2024-02-15
Access the code, data, and analysis at https://github.com/tegorman13/htw

HTW Model

Thomas Gorman
Indiana University
tegorman@iu.edu

Rob Goldstone
Indiana University
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).

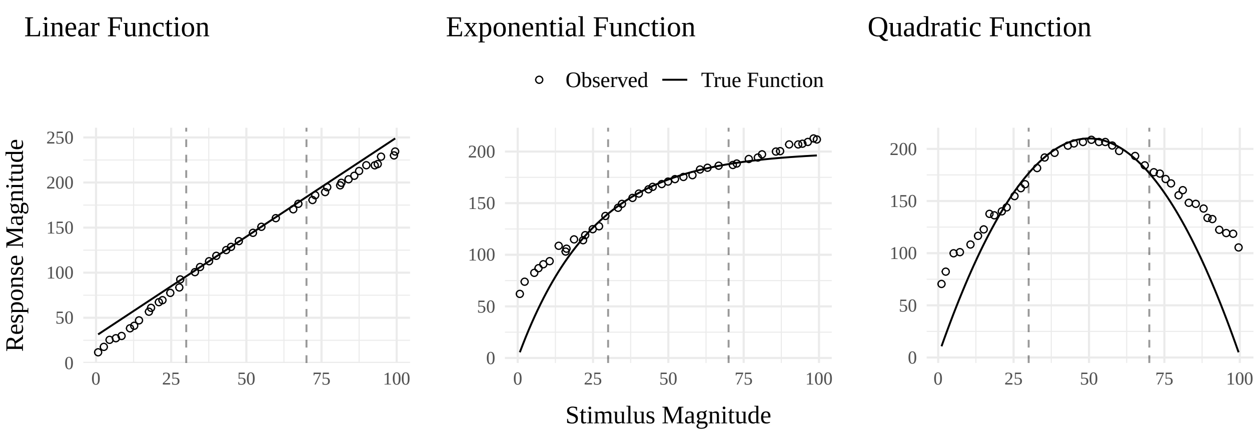


Figure 1: Generalization reproduced patterns from DeLosh et al. (1997) Figure 3. Stimulii that fall within the dashed lines are interpolations of the training examples.

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.

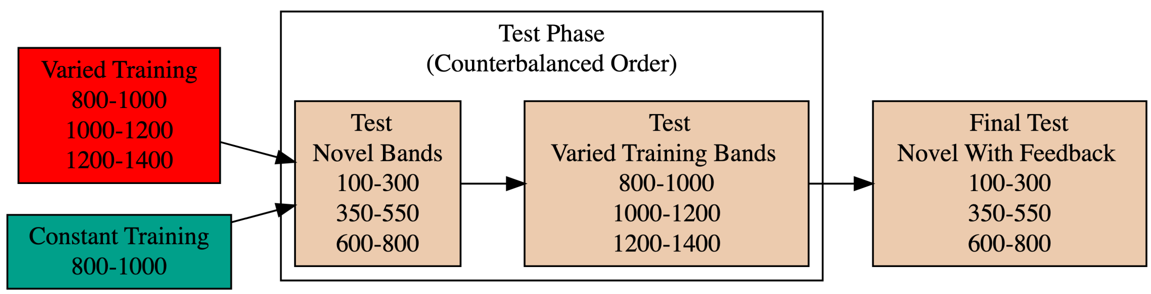


Figure 2: Experiment 1 Design. Constant and Varied participants complete different training conditions.

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 Makowski et al. (2019). 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).

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

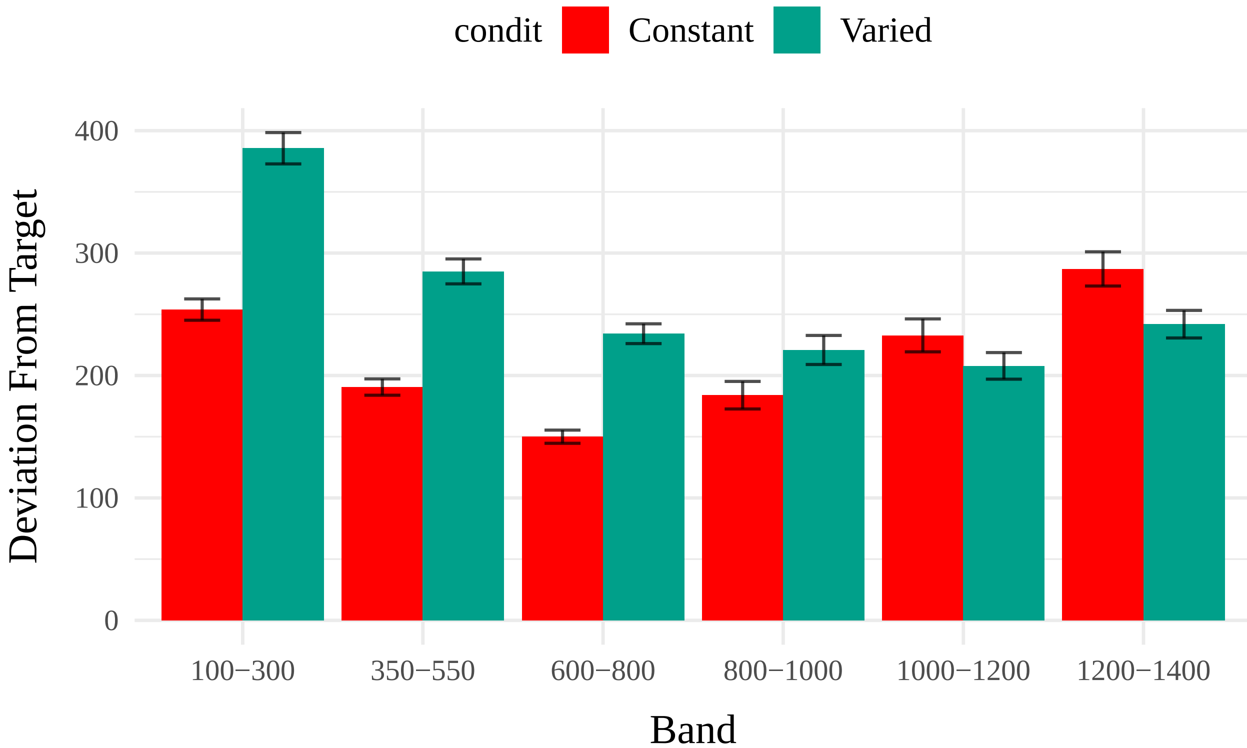


Figure 3: E1. Deviations from target band during testing without feedback stage.

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.

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

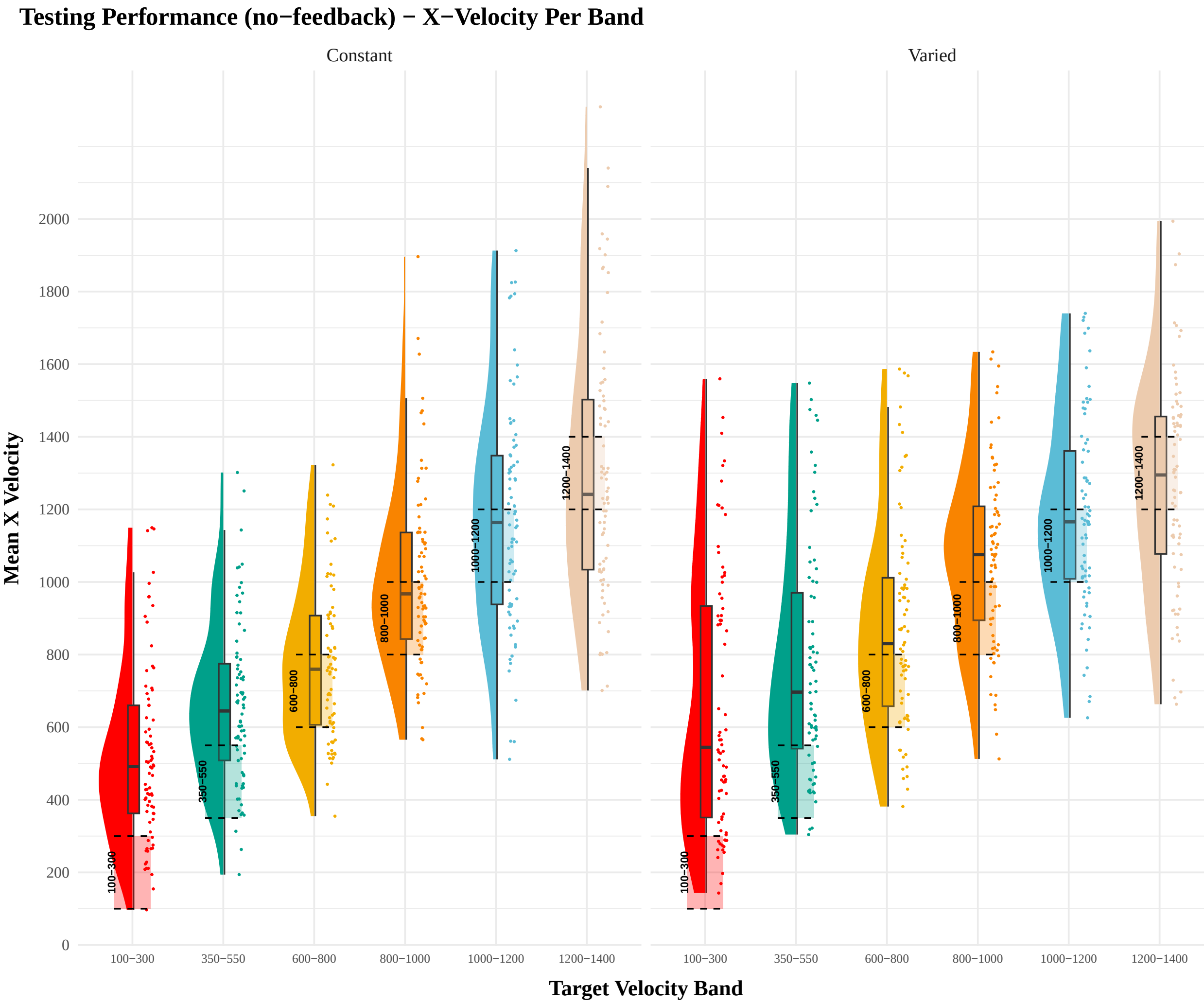


Figure 4: E1 testing x velocities. Translucent bands with dash lines indicate the correct range for each velocity band.

