Tuesday, April 2, 2024
Access the code, data, and analysis at https://github.com/tegorman13/htw

HTW

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

A longstanding issue across both science and instruction has been to understand how various aspects of an educational curriculum or training program influence learning acquisition and generalization. One such aspect, which has received a great deal of research attention, is the variability of examples experienced during training (Raviv et al., 2022). The influence of training variation has been studied in numerous domains, including category learning (Cohen et al., 2001; Posner & Keele, 1968), visuomotor learning (Berniker et al., 2014 ; Schmidt, 1975), language learning (Perry et al., 2010), and education (Braithwaite & Goldstone, 2015; Guo et al., 2014). The pattern of results is complex, with numerous studies finding both beneficial (Braun et al., 2009; Catalano & Kleiner, 1984; Roller et al., 2001), as well as null or negative effects (Brekelmans et al., 2022 ; Hu & Nosofsky, 2024; Van Rossum, 1990). The present study seeks to contribute to the large body of existing research by examining the influence of variability in visuomotor function learning - a domain in which it has been relatively under-studied.

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).

Variability and Function Learning

The influence of variability on function learning tasks has received relatively little attention. The study by DeLosh et al. (1997) (described in detail above) did include a variability manipulation (referred to as density in their paper), wherein participants were trained with either either 8, 20, or 50 unique input-output pairs, with the total number of training trials held constant. They found a minimal influence of variability on training performance, and no difference between groups in interpolation or extrapolation, with all three variability conditions displaying accurate interpolation, and linearly biased extrapolation that was well accounted for by the EXAM model.

In the domain of visuomotor learning, van Dam & Ernst (2015) employed a task which required participants to learn a linear function between the spikiness of shape stimuli and the correct horizontal position to make a rapid pointing response. The shapes ranged from very spiky to completely circular at the extreme ends of the space. Participants trained with intermediate shapes from a lower variation (2 shapes) or higher variation (5 shapes) condition, with the 2 items of the lower varied condition matching the items used on the extreme ends of the higher variation training space. Learning was significantly slower in the higher variation group. However, the two conditions did not differ when tested with novel shapes, with both groups producing extrapolation responses of comparable magnitudes to the most similar training item, rather than in accordance with the true linear function. The authors accounted for both learning and extrapolation performance with a Bayesian learning model. Similar to ALM, the bayesian model assumes that generalization occurs as a Gaussian function of the distance between stimuli. However unlike ALM, the bayesian learning model utilizes more elaborate probabilistic stimulus representations, with a separate Kalman Filter for each shape stimulus.

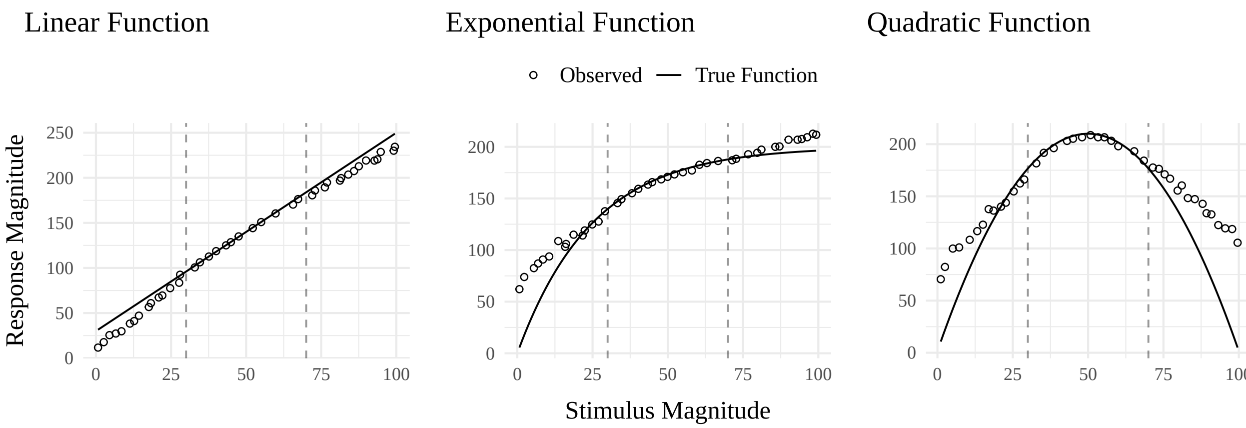


Figure 1: Generalization reproduced patterns from DeLosh et al. (1997) Figure 3. Stimulii that fall within the dashed lines are interpolations of the training examples.

