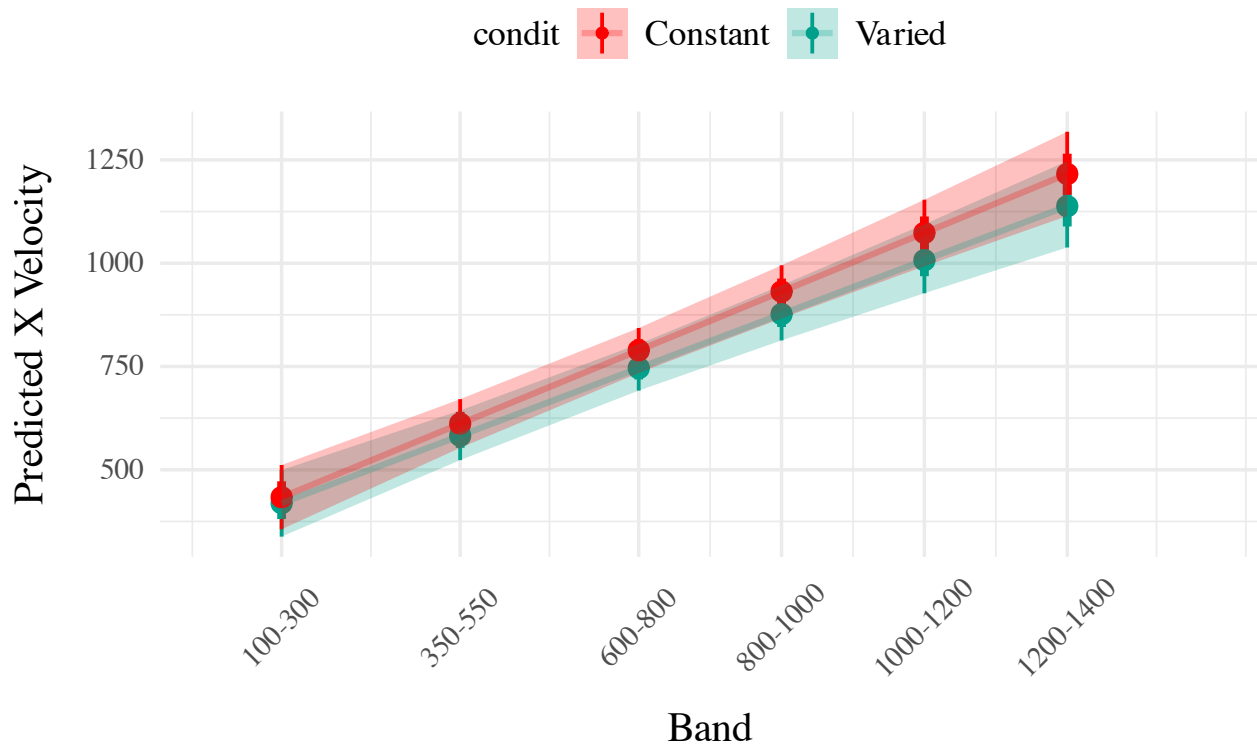
Figure 11: Conditional effect of training condition and Band. Ribbons indicate 95% HDI.

Experiment 3

The major manipulation adjustment of experiment 3 is for participants to receive ordinal feedback during training, in contrast to the continuous feedback of the earlier experiments. Ordinal feedback informs participants whether a throw was too soft, too hard, or fell within the target velocity range. Experiment 3 participants were randomly assigned to both a training condition (Constant vs. Varied) and a Band Order condition (original order used in Experiment 1, or the Reverse order of Experiment 2).

Results

Testing Phase - No feedback.

In the first part of the testing phase, participants are tested from each of the velocity bands, and receive no feedback after each throw. Note that these no-feedback testing trials are identical to those of Experiment 1 and 2, as the ordinal feedback only occurs during the training phase, and final testing phase, of Experiment 3.

Deviation From Target Band   Descriptive summaries testing deviation data are provided in Table 10 and Figure 12. To model differences in accuracy between groups, we fit Bayesian mixed effects regression models to the trial level data from the testing phase. The primary model predicted the absolute deviation from the target velocity band (dist) as a function of training condition (condit), target velocity band (band), and their interaction, with random intercepts and slopes for each participant (id).

Table 10: Testing Deviation - Empirical Summary

(a) Constant Testing - Deviation

Band	Band Type	Mean	Median	Sd
100-300	Extrapolation	396	325	350
350-550	Extrapolation	278	176	299
600-800	Extrapolation	173	102	215
800-1000	Trained	225	126	284
1000-1200	Extrapolation	253	192	271
1200-1400	Extrapolation	277	210	262

(b) Varied Testing - Deviation

Band	Band Type	Mean	Median	Sd
100-300	Extrapolation	383	254	385
350-550	Extrapolation	287	154	318
600-800	Extrapolation	213	140	244
800-1000	Trained	199	142	209
1000-1200	Trained	222	163	221
1200-1400	Trained	281	227	246

Band	Band Type	Mean	Median	Sd
100-300	Extrapolation	403	334	383
350-550	Extrapolation	246	149	287
600-800	Trained	155	82	209
800-1000	Extrapolation	207	151	241
1000-1200	Extrapolation	248	220	222
1200-1400	Extrapolation	322	281	264