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

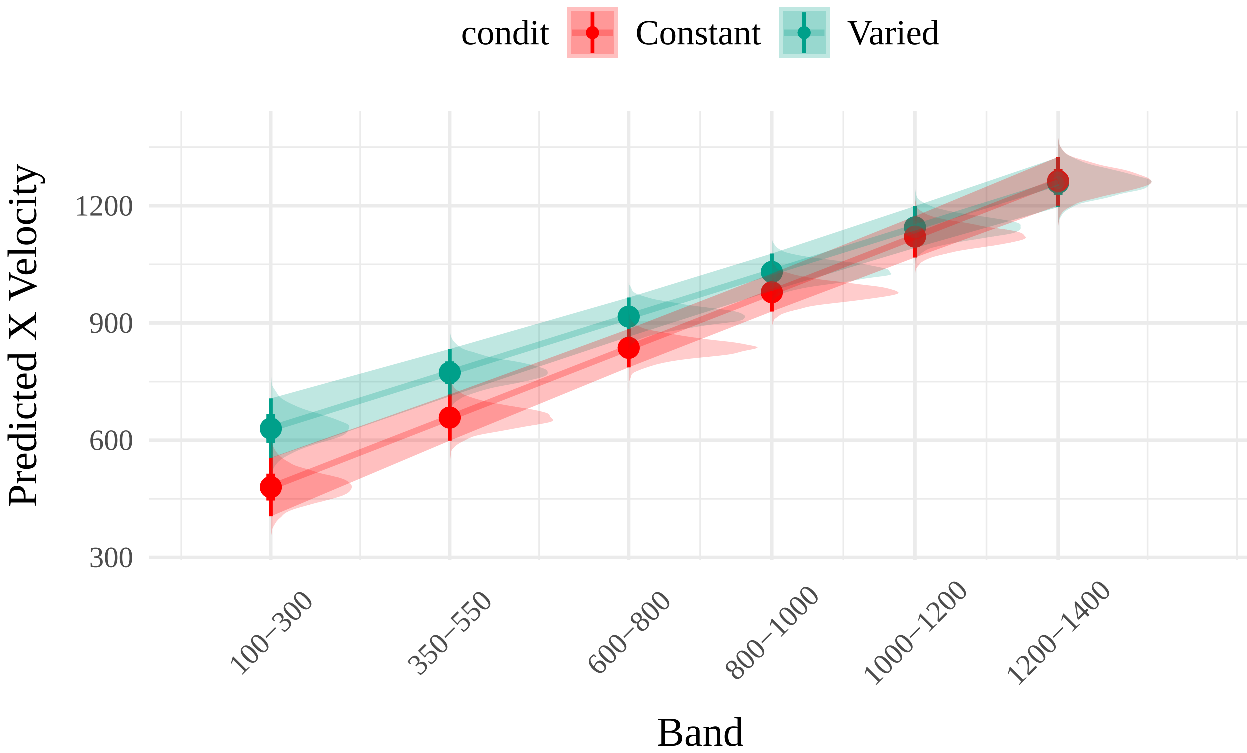
Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	408.55	327.00	490.61	1.00
conditVaried	164.05	45.50	278.85	1.00
Band	0.71	0.62	0.80	1.00
condit*Band	-0.14	-0.26	-0.01	0.98

(b) Model fit to 3 extrapolation bands

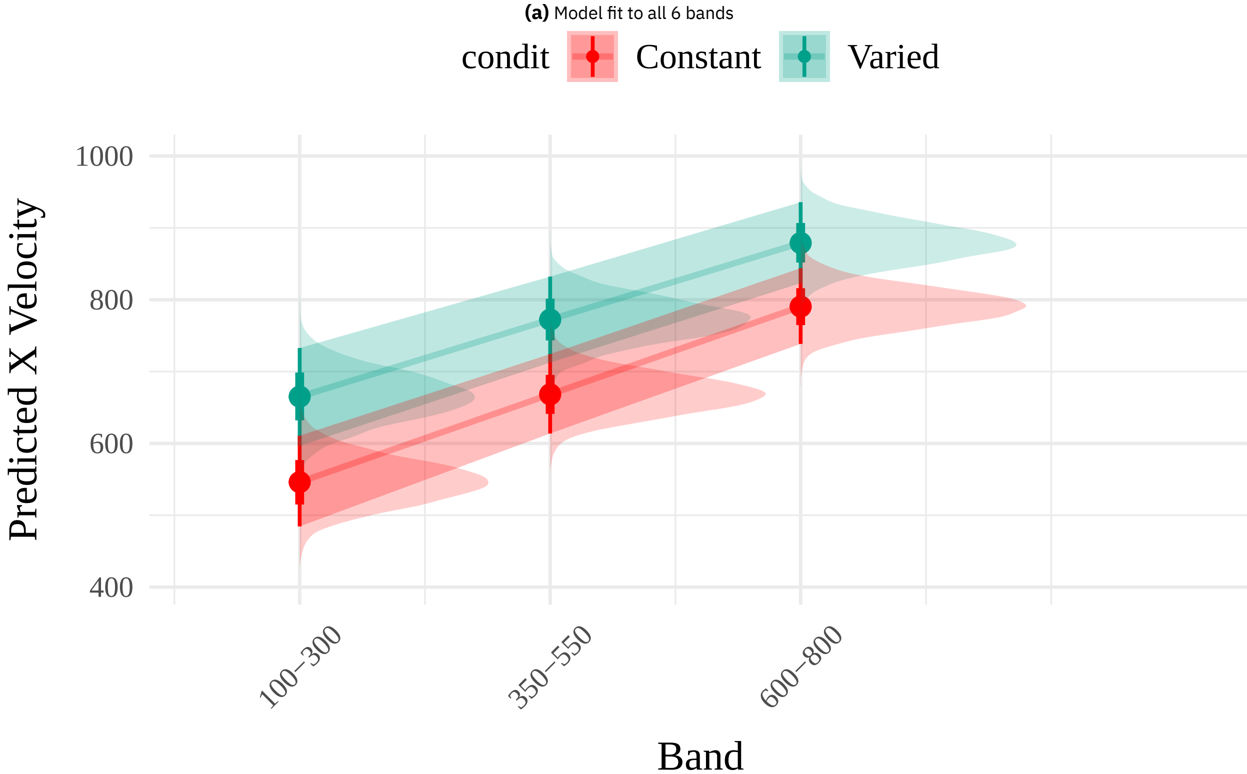
Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	497.49	431.26	566.17	1.00
conditVaried	124.79	26.61	224.75	0.99
Band	0.49	0.42	0.56	1.00
condit*Band	-0.06	-0.16	0.04	0.88

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 = 0.49, 95% CrI [0.42, 0.56]) still demonstrates a high degree of discrimination ability. However, the posterior distribution for interaction term (B = -0.06, 95% CrI [-0.16, 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