Overview Of Present Study

The present study investigates the influence of training variability on learning, generalization, and extrapolation in a uni-dimensional visuomotor function learning task. To the best of our knowledge, this research is the first to employ the classic constant vs. varied training manipulation, commonly used in the literature on the benefits of variability, in the context of a uni-dimensional function learning task. Across three experiments, we compare constant and varied training conditions in terms of learning performance, extrapolation accuracy, and the ability to reliably discriminate between stimuli.

To account for the empirical results, we will apply a series of computational models, including the Associative Learning Model (ALM) and the Extrapolation-Association Model (EXAM). Notably, this study is the first to employ approximate Bayesian computation (ABC) to fit these models to individual subject data, enabling us to thoroughly investigate the full range of posterior predictions of each model, and to examine the the ability of these influential models of function learning to account for both the group level and individual level data.

Methods

Participants A total of 156 participants were recruited from the Indiana University Introductory Psychology Course. Participants were randomly assigned to one of two training conditions: varied training or constant training.

Task. The "Hit The Wall" (HTW) visuomotor extrapolation task task was programmed in Javascript, making heavy use of the phaser.io game library. The HTW task involved launching a projectile such that it would strike the "wall" at target speed indicated at the top of the screen (see Figure 2). The target velocities were given as a range, or band, of acceptable velocity values (e.g. band 800-1000). During the training stage, participants received feedback indicating whether they had hit the wall within the target velocity band, or how many units their throw was above or below from the target band. Participants were instructed that only the x velocity component of the ball was relevant to the task. The y velocity, or the location at which the ball struck the wall, had no influence on the task feedback.

/

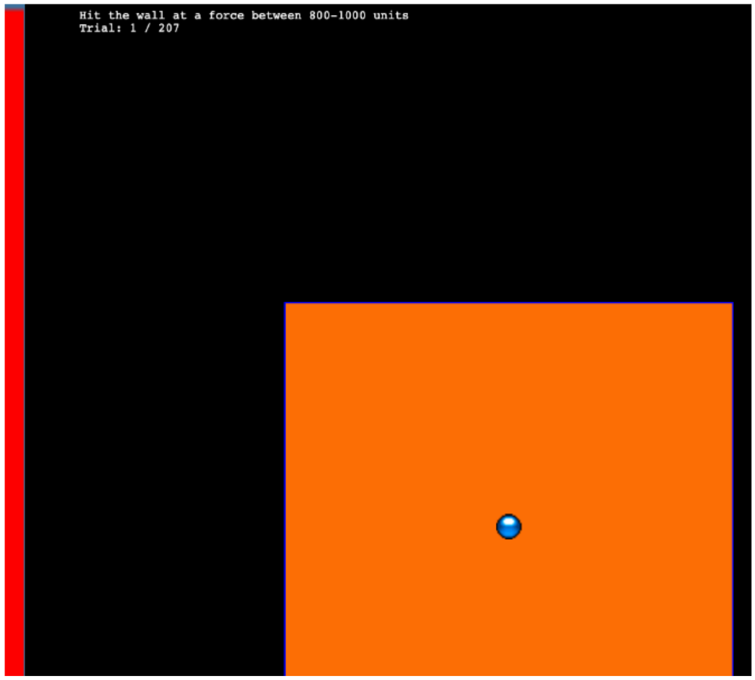


Figure 2: The Hit the wall task. Participants launch the blue ball to hit the red wall at the target velocity band indicated at the top of the screen. The ball must be released from within the orange square - but the location of release, and the location at which the ball strikes the wall are both irrelevant to the task feedback.

Procedure. All participants completed the task online. Participants were provided with a description of the experiment and indicated informed consent. Figure 3 illustrates the general procedure. 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). Participants in the constant training condition trained on only one velocity band (800-1000) - the closest band to what would be the novel extrapolation bands in the testing stage.

Following the training stage, participants proceeded immediately to the testing stage. Participants were tested from all six velocity bands, in two separate stages. In the novel extrapolation testing stage, participants completed "no-feedback" testing from three novel extrapolation bands (100-300, 350-550, and 600-800), with each band consisting of 15 trials. Participants were also tested from the three velocity bands that were trained by the varied condition (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 order in which participants completed the novel-extrapolation and testing-from-3-varied bands was counterbalanced across participants. A final training stage presented participants with "feedback" testing for each of the three extrapolation bands (100-300, 350-550, and 600-800).