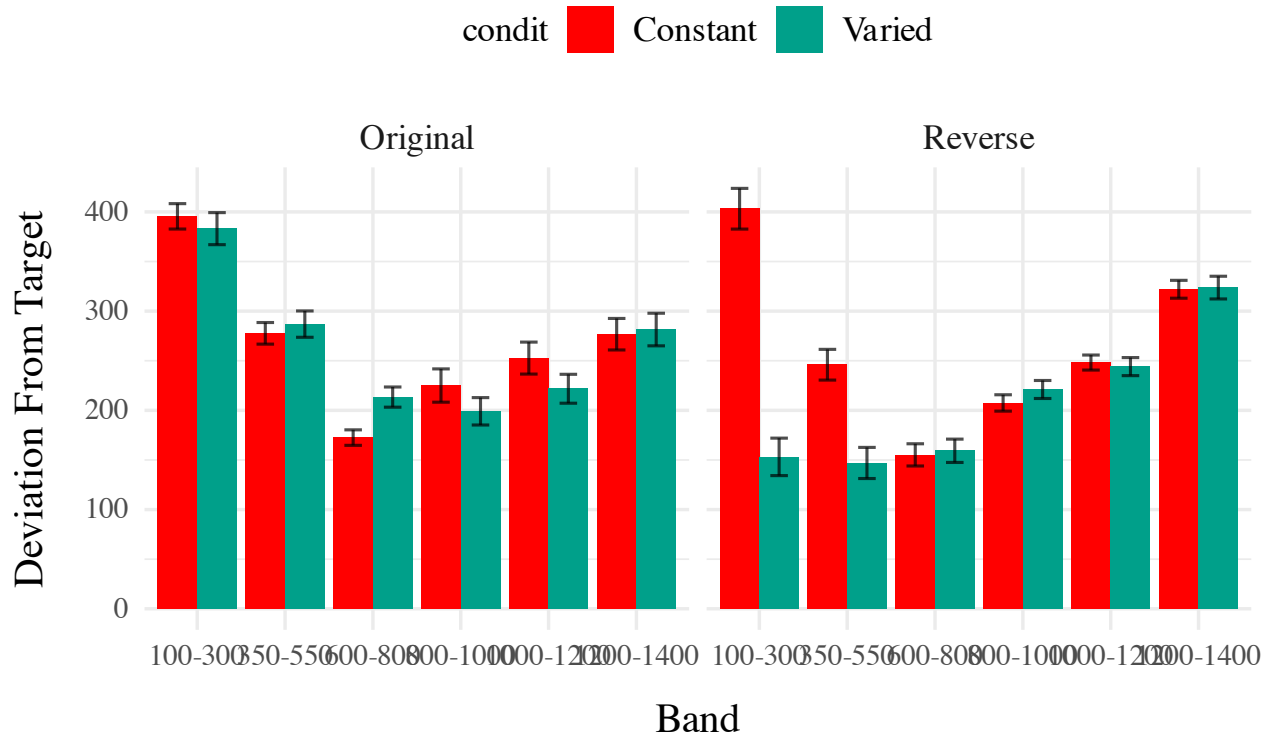| Band	Band Type	Mean	Median	Sd
100-300	Trained	153	0	307
350-550	Trained	147	55	258
600-800	Trained	159	107	192
800-1000	Extrapolation	221	160	235
1000-1200	Extrapolation	244	185	235
1200-1400	Extrapolation	324	264	291

Figure 12: e3. Deviations from target band during testing without feedback stage.

Table 11: Experiment 3. Bayesian Mixed Model predicting absolute deviation as a function of condition (Constant vs. Varied) and Velocity Band

Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	306.47	243.89	368.75	1.00
conditVaried	-90.65	-182.79	3.75	0.97
Band	-0.07	-0.13	0.00	0.97
condit*Band	0.09	-0.01	0.19	0.96

The effect of training condition in Experiment 3 showed a similar pattern to Experiment 2, with the varied group tending to have lower deviation than the constant group (β = -90.65, 95% CrI [-182.79, 3.75]), with 97% of the posterior distribution falling under 0.
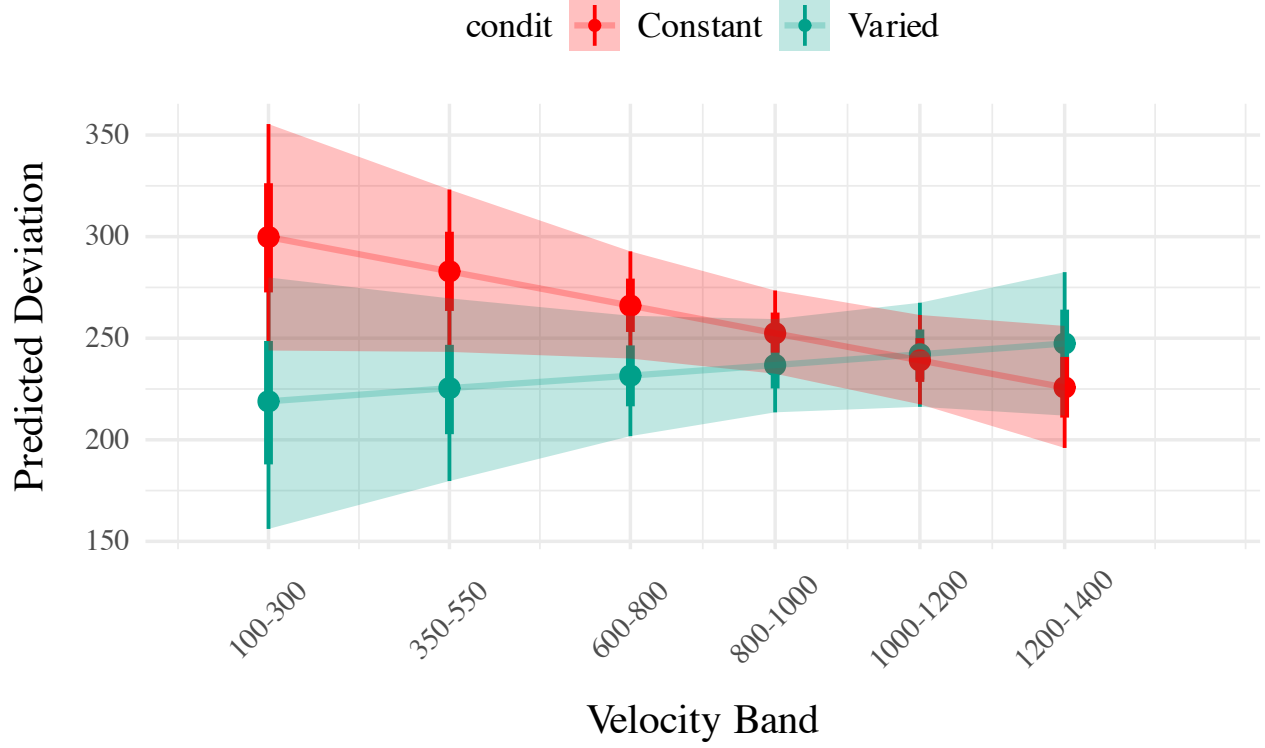
Figure 13: e3. Conditioinal Effect of Training Condition and Band. Ribbon indicated 95% Credible Intervals.