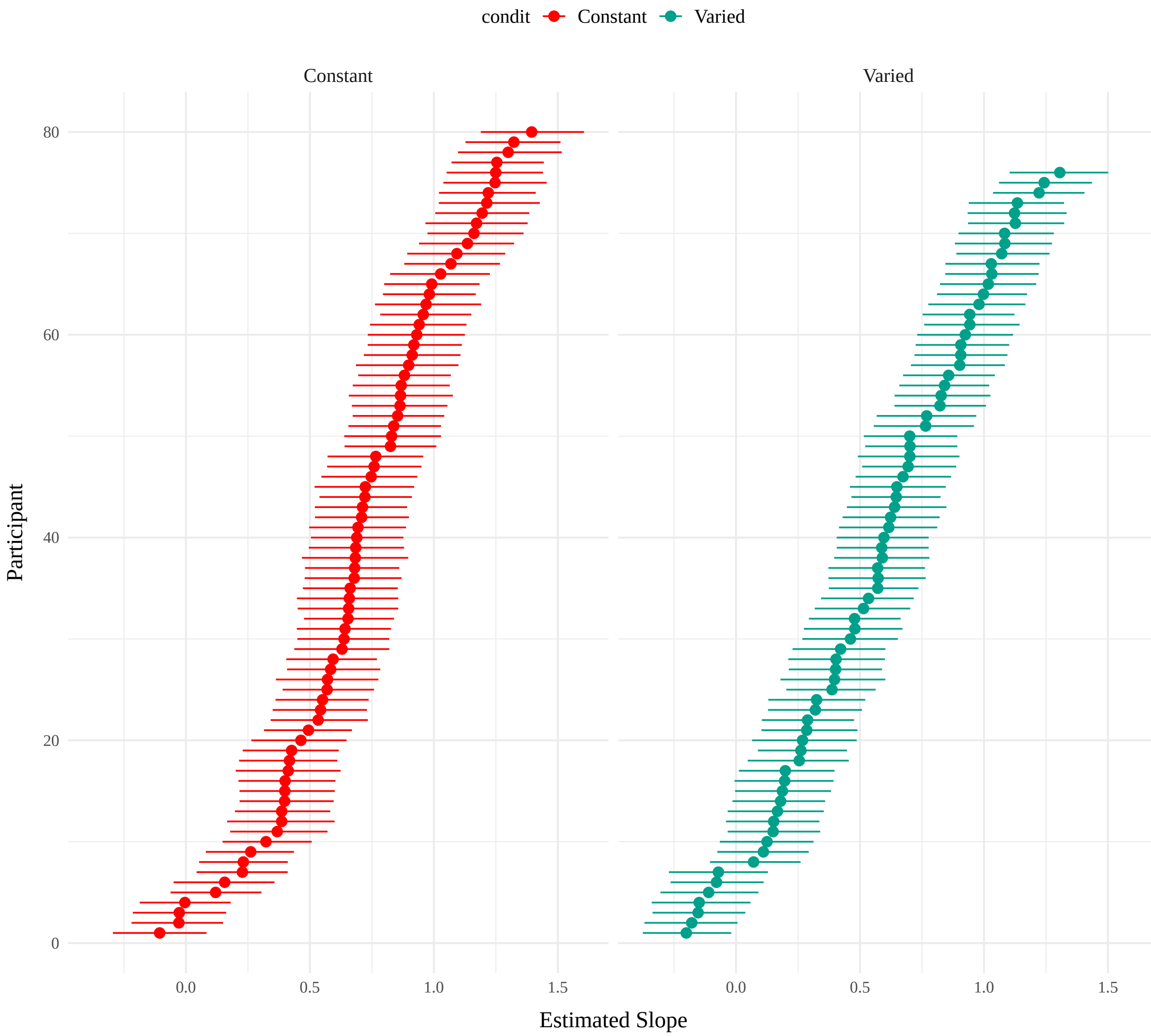


(b) Model fit to only 3 extrapolation bands

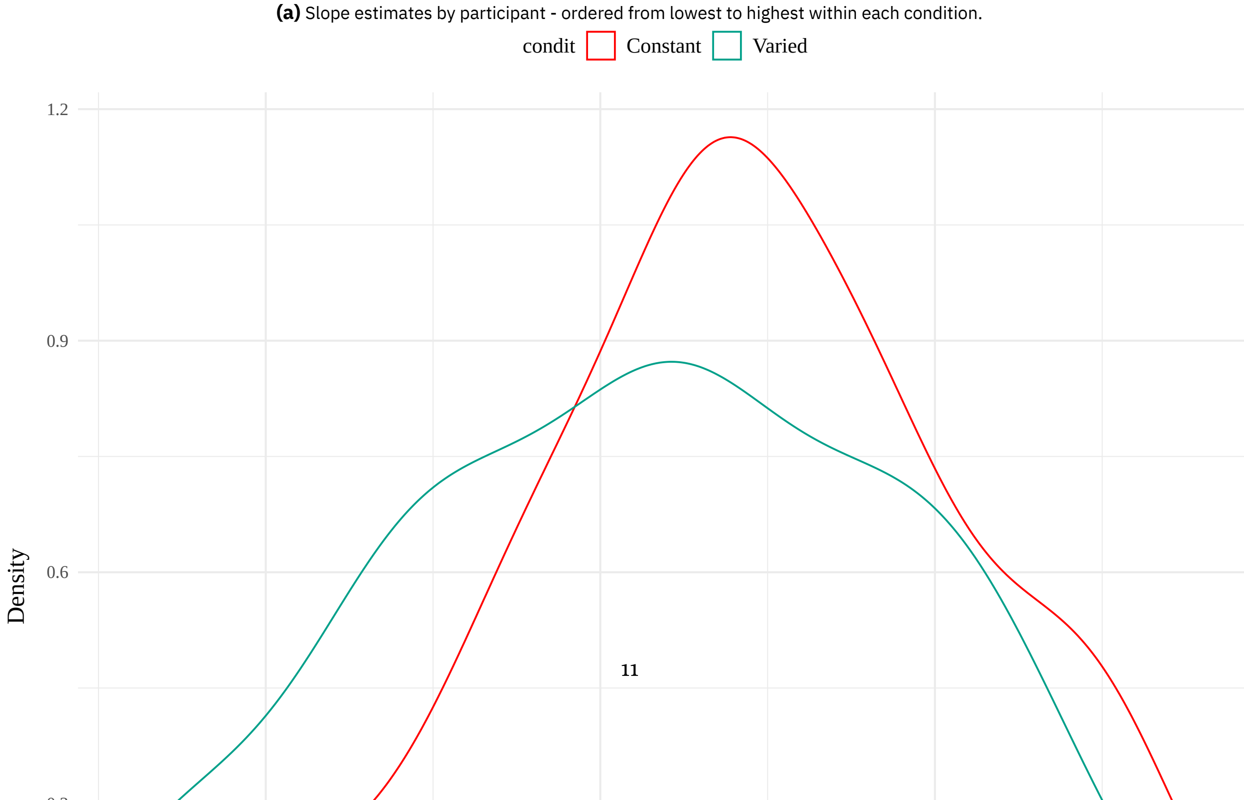
Figure 5: Conditional effect of training condition and Band. Ribbons indicate 95% HDI. The steepness of the lines serves as an indicator of how well participants discriminated between velocity bands.

Table 5: Slope coefficients by quartile, per condition

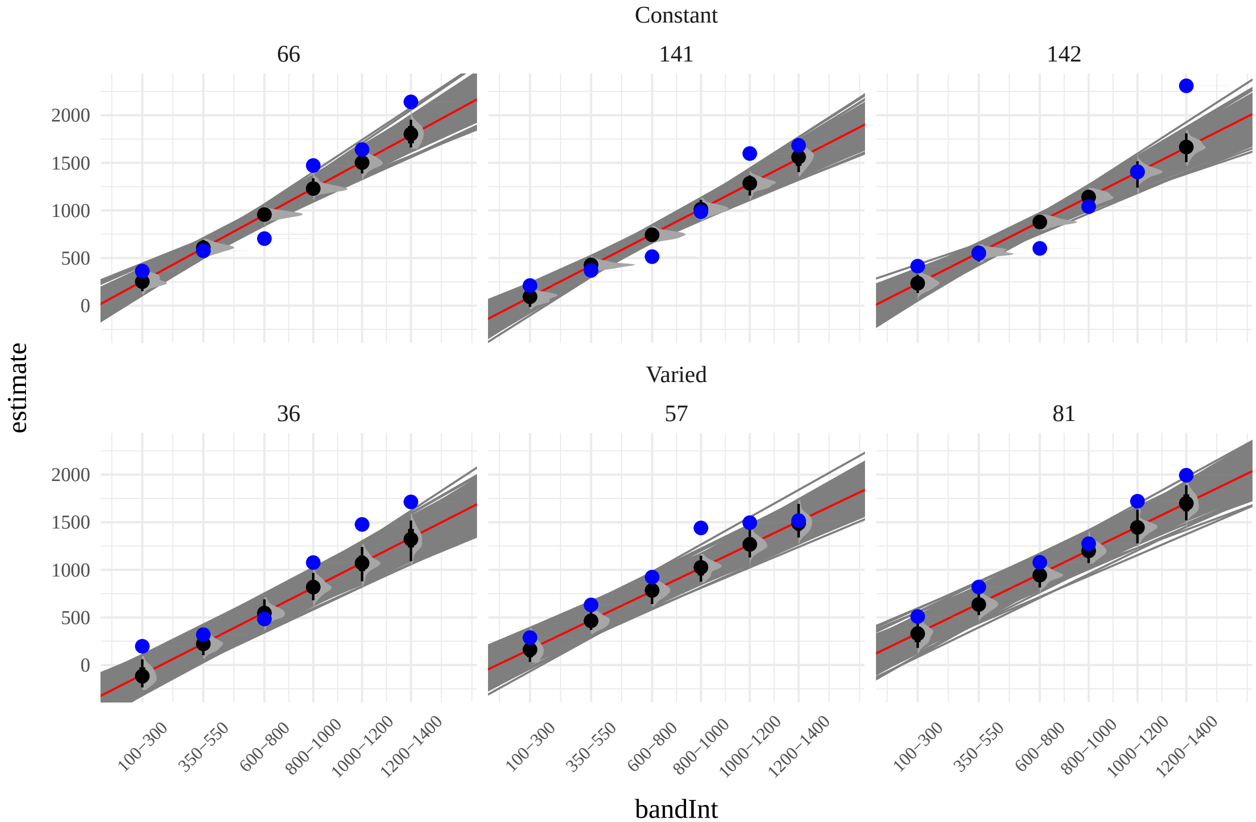
Condition	Q_0%_mean	Q_25%_mean	Q_50%_mean	Q_75%_mean	Q_100%_mean
Constant	-0.106	0.487	0.691	0.933	1.40
Varied	-0.201	0.267	0.589	0.903	1.31



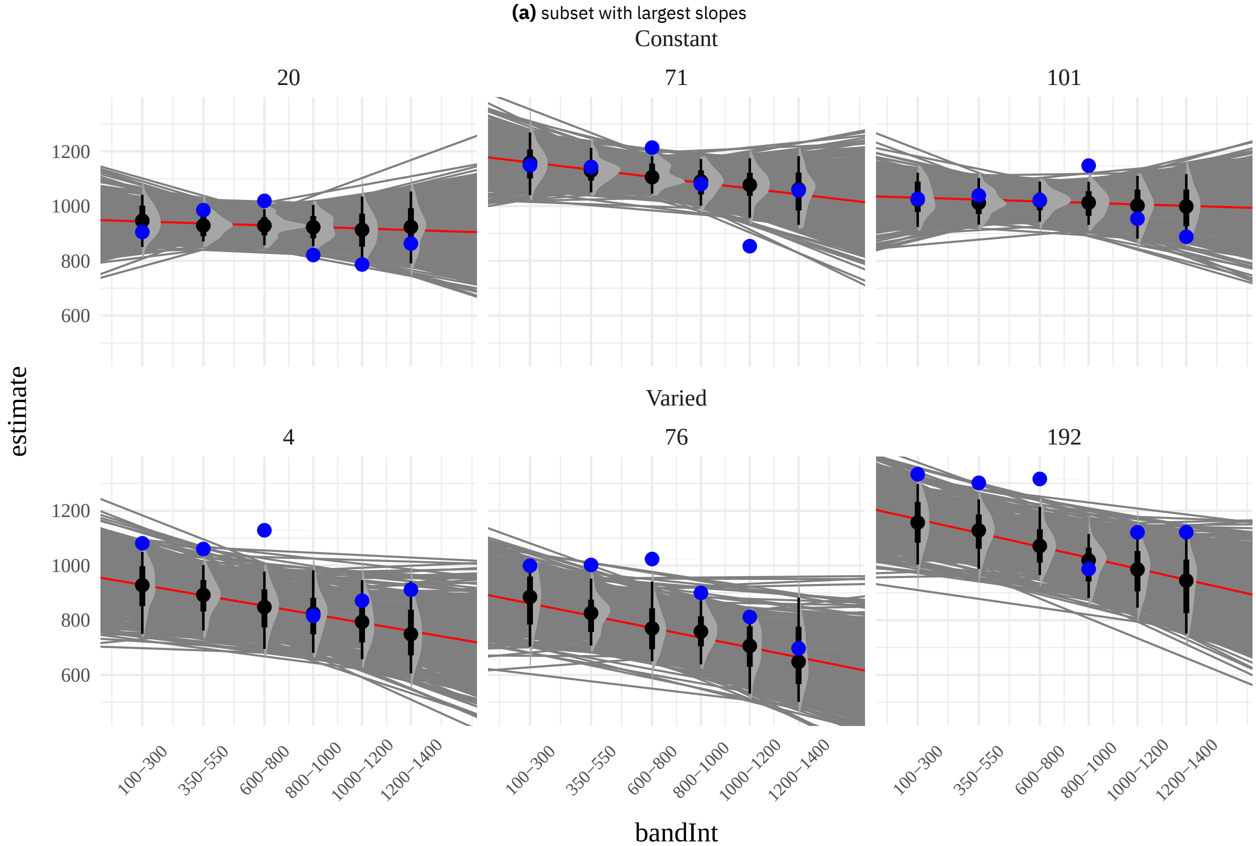
(a) Slope estimates by participant - ordered from lowest to highest within each condition.