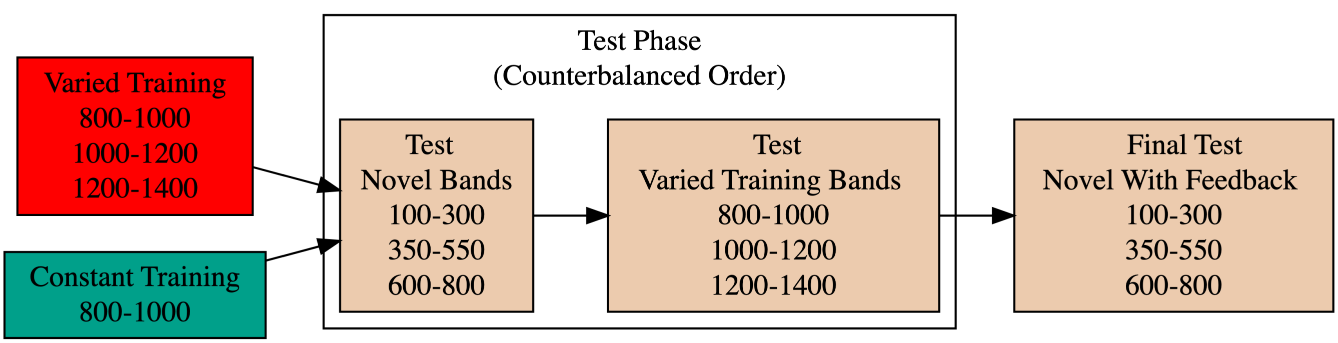


Figure 3: Experiment 1 Design. Constant and Varied participants complete different training conditions.

Analyses Strategy

All data processing and statistical analyses were performed in R version 4.32 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 avoiding convergence issues common to the frequentist analogues of our mixed models.

Each model was set to run with 4 chains, 5000 iterations per chain, with the first 2500 discarded as warmup chains. Rhat values were 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. For each model, we report 1) the mean values of the posterior distribution for the parameters of interest, 2) the lower and upper credible intervals (CrI), and the probability of direction value (pd).

Group Comparison	Code	Data
End of Training Accuracy	`brm(dist ~ condit)`	Final Training Block
Test Accuracy	`brm(dist ~ condit * bandType + (1|id) + (1|bandInt)`	All Testing trials
Band Discrimination	`brm(vx ~ condit * band +(1 + bandInt|id)`	All Testing Trials

In each experiment we compare varied and constant conditions in terms of 1) accuracy in the final training block; 2) testing accuracy as a function of band type (trained vs. extrapolation bands); 3) extent of discrimination between all six testing bands. We quantified accuracy as the absolute deviation between the response velocity and the nearest boundary of the target band. Thus, when the target band was velocity 600-800, throws of 400, 650, and 900 would result in deviation values of 200, 0, and 100, respectively. The degree of discrimination between bands was index by fitting a linear model predicting the response velocity as a function of the target velocity. Participants who reliably discriminated between velocity bands tended to haves slope values ~1, while participants who made throws irrespective of the current target band would have slopes ~0.

Results

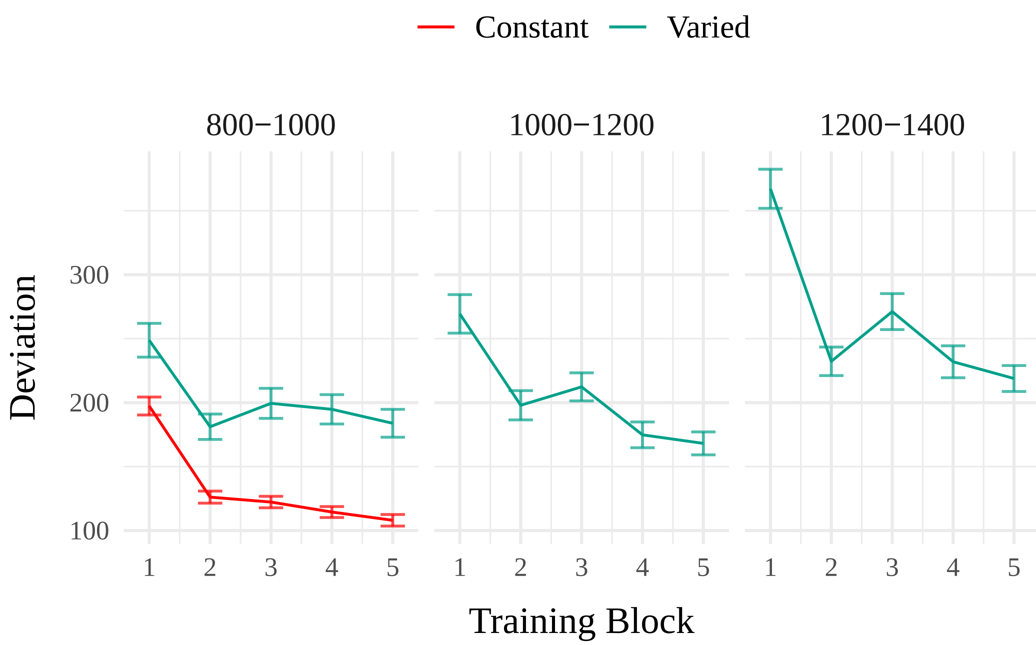


Figure 4: Experiment 1 Training Stage. Deviations from target band across training blocks. Lower values represent greater accuracy.