Discrimination between Velocity Bands   In addition to accuracy/deviation. We also assessed the ability of participants to reliably discriminate between the velocity bands (i.e. responding differently when prompted for band 600-800 than when prompted for band 150-350). Table 12 shows descriptive statistics of this measure, and Figure 1 visualizes the full distributions of throws for each combination of condition and velocity band. To quantify discrimination, we again fit Bayesian Mixed Models as above, but this time the dependent variable was the raw x velocity generated by participants.

$$vx_{ij} = \beta_0 + \beta_1 \cdot condit_{ij} + \beta_2 \cdot bandInt_{ij} + \beta_3 \cdot condit_{ij} \cdot bandInt_{ij} + b_{0i} + b_{1i} \cdot bandInt_{ij} + \epsilon_{ij} \tag{5}$$

Testing Performance (no-feedback) - X-Velocity Per Band
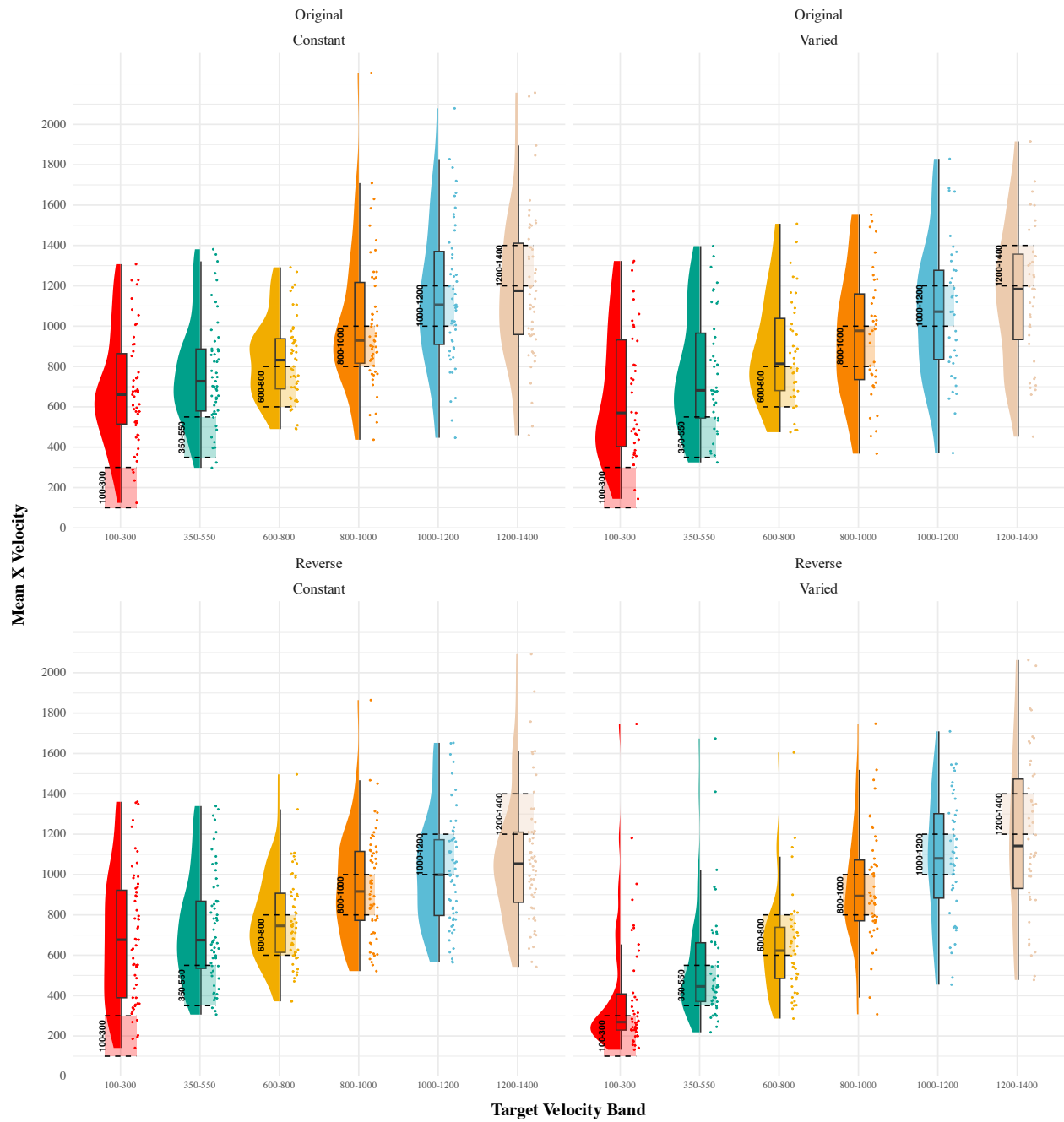