test



(a) subset with largest slopes

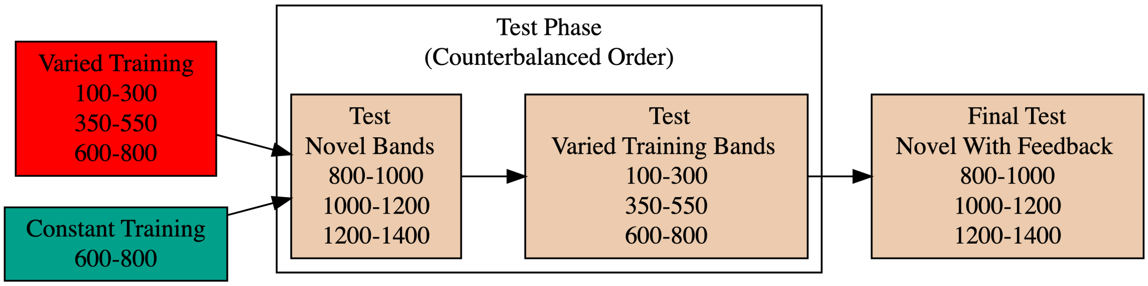


(b) subset with smallest slopes

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).



igure 8: Experiment 2 Design. Constant and Varied participants complete different training conditions. The training and testing bands are the reverse of Experiment 1.

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).

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

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

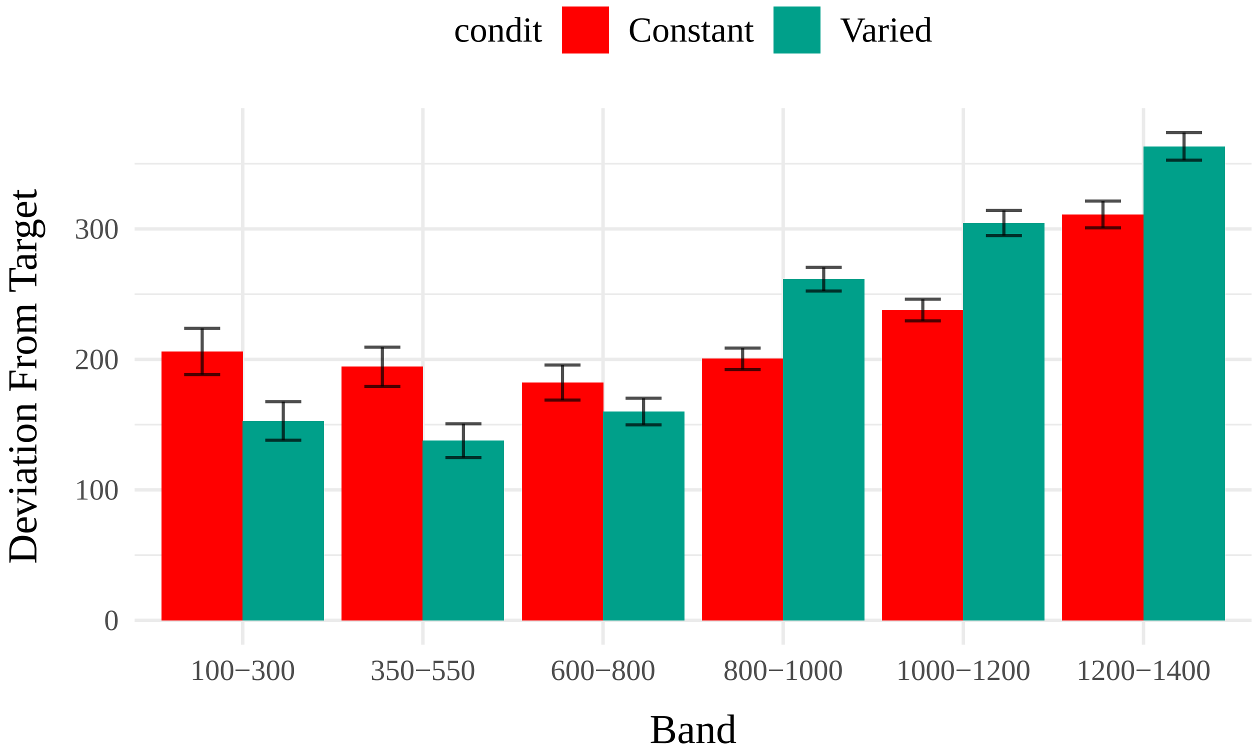


Figure 9: E2. Deviations from target band during testing without feedback stage.

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.

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

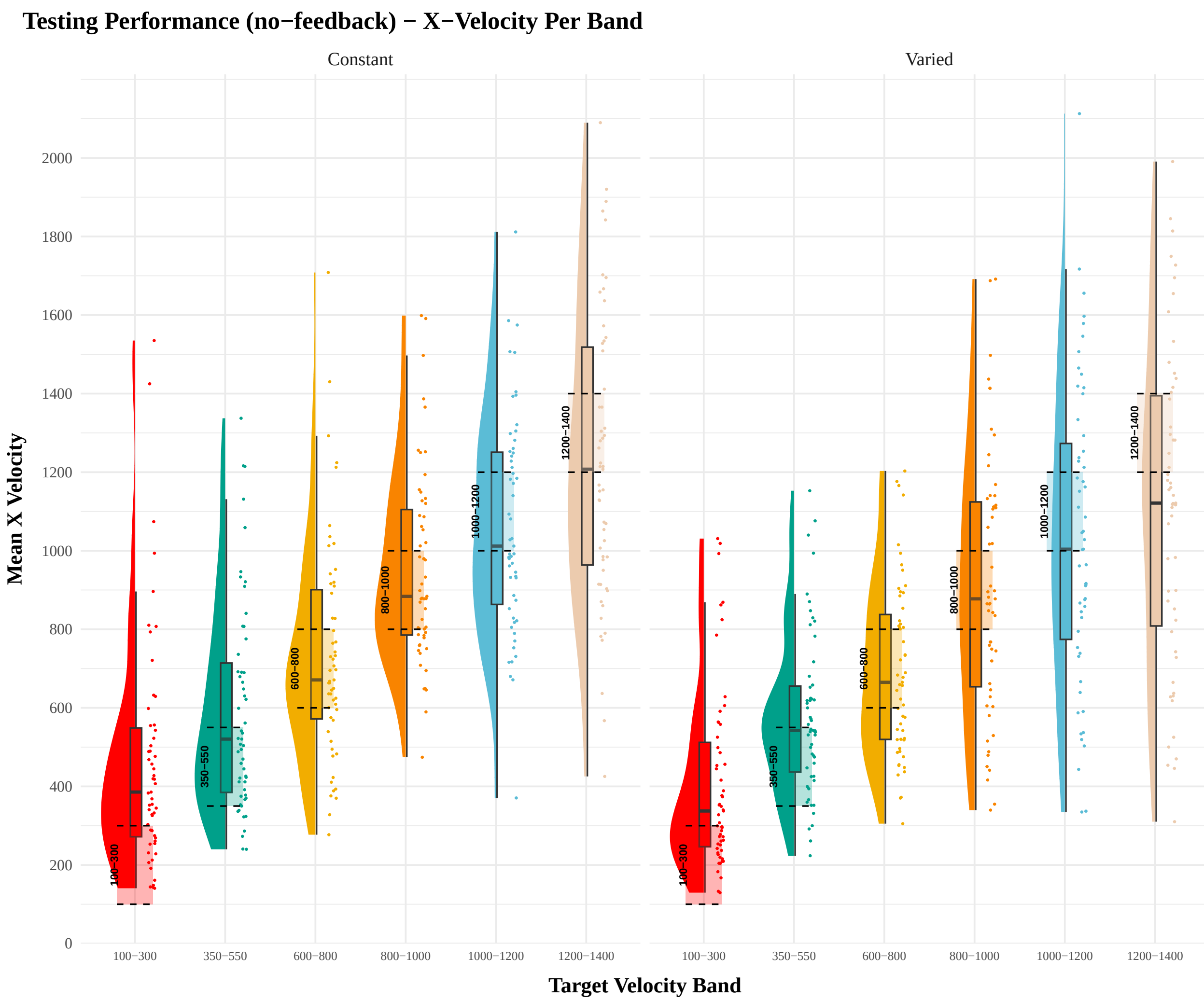


Figure 10: E2 testing x velocities. Translucent bands with dash lines indicate the correct range for each velocity band.

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

Table 9: Experiment 2. 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	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.