Table 2: Experiment 1 - End of training performance. The Intercept represents the average of the baseline (constant condition), and the conditVaried coefficient reflects the difference between the constant and varied groups. A larger positive estimates indicates a greater deviation (lower accuracy) for the varied group.

Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	106.34	95.46	117.25	1
conditVaried	79.64	57.92	101.63	1

Training. Figure 4 displays the average deviations across training blocks for the varied group, which trained on three velocity bands, and the constant group, which trained on one velocity band. To compare the training conditions at the end of training, we analyzed performance on the 800-1000 velocity band, which both groups trained on. The full model results are shown in Table 1. The varied group had a significantly greater deviation than the constant group in the final training block, (β = 79.64, 95% CrI [57.92, 101.63]; pd = 100%).

Table 3: Experiment 1 testing accuracy. Main effects of condition and band type (training vs. extrapolation), and the interaction between the two factors. Larger coefficients indicate larger deviations from the baselines (Condition=constant & bandType=Trained) - and a positive interaction coefficient indicates disproporionate deviation for the varied condition on the extrapolation bands

Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	152.55	70.63	229.85	1.0
conditVaried	39.00	-21.10	100.81	0.9
bandTypeExtrapolation	71.51	33.24	109.60	1.0
conditVaried:bandTypeExtrapolation	66.46	32.76	99.36	1.0

Testing. To compare accuracy between groups in the testing stage, we fit a Bayesian mixed effects model predicting deviation from the target band as a function of training condition (varied vs. constant) and band type (trained vs. extrapolation), with random intercepts for participants and bands. The model results are shown in Table 3. The main effect of training condition was not significant (β = 39, 95% CrI [-21.1, 100.81]; pd = 89.93%). The extrapolation testing items had a significantly greater deviation than the training bands (β = 71.51, 95% CrI [33.24, 109.6]; pd = 99.99%). Most importantly, the interaction between training condition and band type was significant (β = 66.46, 95% CrI [32.76, 99.36]; pd = 99.99%), As shown in Figure 5, the varied group had disproportionately larger deviations compared to the constant group in the extrapolation bands.

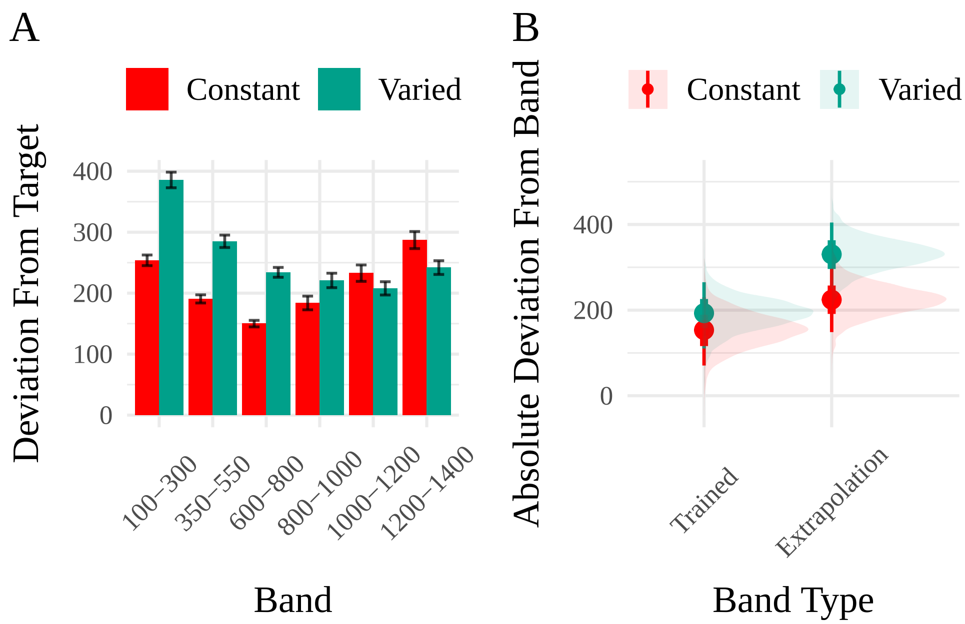


Figure 5: A) Deviations from target band during testing without feedback stage. B) Conditional effect of condition (Constant vs. Varied) and testing band type (training vs. extrapolation) on testing accuracy. Error bars represent 95% credible intervals.