Figure 14: e3 testing x velocities. Translucent bands with dash lines indicate the correct range for each velocity band.

Table 12: Testing vx - Empirical Summary

(a) Constant Testing - vx

Band	Band Type	Mean	Median	Sd
100-300	Extrapolation	680	625	370
350-550	Extrapolation	771	716	357
600-800	Extrapolation	832	786	318
800-1000	Trained	1006	916	417
1000-1200	Extrapolation	1149	1105	441
1200-1400	Extrapolation	1180	1112	443

(b) Varied Testing - vx

Band	Band Type	Mean	Median	Sd
100-300	Extrapolation	667	554	403
350-550	Extrapolation	770	688	383
600-800	Extrapolation	869	814	358
800-1000	Trained	953	928	359
1000-1200	Trained	1072	1066	388
1200-1400	Trained	1144	1093	426

Band	Band Type	Mean	Median	Sd
100-300	Extrapolation	684	634	406
350-550	Extrapolation	729	679	350
600-800	Trained	776	721	318
800-1000	Extrapolation	941	883	387
1000-1200	Extrapolation	1014	956	403
1200-1400	Extrapolation	1072	1014	442

Band	Band Type	Mean	Median	Sd
100-300	Trained	392	270	343
350-550	Trained	540	442	343
600-800	Trained	642	588	315
800-1000	Extrapolation	943	899	394
1000-1200	Extrapolation	1081	1048	415
1200-1400	Extrapolation	1185	1129	500

Table 13: Experiment 3. Bayesian Mixed Model Predicting Vx as a function of condition (Constant vs. Varied) and Velocity Band

Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	607.67	536.02	679.87	1
conditVaried	-167.76	-277.14	-64.08	1
Band	0.44	0.35	0.52	1
condit*Band	0.18	0.06	0.31	1

See Table 13 for the full model results.

Slope estimates for experiment 3 suggest that participants were capable of distinguishing between velocity bands even when provided only ordinal feedback during training ($\beta = 0.44$, 95% CrI [0.35, 0.52]). Unlike the previous two experiments, the posterior distribution for the interaction between condition and band was consistently positive, suggestive of superior discrimination for the varied participants $\beta = 0.18$, 95% CrI [0.06, 0.31].
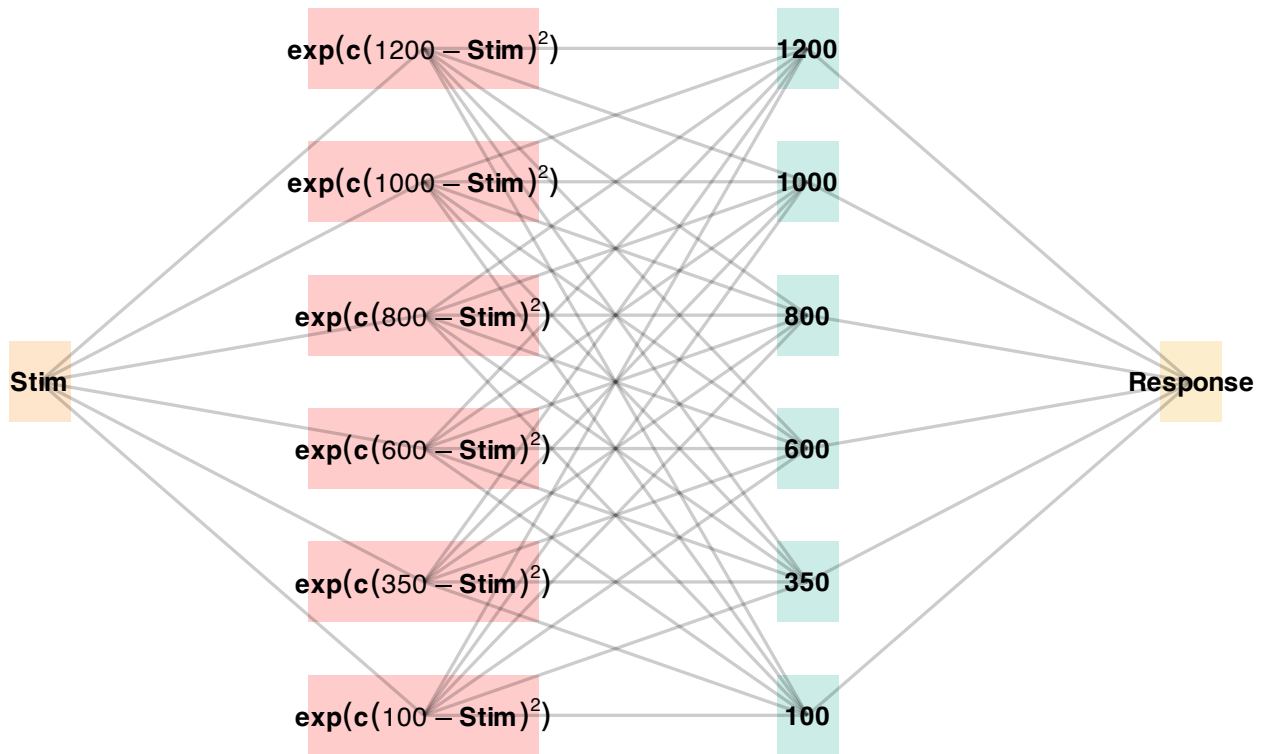

Figure 15: The basic structure of the ALM model.