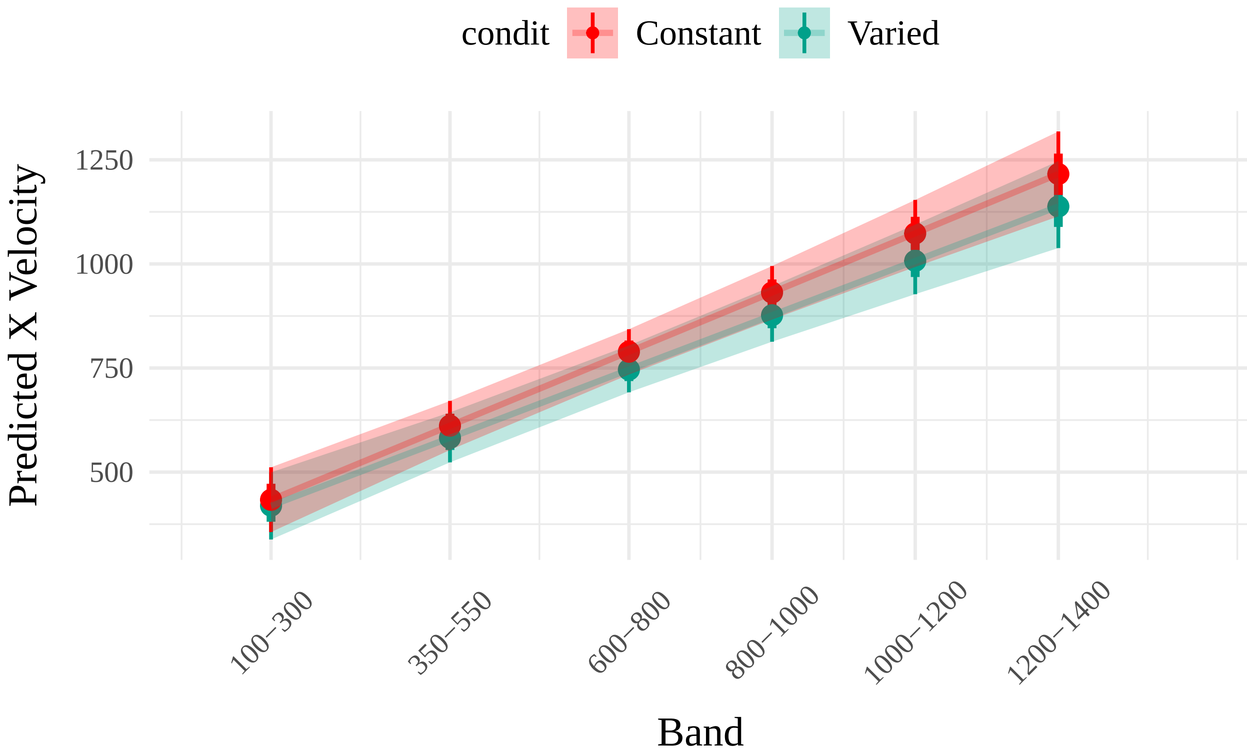


Figure 11: Conditional effect of training condition and Band. Ribbons indicate 95% HDI.

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 **?@tbl-e3-test-nf-deviation** 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).

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

Band	Band Type	Mean	Median	Sd
1200-1400	Extrapolation	277	210	262

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

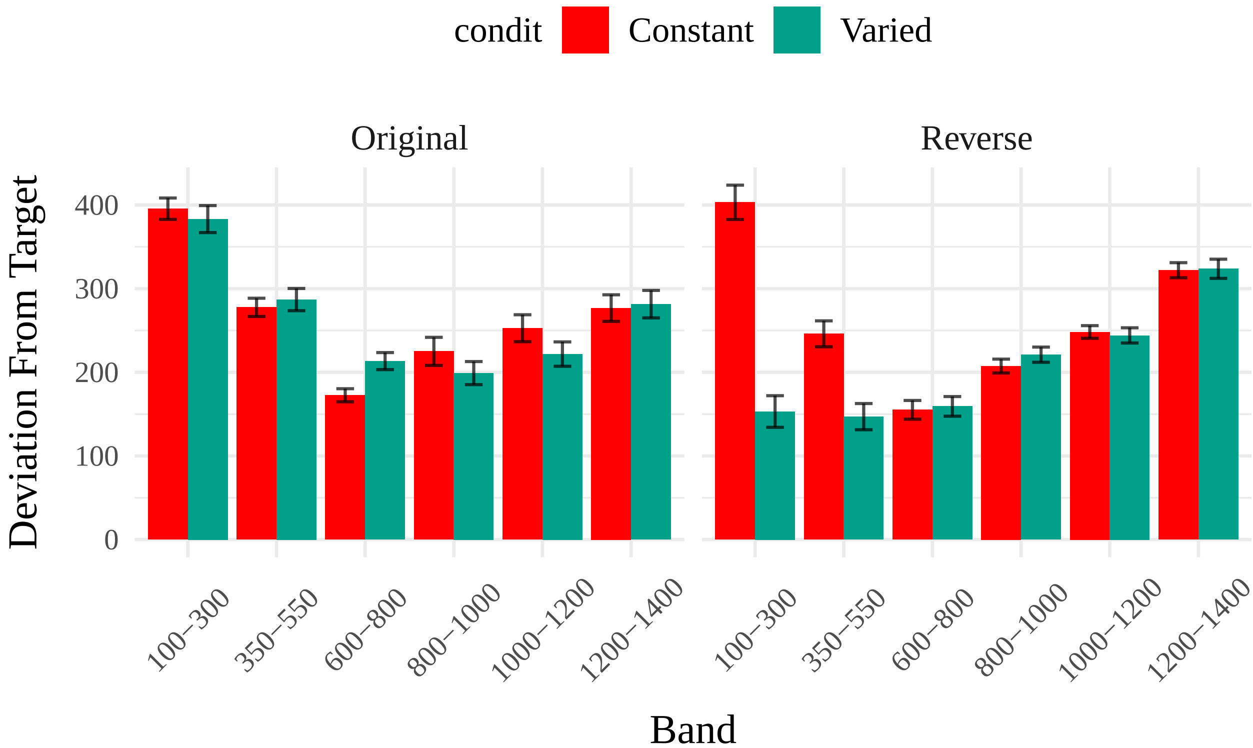


Figure 12

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.

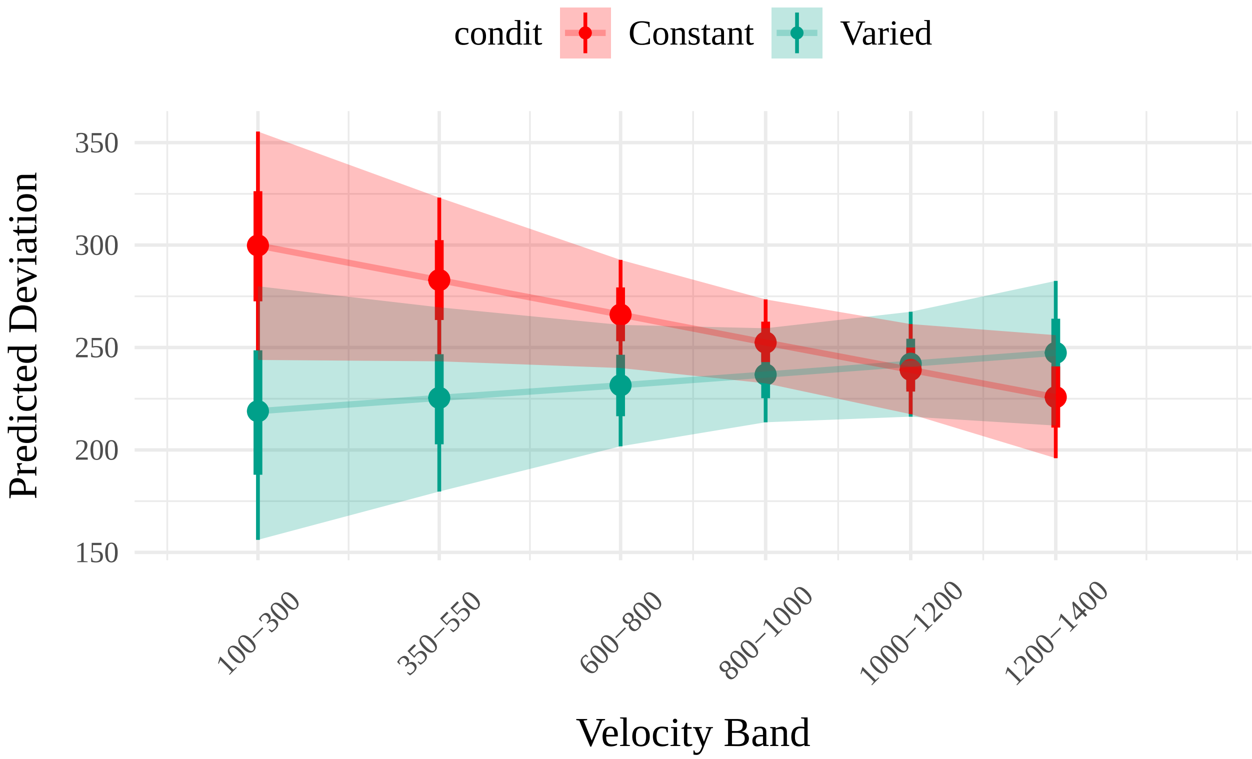


Figure 13

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). **?@tbl-e3-test-nf-vx** 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.