Table 4: Experiment 1. Bayesian Mixed Model Predicting velocity as a function of condition (Constant vs. Varied) and Velocity Band. Larger coefficients on Band represent greater sensitivity/discrimination.

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

Finally, to assess the ability of both conditions to discriminate between velocity bands, we fit a model predicting velocity as a function of training condition and velocity band, with random intercepts and random slopes for each participant. See Table 5 for the full model results. The estimated coefficient for training condition (β = 164.05, 95% CrI [45.5, 278.85], pd = 99.61%) 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 coefficient on the Band predictor (β = 0.71 95% CrI [0.62, 0.8], pd = 100%). 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 (β = -0.14, 95% CrI [-0.26, -0.01], pd = 98.39%), suggests that the discrimination was somewhat modulated by training condition, with the varied participants showing less sensitivity between bands than the constant condition. This difference is depicted visually in Figure 6.

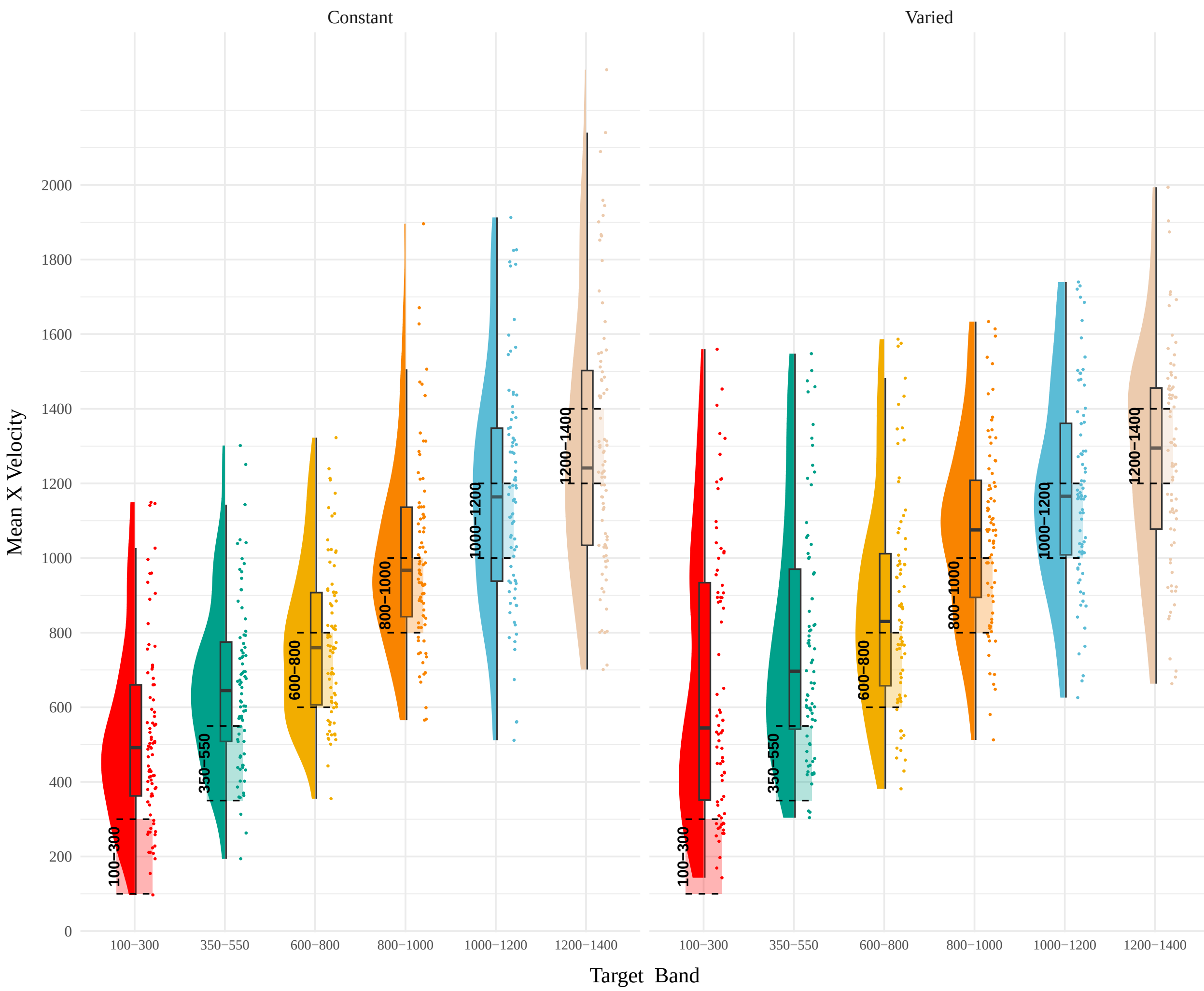
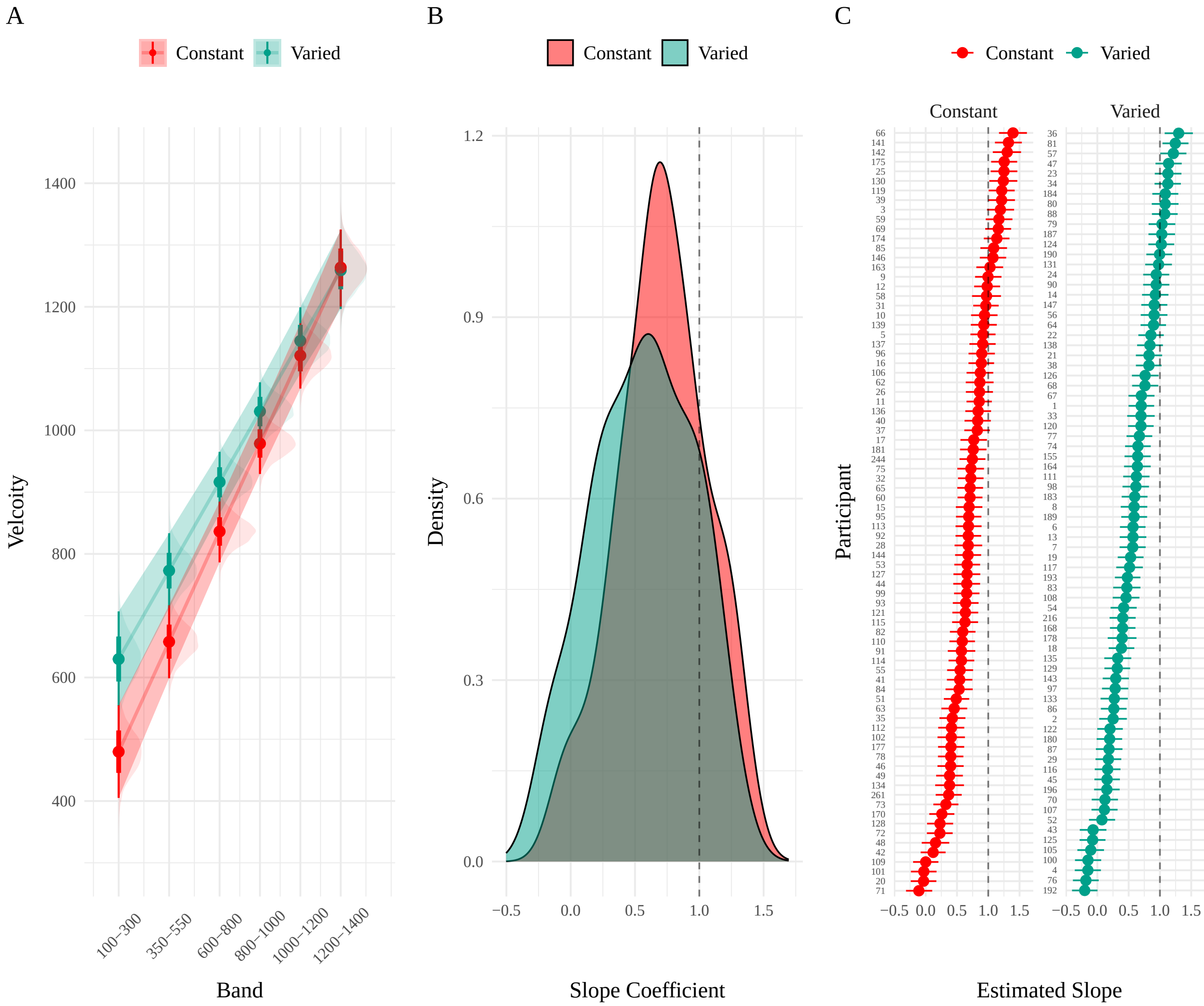


Figure 6: Empirical distribution of velocities producing in testing stage. Translucent bands with dash lines indicate the correct range for each velocity band.

Table 5



(a) Experiment 1. 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.