Table 14: ALM & EXAM Equations

	ALM Response Generation	
Input Activation	$a_i(X) = \frac{e^{-c(X-X_i)^2}}{\sum_{k=1}^{M} e^{-c(X-X_k)^2}}$	Input nodes activate as a function of Gaussian similarity to stimulus
Output Activation	$O_j(X) = \sum_{k=1}^{M} w_{ji} \cdot a_i(X)$	Output unit O_j activation is the weighted sum of input activations and association weights
Output Probability	$P[Y_j|X] = \frac{O_j(X)}{\sum_{k=1}^{M} O_k(X)}$	The response, Y_j probabilites computed via Luce's choice rule
Mean Output	$m(X) = \sum_{j=1}^{L} Y_j \cdot \frac{O_j(x)}{\sum_{k=1}^{M} O_k(X)}$	Weighted average of probabilities determines response to X
	ALM Learning	
Feedback	$f_j(Z) = e^{-c(Z-Y_j)^2}$	feedback signal Z computed as similarity between ideal response and observed response
magnitude of error	$\Delta_{ji} = (f_j(Z) - o_j(X))a_i(X)$	Delta rule to update weights.
Update Weights	$w_{ji}^{new} = w_{ji} + \eta\Delta_{ji}$	Updates scaled by learning rate parameter η.
	EXAM Extrapolation	
Instance Retrieval	$P[X_i|X] = \frac{a_i(X)}{\sum_{k=1}^{M} a_k(X)}$	Novel test stimulus X activates input nodes X_i
Slope Computation	$S = \frac{m(X_1)-m(X_2)}{X_1-X_2}$	Slope value, S computed from nearest training instances
Response	$E[Y|X_i] = m(X_i) + S \cdot [X - X_i]$	ALM response $m(X_i)$ adjusted by slope.

Modeling

In project 1, I applied model-based techniques to quantify and control for the similarity between training and testing experience, which in turn enabled us to account for the difference between varied and constant training via an extended version of a similarity based generalization model. In project 2, I will go a step further, implementing a full process model capable of both 1) producing novel responses and 2) modeling behavior in both the learning and testing stages of the experiment. For this purpose, we will apply the associative learning model (ALM) and the EXAM model of function learning (DeLosh 1997). ALM is a simple connectionist learning model which closely resembles Kruschke's ALCOVE model (Kruscke 1992), with modifications to allow for the generation of continuous responses.

ALM & Exam Description

DeLosh et al. (1997) introduced the associative learning model (ALM), a connectionist model within the popular class of radial-basis networks. ALM was inspired by, and closely resembles Kruschke's influential ALCOVE model of categorization (Kruschke, 1992).

ALM is a localist neural network model, with each input node corresponding to a particular stimulus, and each output node corresponding to a particular response value. The units in the input layer activate as a function of their Gaussian similarity to the input stimulus. So, for example, an input stimulus of value 55 would induce maximal activation of the input unit tuned to 55. Depending on thevalue of the generalization parameter, the nearby units (e.g. 54 and 56; 53 and 57) may also activate to some degree. ALM is structured with input and output nodes that correspond to regions of the stimulus space, and response space, respectively. The units in the input layer activate as a function of their similarity to a presented stimulus. As was the case with the exemplar-based models, similarity in ALM is exponentially decaying function of distance. The input layer is fully connected to the output layer, and the activation for any particular output node is simply the weighted sum of the connection weights between that node and the input activations. The network then produces a response by taking the weighted average of the output units (recall that each output unit has a value corresponding to a particular response). During training, the network receives feedback which activates each output unit as a function of its distance from the ideal level of activation necessary to produce the correct response. The connection weights between input and output units are then updated via the standard delta learning rule, where the magnitude of weight changes are controlled by a learning rate parameter.

See Table 14 for a full specification of the equations that define ALM and EXAM.

Model Fitting and Comparison

Following the procedure used by Mcdaniel et al. (2009), we will assess the ability of both ALM and EXAM to account for the empirical data when fitting the models to 1) only the training data, and 2) both training and testing data. Models were fit to the aggregated participant data by minimizing the root-mean squared deviation (RMSE). Because ALM has been shown to do poorly at accounting for human patterns extrapolation (DeLosh et al., 1997), we will also generate predictions from the EXAM model for the testing stage. EXAM which operates identically to ALM during training, but includes a linear extrapolation mechanism for generating novel responses during testing.

For the hybrid model, predictions are computed by first generating separate predictions from ALM and EXAM, and then combining them using the following equation: $\hat{y} = (1 - w) \cdot alm.pred + w \cdot exam.pred$. For the grid search, the weight parameter is varied from 0 to 1, and the resulting RMSE is recorded.