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

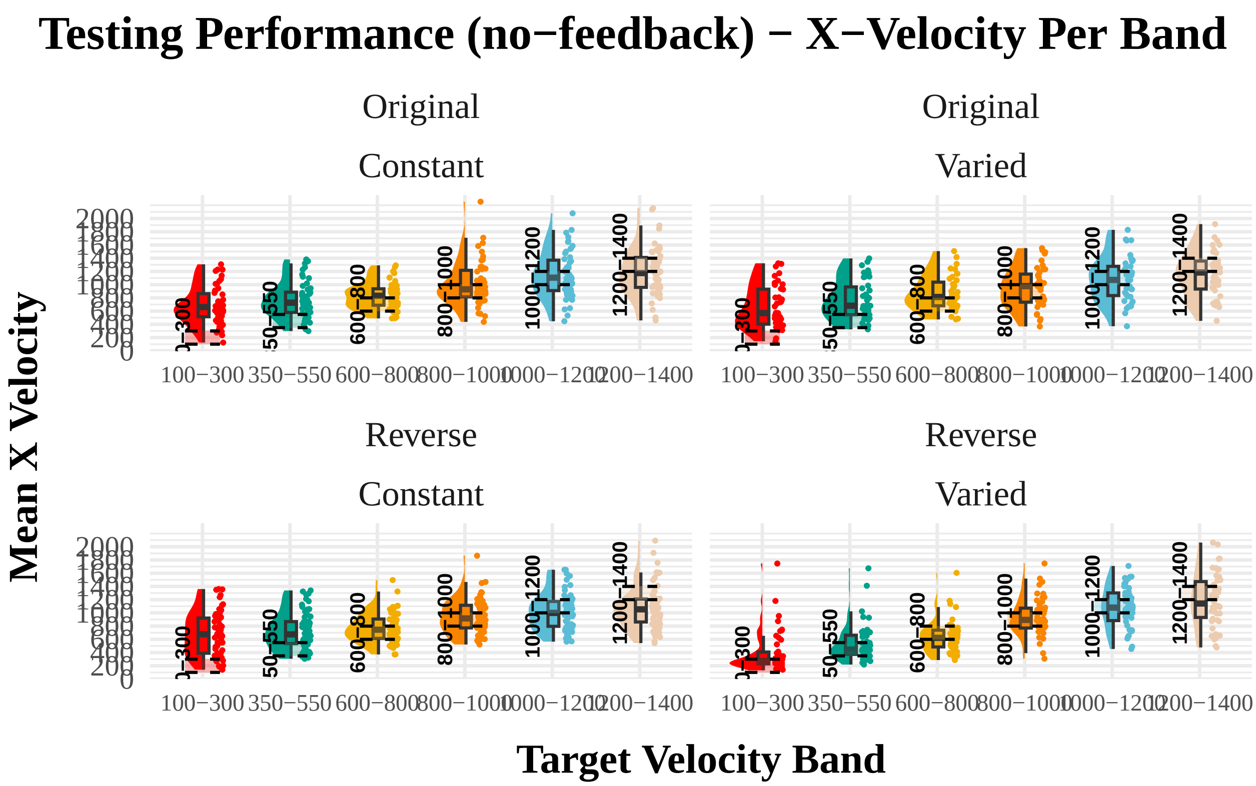


Figure 14

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

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

Band	Band Type	Mean	Median	Sd
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

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 **?@tbl-e3-bmm-vx** 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 (β = 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 β = 0.18, 95% CrI [0.06, 0.31].

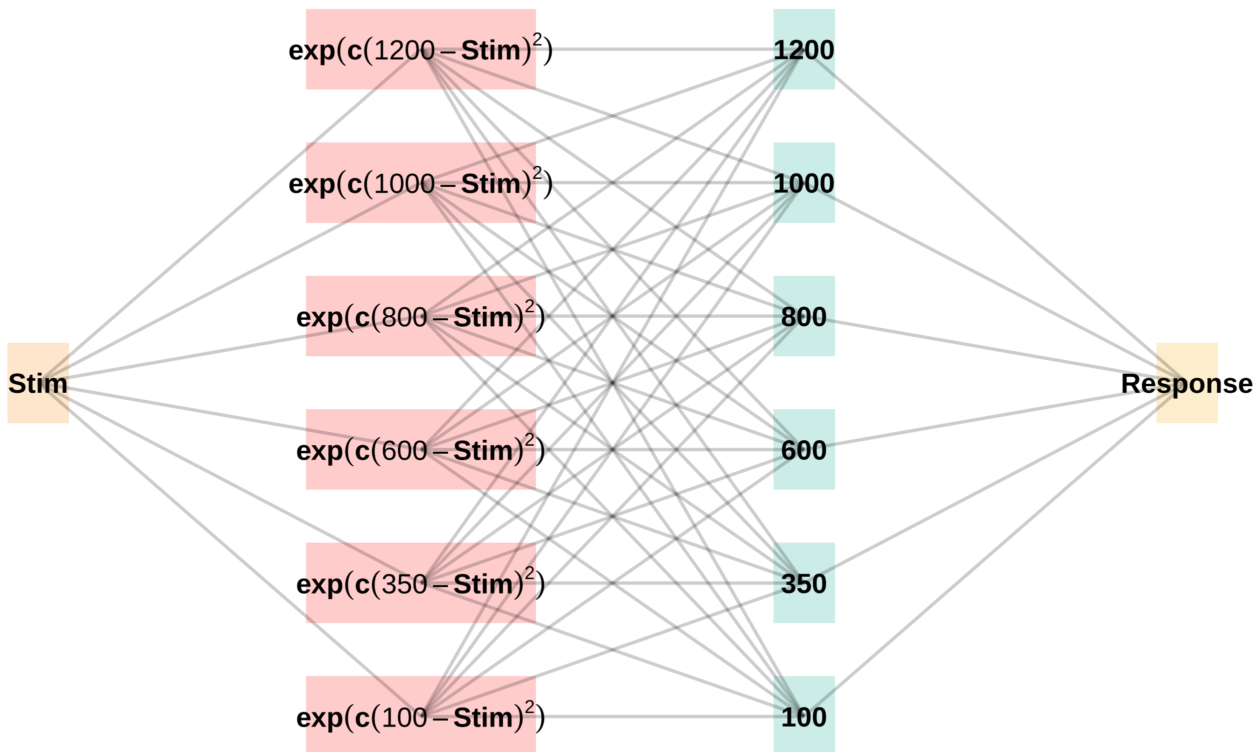


Figure 15: The basic structure of the ALM model.

Modeling Approach

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 et al., 1997). ALM is a simple connectionist learning model which closely resembles Kruschke's ALCOVE model (Kruschke, 1992), with modifications to allow for the generation of continuous responses.

ALM & Exam Description

ALM is a localist neural network model (Page, 2000), 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 the value 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. The EXAM model is an extension of ALM, with the same learning rule and representational scheme for input and output units. The primary difference is that EXAM includes a linear extrapolation mechanism for generating novel responses during testing, a modification necessary to account for human extrapolation patterns in past research Brown & Lacroix (2017). Although this extrapolation rule departs from a strictly similarity-based generalization mechanism, EXAM is distinct from pure rule-based models in that it remains constrained by the weights learned during training.

See Table 20 for a full specification of the equations that define ALM and EXAM.

Model Fitting Strategy

To fit ALM and EXAM to our participant data, we employ a similar method to Mcdaniel et al. (2009), wherein we examine the performance of each model after being fit to various subsets of the data. Each model was fit to the data in with separate procedures: 1) fit to maximize predictions of the testing data, 2) fit to maximize predictions of both the training and testing data, 3) fit to maximize predictions of the just the training data. We refer to this fitting manipulations as "Fit Method" in the tables and figures below. It should be emphasized that for all three fit methods, the ALM and EXAM models behave identically - with weights updating only during the training phase.Models to were fit separately to the data of each individual participant. The free parameters for both models are the generalization (c) and learning rate (lr) parameters. Parameter estimation was performed using approximate bayesian computation (ABC), which we describe in detail below.

Approximate Bayesian Computation

To estimate parameters, we used approximate bayesian computation (ABC), enabling us to obtain an estimate of the posterior distribution of the generalization and learning rate parameters for each individual. ABC belongs to the class of simulation-based inference methods (Cranmer et al., 2020), which have begun being used for parameter estimation in cognitive modeling relatively recently (Kangasrääsiö et al., 2019; Turner et al., 2016; Turner & Van Zandt, 2012). Although they can be applied to any model from which data can be simulated, ABC methods are most useful for complex models that lack an explicit likelihood function (e.g. many neural network and evidence accumulation models).

Table 20: 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.

The general ABC procedure is to 1) define a prior distribution over model parameters. 2) sample candidate parameter values, θ^*, from the prior. 3) Use θ^* to generate a simulated dataset, $Data_{sim}$. 4) Compute a measure of discrepancy between the simulated and observed datasets, $discrep(Data_{sim}, Data_{obs})$. 5) Accept θ^* if the discrepancy is less than the tolerance threshold, ϵ, otherwise reject θ^*. 6) Repeat until desired number of posterior samples are obtained.

Although simple in the abstract, implementations of ABC require researchers to make a number of non-trivial decisions as to i) the discrepancy function between observed and simulated data, ii) whether to compute the discrepancy between trial level data, or a summary statistic of the datasets, iii) the value of the minimum tolerance ϵ. For the present work, we follow the guidelines from published ABC tutorials (Farrell & Lewandowsky, 2018; Turner & Van Zandt, 2012). For the test stage, we summarized datasets with mean velocity of each band in the observed dataset as $V_{obs}^{(k)}$ and in the simulated dataset as $V_{sim}^{(k)}$, where $k \in \{1, 2, 3, 4, 5, 6\}$ represents each of the six velocity bands. For computing the discrepancy between datasets in the training stage, we aggregated training trials into three equally sized blocks (separately for each velocity band in the case of the varied group). After obtaining the summary statistics of the simulated and observed datasets, the discrepancy was computed as the mean of the absolute difference between simulated and observed datasets (Equation 6 and Equation 7).

$$discrep_{Test}(Data_{sim}, Data_{obs}) = \frac{1}{6} \sum_{k=1}^{6} |V_{obs}^{(k)} - V_{sim}^{(k)}| \tag{6}$$

$$\begin{aligned} discrep_{Train,constant}(Data_{sim}, Data_{obs}) &= \frac{1}{N_{blocks}} \sum_{j=1}^{N_{blocks}} |V_{obs,constant}^{(j)} - V_{sim,constant}^{(j)}| \\ discrep_{Train,varied}(Data_{sim}, Data_{obs}) &= \frac{1}{N_{blocks} \times 3} \sum_{j=1}^{N_{blocks}} \sum_{k=1}^{3} |V_{obs,varied}^{(j,k)} - V_{sim,varied}^{(j,k)}| \end{aligned} \tag{7}$$

The final component of our ABC implementation is the determination the appropriate value of ϵ. The setting of ϵ exerts strong influence on the approximated posterior distribution. Smaller values of ϵ increase the rejection rate, and improve the fidelity of the approximated posterior, while larger values result in an ABC sampler that reproduces the prior distribution. Because the individual participants in our dataset differed substantially in terms of the noisiness of their data, we employed an adaptive tolerance setting strategy to tailor ϵ to each individual. The initial value of ϵ was set to the overall standard deviation of each individuals velocity values. Thus, sampled parameter values that generated simulated data within a standard deviation of the observed data were accepted, while worse performing parameters were rejected. After every 300 samples the tolerance was allowed to increase only if the current acceptance rate of the algorithm was less than 1%. In such cases, the tolerance was shifted towards the average discrepancy of the 5 best samples obtained thus far. To ensure the acceptance rate did not become overly permissive, ϵ was also allowed to decrease every time a sample was accepted into the posterior.