E1 Summary

In Experiment 1, we investigated how variability in training influenced participants' ability learn and extrapolate in a visuomotor task. Our findings that training with variable conditions rresulted in lower final training performance is consistent with much of the prior researchon the influence of training variability (Raviv et al., 2022; Soderstrom & Bjork, 2015), and is particularly unsurprising in the present work, given that the constant group received three times the amount of training on the velocity band common to the two conditions.

More importantly, the varied training group exhibited significantly larger deviations from the target velocity bands during the testing phase, particularly for the extrapolation bands that were not encountered by either condition during training.

Experiment 2

Methods & Procedure

The task and procedure of Experiment 2 was identical to Experiment 1, with the exception that the training and testing bands were reversed (see Figure 8). The Varied group trained on bands 100-300, 350-550, 600-800, and the constant group trained on band 600-800. Both groups were tested from all six bands. A total of 110 participants completed the experiment (Varied: 55, Constant: 55).

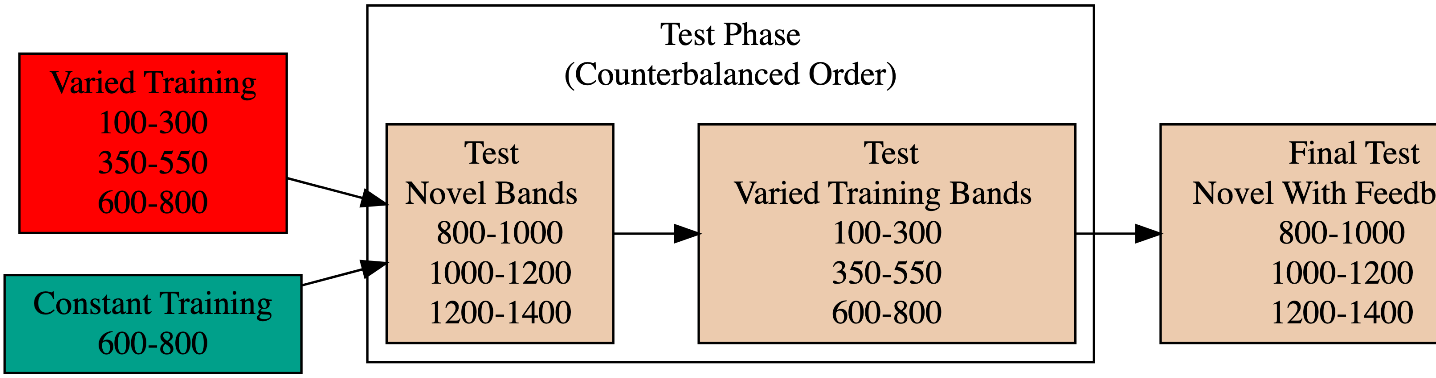


Figure 8: Experiment 2 Design. Constant and Varied participants complete different training conditions. The training and testing bands are the reverse of Experiment 1.

Results

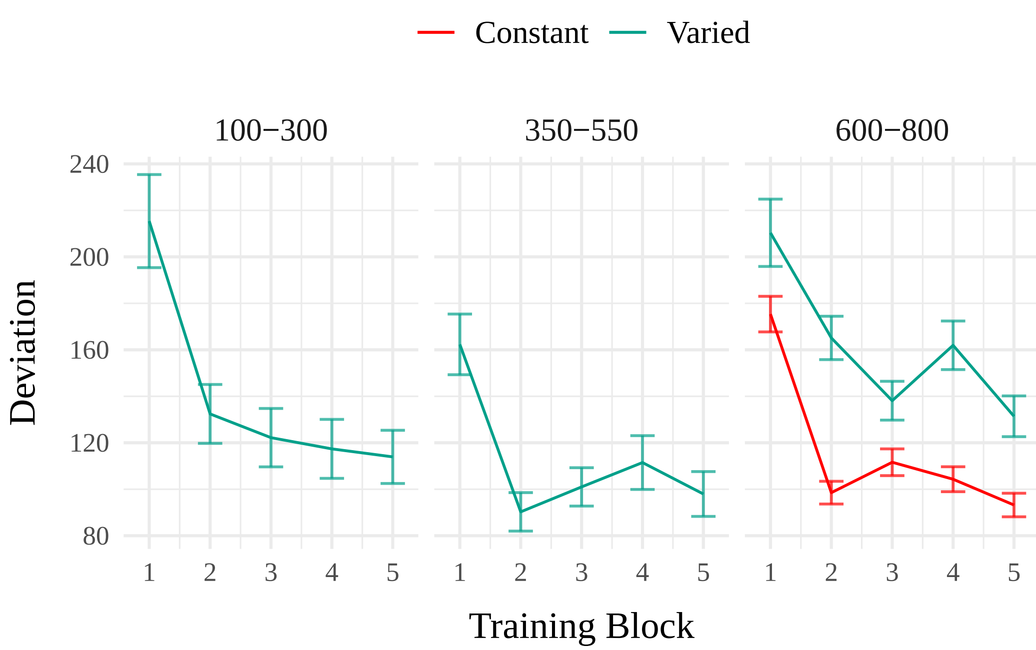


Figure 9: Experiment 2 Training Stage. Deviations from target band across training blocks. Lower values represent greater accuracy.