Each model was fit to the data in 3 different ways. 1) To just the testing data, 2) Both the training and testing data, 3) Only the training data. In all cases, the model only updates its weights during the training phase, and the weights are frozen during the testing phase. In all cases, only the ALM model generates predictions during the training phase. For the testing phase, all 3 models are used to generate predictions.

Table 15: Fit Parameters and Model RMSE. The Test_RMSE column is the main performance indicator of interest, and represents the RMSE for just the testing data. The Fit_Method column indicates the data used to fit the model. The w parameter determines the balance between the ALM and EXAM response generation processes, and is only included for the hybrid model. A weight of .5 would indicate equal contribution from both models. w values approaching 1 indicate stronger weight for EXAM.

| Model	Fit Method	Constant			Varied		
		c	lr	w	c	lr	w
ALM	Test Only	0.000	0.100		0.134	2.030	
ALM	Test & Train	0.047	0.080		0.067	0.100	
ALM	Train Only	0.060	0.100		0.047	0.080	
EXAM	Test Only	0.007	1.327		0.409	1.910	
EXAM	Test & Train	0.081	0.161		0.074	0.100	
EXAM	Train Only	0.060	0.100		0.047	0.080	
Hybrid	Test Only	0.008	1.580	1	0.395	2.017	0.64
Hybrid	Test & Train	0.067	0.134	1	0.134	2.017	0.79
Hybrid	Train Only	0.042	0.067	0	0.042	0.067	0.00

Testing Observations vs. Predictions

Table 16: Model Perforamnce - averaged over all X values/Bands. ME=Mean Average Error, RMSE = Root mean squared error.
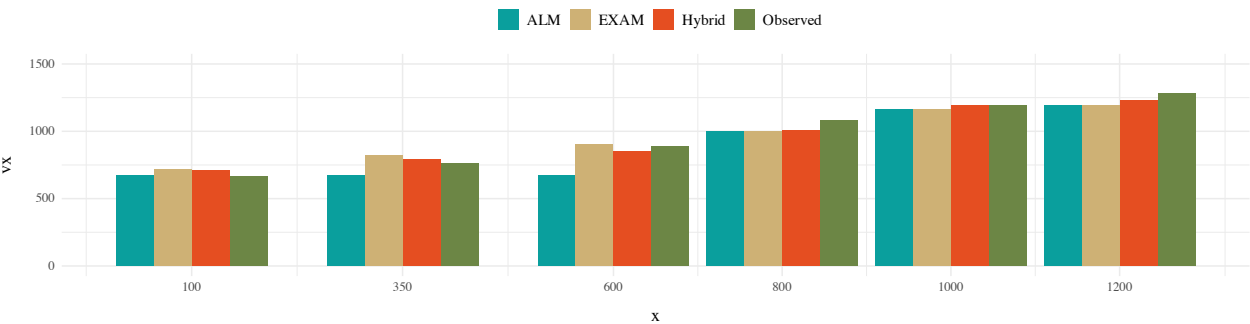
Fit_Method	Model	Constant		Varied	
		ME	RMSE	ME	RMSE
Test Only	ALM	223.8	348.0	56.3	95.4
	EXAM	-59.2	127.5	-6.0	45.9
	Hybrid	-58.2	127.4	-3.0	33.8
Test & Train	ALM	193.2	328.7	82.3	106.6
	EXAM	-28.8	132.1	13.2	60.2
	Hybrid	-16.7	136.7	16.7	46.5
Train Only	ALM	194.5	329.2	86.3	109.1
	EXAM	75.3	199.9	17.5	65.4
	Hybrid	197.5	330.4	88.3	110.3

Varied Testing Predictions
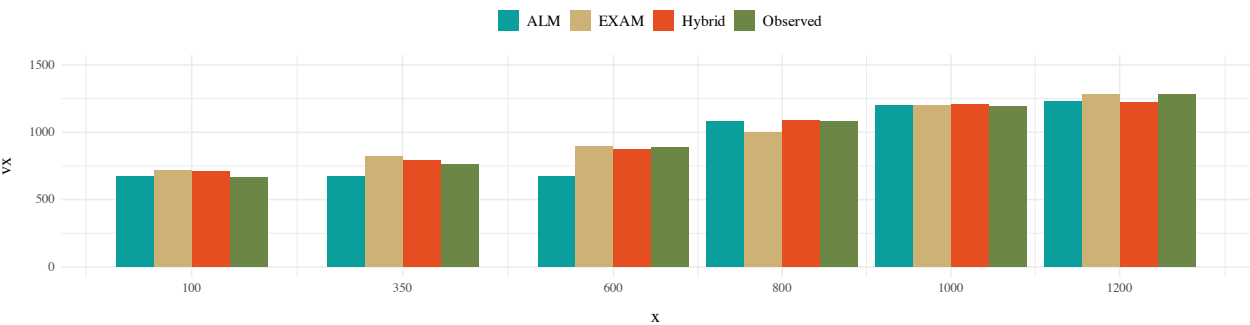
Fit to Test Only