For each of the 156 participants from Experiment 1, the ABC algorithm was run until 200 samples of parameters were accepted into the posterior distribution. Obtaining this number of posterior samples required an average of 205,000 simulation runs per participant. Fitting each combination of participant, Model (EXAM & ALM), and fitting method (Test only, Train only, Test & Train) required a total of 192 million simulation runs. To facilitate these intensive computational demands, we used the Future Package in R (Bengtsson, 2021), allowing us to parallelize computations across a cluster of ten M1 iMacs, each with 8 cores.

Modelling Results

Group level fits

Table 21: Mean model errors predicting testing data, aggregated over all participants and velocity bands. Note that Fit Method refers to how model parameters were optimized, while error values reflect mean absolute error for the 6 testing bands

Fit_Method	Model	Constant	Varied
	ALM	276.7	231.2

Fit_Method	Model	Constant	Varied
Test	EXAM	215.9	215.0
Test_Train	ALM	288.2	268.3
	EXAM	228.6	250.7
Train	ALM	528.1	368.7
	EXAM	340.3	370.9

The posterior distributions of the *c* and *lr* parameters are shown Figure 17 (i.e. these plots combine all the posterior samples from all of the subjects). There were substantial individual differences in the posteriors of both parameters, with the within-group individual differences generally swamped any between-group or between-model differences. The magnitude of these individual differences remains even if we consider only the single best parameter set for each subject.

We used the posterior distribution of *c* and *lr* parameters to generate a posterior predictive distribution of the observed data for each participant, which then allows us to compare the empirical data to the full range of predictions from each model. Model residuals are shown in the upper panels of Figure 16. The pattern of training stage residual errors are unsurprising across the combinations of models and fitting method . Differences between ALM and EXAM are generally minor (the two models have identical learning mechanisms). The differences in the magnitude of residuals across the three fitting methods are also straightforward, with massive errors for the 'fit to Test Only' model, and the smallest errors for the 'fit to train only' models. It is also noteworthy that the residual errors are generally larger for the first block of training, which is likely due to the initial values of the ALM weights being unconstrained by whatever initial biases participants tend to bring to the task. Future work may explore the ability of the models to capture more fine grained aspects of the learning trajectories. However for the present purposes, our primary interest is in the ability of ALM and EXAM to account for the testing patterns while being constrained, or not constrained, by the training data. All subsequent analyses and discussion will thus focus on the testing stage.

The residuals of model predictions for the testing stage (Figure 16) show the opposite pattern of fitting method - with models fit only to the test data showing the best performance, followed by models fit to both training and test data, and with models fit only to the training data showing the worst performance (note that y axes are scaled different between plots). Unsurprisingly, the advantage of EXAM is strongest for extrapolation positions (the three smallest bands for both groups - as well as the two highest bands for the Constant group). Although EXAM tends to perform better for both Constant and Varied participants (see also Table 21), the relative advantage of EXAM is generally larger for the Constant group - a pattern consistent across all three fitting methods. Panel B of Figure 16 directly compares the aggregated observed data to the posterior predictive distributions for the testing stage. Of interest are a) the extent to which the median estimates of the ALM and EXAM posteriors deviate from the observed medians for each velocity band; b) the ability of ALM and EXAM to discriminate between velocity bands; c) the relative performance of models that are constrained by the training data (i.e. the 'fit to train only' and 'fit to both' models) compared to the 'fit to test only' models; and d) the extent to which the variance of the posterior predictive distributions mimics the variance of the observed data.

- ** explain how the constant group ALM predictions for band 100 look deceptively good due to aggregation of a large subset of subjects having ALM predictions of 0 for vb100, and a large subset with ALM predictions close to their position 800 value. This is relected by much greater variance of the ALM esimates in the posterior predictive plot

- ** comment on how much constrained by the training data has a worse impact on the EXAM predictions for varied than for constant - perhaps due to the varied training data being much noisier than the constant training data.

- ** comment on EXAM doing a better job mimicing the within-condition variance of the observed data

- ** comment on the % of Constant subjects being best accounted for by EXAM being higher.

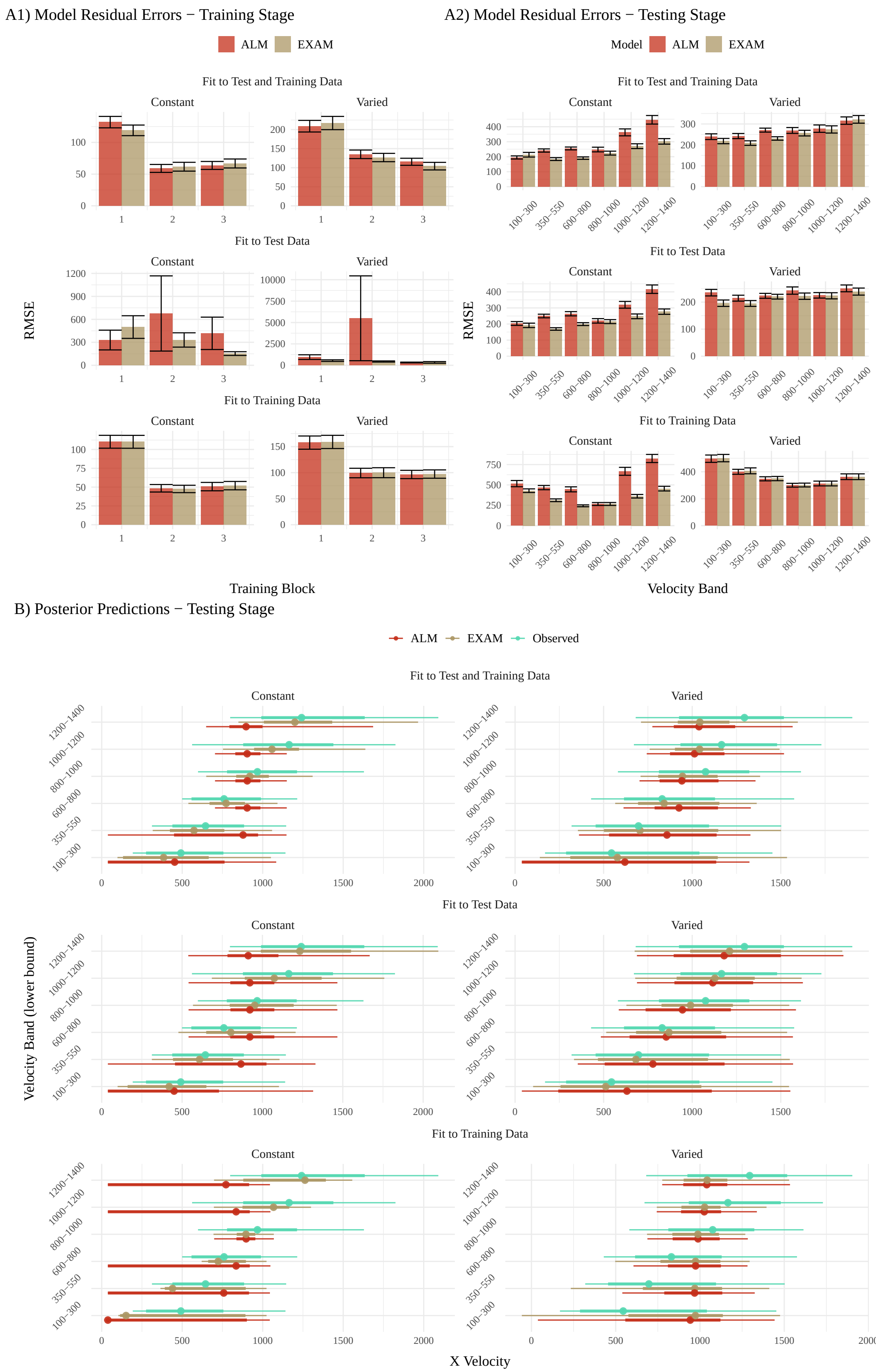


Figure 16: A) Model residuals for each combination of training condition, fit method, and model. Residuals reflect the difference between observed and predicte values. Lower values indicate better model fit. Note that y axes are scaled differently between facets. B) Full posterior predictive distributions vs. observed data fror participants.Points represent median values, thicker intervals represent 66% credible intervals and thin intervals represent 95% credible intervals around the median.