Fit to Test and Train
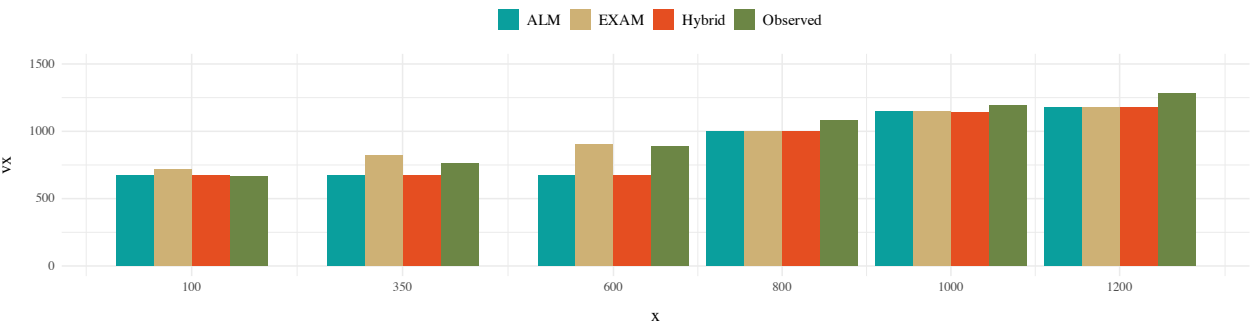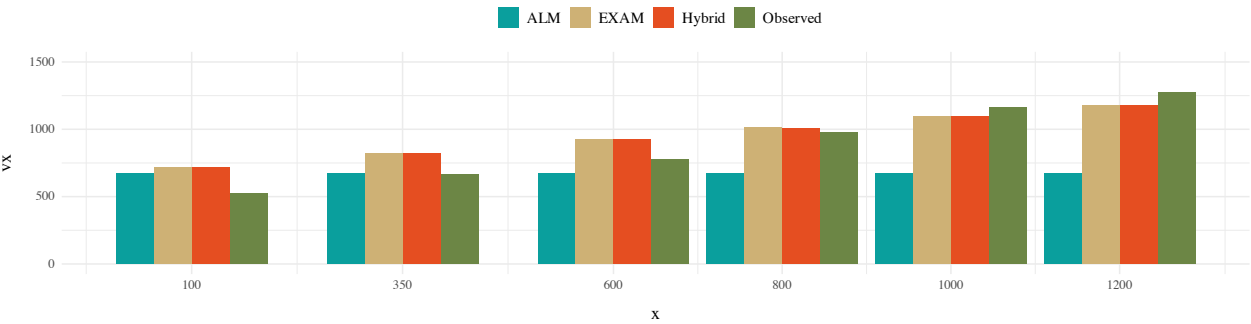

Fit to Train Only

Figure 16: Varied Group - Mean Model predictions vs. observations

Table 17: Varied group - mean model predictions vs. observations. Extrapolation Bands are bolded. For each Modelling fitting and band combination, the model with the smallest residual is highlighted. Only the lower bound of each velocity band is shown (bands are all 200 units).

Fit Method	X	Observed	ALM Predicted	ALM Residual	EXAM Predicted	EXAM Residual	Hybrid Predicted	Hybrid Residual
Test Only	**100**	663	675	-12	716	-53	708	-45
Test Only	**350**	764	675	89	817	-53	792	-28
Test Only	**600**	884	675	209	895	-11	875	9
Test Only	800	1,083	1,078	5	1,000	83	1,091	-8
Test Only	1,000	1,196	1,202	-6	1,199	-3	1,204	-8
Test Only	1,200	1,283	1,230	53	1,282	1	1,221	62
Test & Train	**100**	663	675	-12	716	-53	707	-44
Test & Train	**350**	764	675	89	817	-53	788	-24
Test & Train	**600**	884	675	209	902	-18	851	33
Test & Train	800	1,083	1,000	83	1,000	83	1,004	79
Test & Train	1,000	1,196	1,163	33	1,165	31	1,196	0
Test & Train	1,200	1,283	1,191	92	1,194	89	1,227	56
Train Only	**100**	663	675	-12	716	-53	675	-12
Train Only	**350**	764	675	89	817	-53	675	89
Train Only	**600**	884	675	209	905	-21	675	209
Train Only	800	1,083	1,000	83	1,000	83	999	84
Train Only	1,000	1,196	1,150	46	1,150	46	1,143	53
Train Only	1,200	1,283	1,180	103	1,180	103	1,176	107