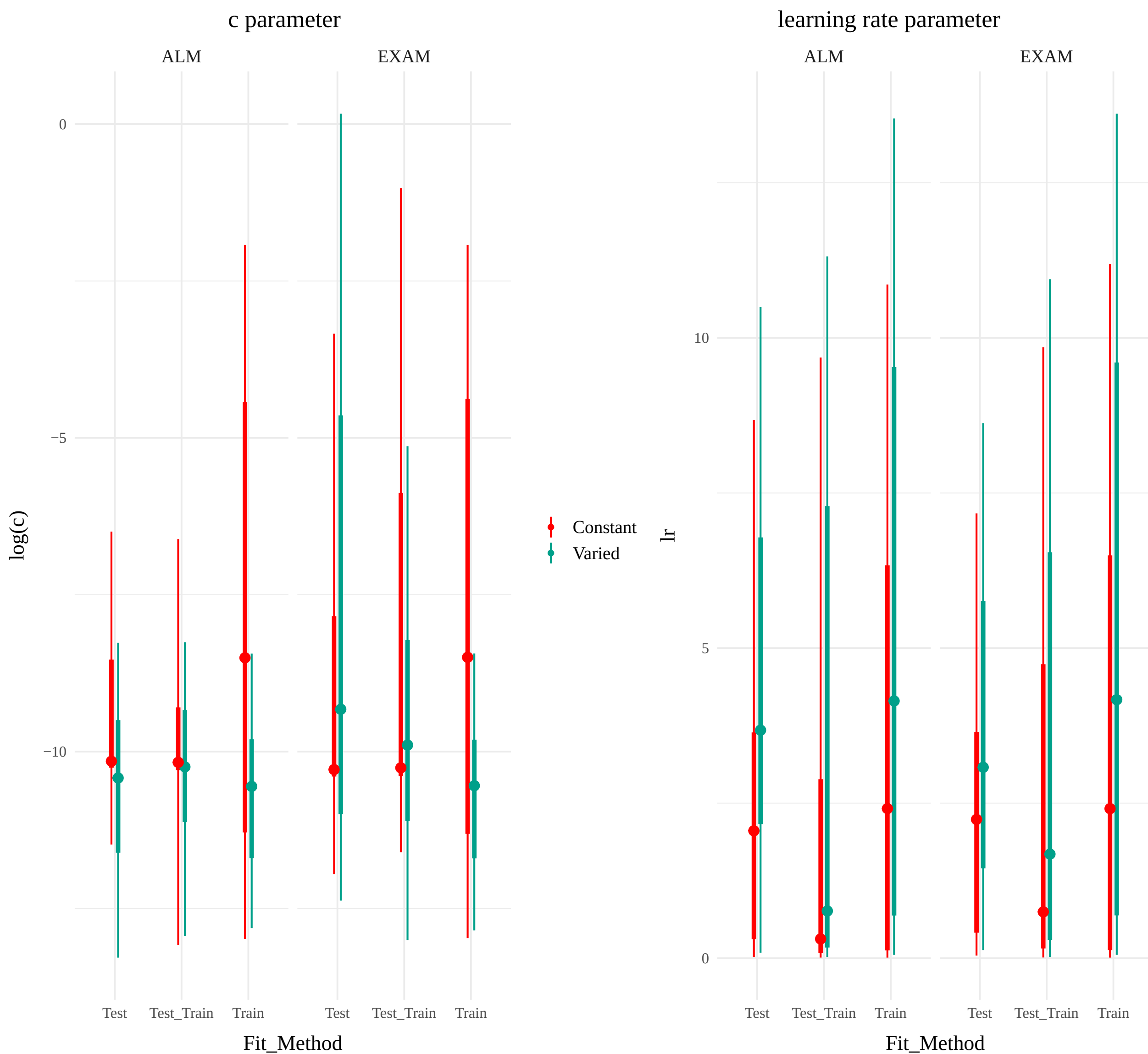


Figure 17: Posterior Distributions of c and lr parameters. Points represent median values, thicker intervals represent 66% credible intervals and thin intervals represent 95% credible intervals around the median. Note that the y axes of the plots for the c parameter are scaled logarithmically.

To more accurately assess the relative abilities of ALM and EXAM to capture important empirical patterns - we will now examine the predictions of both models for the subset of individual participants shown in Figure 18. Panel A presents three varied and constant participants who demonstrated a reasonable degree of discrimination between the 6 velocity bands during testing.

- ** comment on the different ways ALM can completely fail to mimic discrimination patterns (sbj. 35; sbj. 137),and on how it can sometimes partially succeed (sbj. 11; 14,74)
- ** comment on how EXAM can somtimes mimic non-monotonic spacing between bands due to associative stregth from training (i.e. subject 47)
- ** compare c values to slope parameters from the statistical models earlier in paper

Figure 18: Model predictions alongside observed data for a subset of individual participants. A) 3 constant and 3 varied participants fit to both the test and training data. B) 3 constant and 3 varied subjects fit to only the trainign data.

To add to appendix

		Constant		Varied	
Fit_Method	x	ALM	EXAM	ALM	EXAM
Test	100	203.3	191.4	233.5	194.8
	350	249.8	169.0	213.2	193.5
	600	264.1	199.5	222.4	219.2
	800	218.2	214.3	243.9	222.9
	1,000	315.9	245.3	224.4	222.3
	1,200	409.1	275.9	249.8	237.2
Test_Train	100	195.0	213.2	238.1	217.2
	350	241.4	183.9	241.0	207.1
	600	255.3	190.5	270.5	230.0
	800	244.9	222.0	270.3	257.9
	1,000	355.3	265.1	276.0	272.2
	1,200	437.3	297.0	313.8	319.9
Train	100	519.3	430.2	495.7	498.8
	350	466.6	310.9	398.6	405.2
	600	445.4	243.0	347.3	349.0
	800	260.9	261.2	298.5	300.0
	1,000	667.3	352.9	311.0	311.0
	1,200	809.3	443.5	361.3	361.3

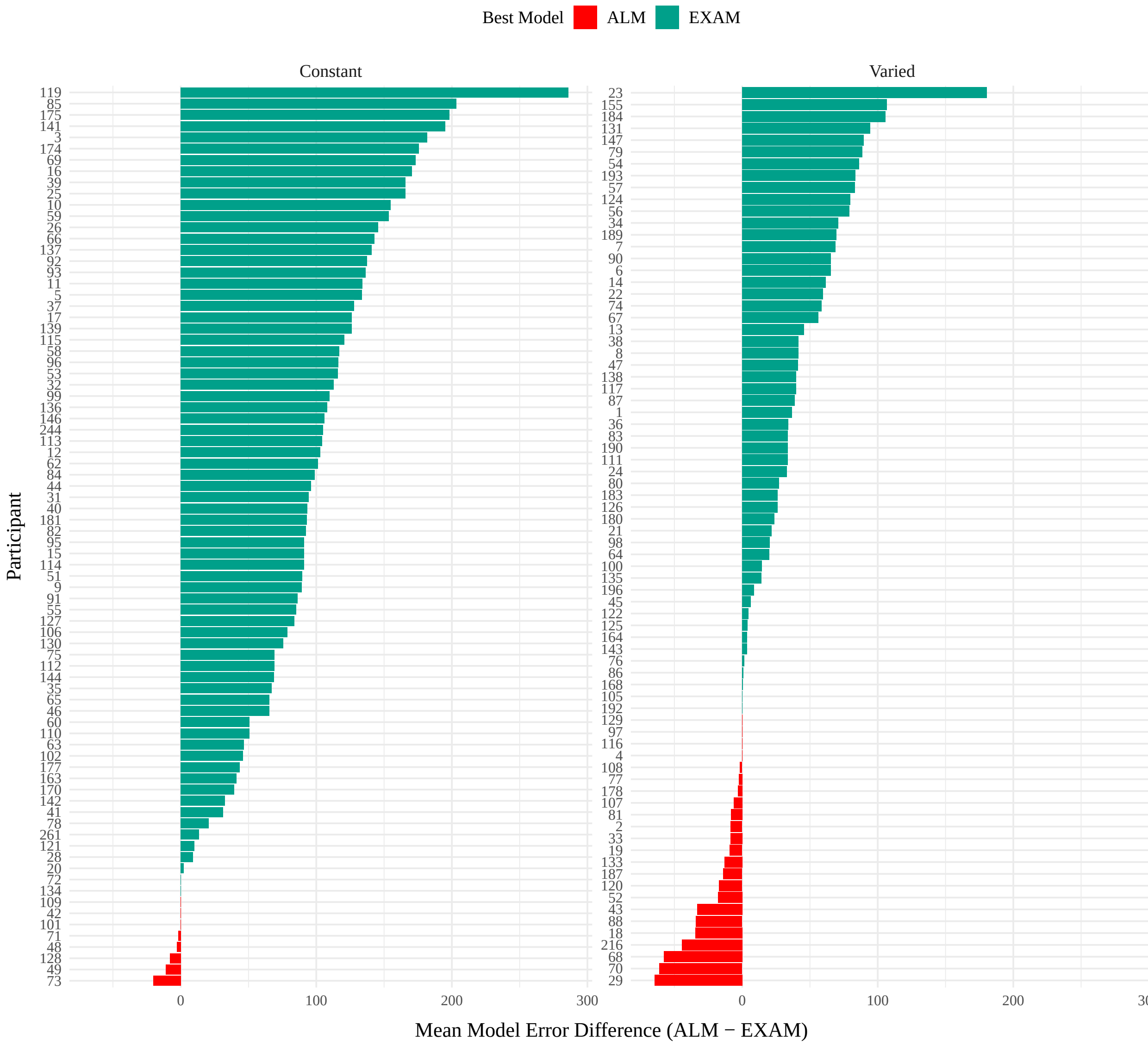


Figure 19: Difference in model errors for each participant, with models fit to both train and test data. Positive values favor EXAM, while negative values favor ALM.

References

Bengtsson, H. (2021). A Unifying Framework for Parallel and Distributed Processing in R using Futures. *The R Journal*, *13*(2), 208. https://doi.org/10.32614/RJ-2021-048

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

Brown, M. A., & Lacroix, G. (2017). Underestimation in linear function learning: Anchoring to zero or x-y similarity? *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *71*(4), 274–282. https://doi.org/10.1037/cep0000129

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

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, *117*(48), 30055–30062. https://doi.org/10.1073/pnas.1912789117

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

Farrell, S., & Lewandowsky, S. (2018). *Computational Modeling of Cognition and Behavior:* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781316272503

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

Kangasrääsiö, A., Jokinen, J. P. P., Oulasvirta, A., Howes, A., & Kaski, S. (2019). Parameter Inference for Computational Cognitive Models with Approximate Bayesian Computation. *Cognitive Science*, *43*(6), e12738. https://doi.org/10.1111/cogs.12738

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

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541

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

Page, M. (2000). Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences*, *23*(4), 443–467. https://doi.org/10.1017/S0140525X00003356

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.

Turner, B. M., Sederberg, P. B., & McClelland, J. L. (2016). Bayesian analysis of simulation-based models. *Journal of Mathematical Psychology*, *72*, 191–199. https://doi.org/10.1016/j.jmp.2014.10.001

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, *56*(2), 69–85. https://doi.org/10.1016/j.jmp.2012.02.005