Constant Testing Predictions

Fit to Test Only

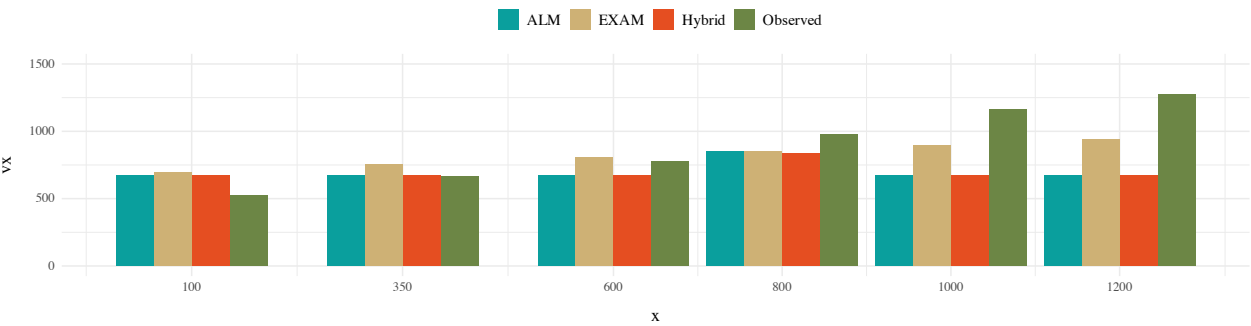

Fit to Test and Train
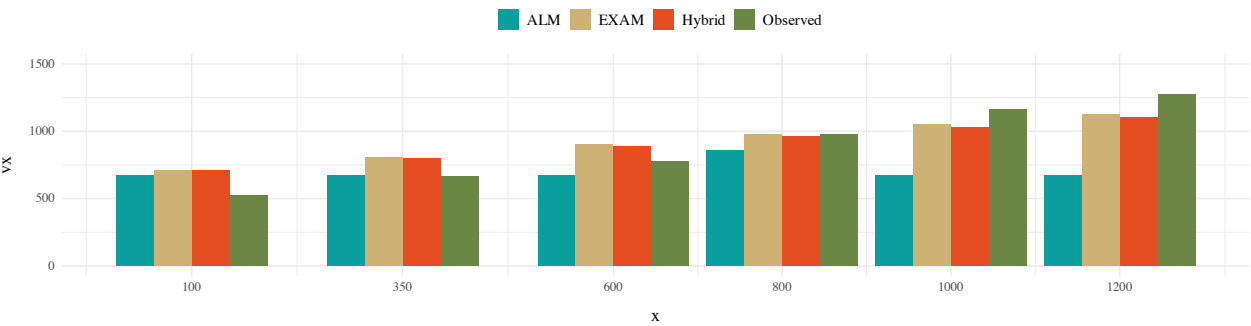
Fit to Train Only

Figure 17: Constant Group - Mean Model predictions vs. observations

Table 18: Constant group - mean model predictions vs. observations. The X values of Extrapolation Bands are bolded. For each Modelling fitting and band combination, the model with the smallest residual is highlighted. Only the lower bound of each velocity band is shown (bands are all 200 units).

Fit Method	X	Observed	ALM Predicted	ALM Residual	EXAM Predicted	EXAM Residual	Hybrid Predicted	Hybrid Residual
Test Only	**100**	527	675	-148	717	-190	717	-190
Test Only	**350**	666	675	-9	822	-156	821	-155
Test Only	**600**	780	675	105	927	-147	926	-146
Test Only	800	980	675	305	1,010	-30	1,009	-29
Test Only	**1,000**	1,163	675	488	1,094	69	1,093	70
Test Only	**1,200**	1,277	675	602	1,178	99	1,176	101
Test & Train	**100**	527	675	-148	712	-185	711	-184
Test & Train	**350**	666	675	-9	806	-140	800	-134
Test & Train	**600**	780	675	105	900	-120	889	-109
Test & Train	800	980	859	121	975	5	960	20
Test & Train	**1,000**	1,163	675	488	1,049	114	1,031	132
Test & Train	**1,200**	1,277	675	602	1,124	153	1,102	175
Train Only	**100**	527	675	-148	697	-170	675	-148
Train Only	**350**	666	675	-9	752	-86	675	-9
Train Only	**600**	780	675	105	807	-27	675	105
Train Only	800	980	851	129	851	129	833	147
Train Only	**1,000**	1,163	675	488	895	268	675	488
Train Only	**1,200**	1,277	675	602	939	338	675	602

EXAM fit learning curves

References

Bott, L., & Heit, E. (2004). Nonmonotonic Extrapolation in Function Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 38–50. https://doi.org/10.1037/0278-7393.30.1.38

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance, 11*(1), 1–27. https://doi.org/10.1016/0030-5073(74)90002-6

Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Carroll, J. D. (1963). Functional Learning: The Learning of Continuous Functional Mappings Relating Stimulus and Response Continua. *ETS Research Bulletin Series*, *1963*(2), i–144. https://doi.org/10.1002/j.2333-8504.1963.tb00958.x

Ciccione, L., & Dehaene, S. (2021). Can humans perform mental regression on a graph? Accuracy and bias in the perception of scatterplots. *Cognitive Psychology*, *128*, 101406. https://doi.org/10.1016/j.cogpsych.2021.101406

Courrieu, P. (2012). Quick approximation of bivariate functions. *British Journal of Mathematical and Statistical Psychology*, *65*(1), 89–121. https://doi.org/10.1111/j.2044-8317.2011.02016.x

DeLosh, E. L., McDaniel, M. A., & Busemeyer, J. R. (1997). Extrapolation: The Sine Qua Non for Abstraction in Function Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 19. https://doi.org/10.1037/0278-7393.23.4.968

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of Linear Experts: Knowledge Partitioning and Function Learning. *Psychological Review*, *111*(4), 1072–1099. https://doi.org/10.1037/0033-295X.111.4.1072

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 811. https://doi.org/10.1037/0278-7393.17.5.811

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of Category Learning. *Psychological Review*, *99*(1). https://doi.org/10.1037/0033-295X.99.1.22

Mcdaniel, M. A., Dimperio, E., Griego, J. A., & Busemeyer, J. R. (2009). Predicting transfer performance: A comparison of competing function learning models. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*, 173–195. https://doi.org/10.1037/a0013982

McDaniel, M. A., Fadler, C. L., & Pashler, H. (2013). Effects of spaced versus massed training in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1417–1432. https://doi.org/10.1037/a0032184

Said, N., & Fischer, H. (2021). Extrapolation accuracy underestimates rule learning: Evidence from the function-learning paradigm. *Acta Psychologica*, *218*, 103356. https://doi.org/10.1016/j.actpsy.2021.103356

Schulz, E., Quiroga, F., & Gershman, S. J. (2020). Communicating Compositional Patterns. *Open Mind*, *4*, 25–39. https://doi.org/10.1162/opmi_a_00032

Team, R. C. (2020). *R: A Language and Environment for Statistical Computing.* R: A Language and Environment for Statistical Computing.