Table 6: Experiment 2 - End of training performance. The Intercept represents the average of the baseline (constant condition), and the conditVaried coefficient reflects the difference between the constant and varied groups. A larger positive coefficient indicates a greater deviation (lower accuracy) for the varied group.

Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	91.01	80.67	101.26	1
conditVaried	36.15	16.35	55.67	1

Training. Figure 9 presents the deviations across training blocks for both constant and varied training groups. We again compared training performance on the band common to both groups (600-800). The full model results are shown in Table 1. The varied group had a significantly greater deviation than the constant group in the final training block, (β = 36.15, 95% CrI [16.35, 55.67]; pd = 99.95%).

Table 7: Experiment 2 testing accuracy. Main effects of condition and band type (training vs. extrapolation), and the interaction between the two factors. Larger coefficient estimates indicate larger deviations from the baselines (constant & trained bands) - and a positive interaction coefficient indicates disproporionate deviation for the varied condition on the extrapolation bands

Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	190.91	125.03	259.31	1.00
conditVaried	-20.58	-72.94	33.08	0.78
bandTypeExtrapolation	38.09	-6.94	83.63	0.95
conditVaried:bandTypeExtrapolation	82.00	41.89	121.31	1.00

Testing Accuracy. The analysis of testing accuracy examined deviations from the target band as influenced by training condition (Varied vs. Constant) and band type (training vs. extrapolation bands). The results, summarized in Table 7, reveal no significant main effect of training condition (β = -20.58, 95% CrI [-72.94, 33.08]; pd = 77.81%). However, the interaction between training condition and band type was significant (β = 82, 95% CrI [41.89, 121.31]; pd = 100%), with the varied group showing disproportionately larger deviations compared to the constant group on the extrapolation bands (see Figure 10).

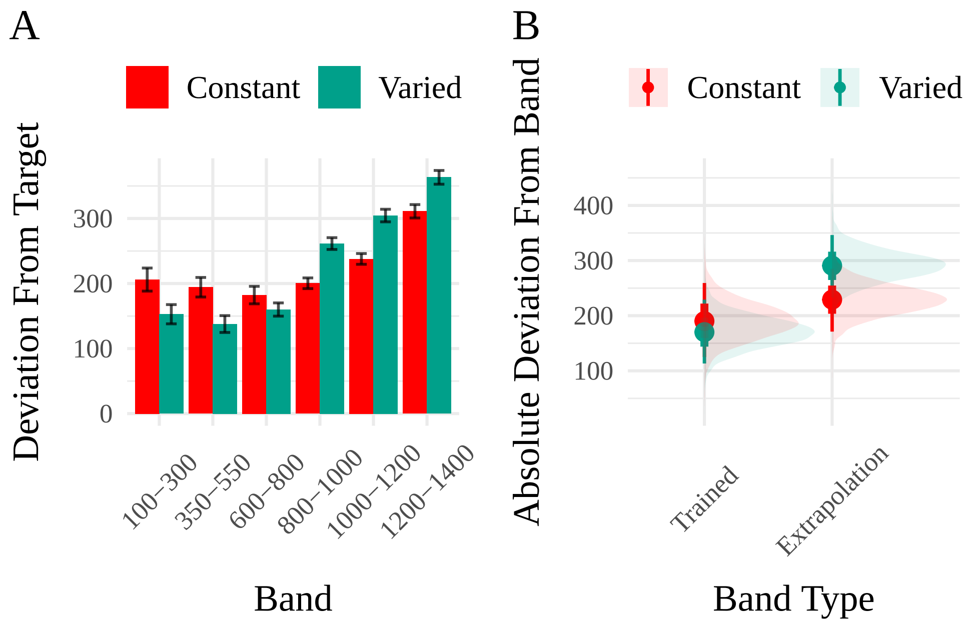


Figure 10: A) Deviations from target band during testing without feedback stage. B) Estimated marginal means for the interaction between training condition and band type. Error bars represent 95% confidence intervals.

Table 8: 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

Testing Discrimination. Finally, to assess the ability of both conditions to discriminate between velocity bands, we fit a model predicting velocity as a function of training condition and velocity band, with random intercepts and random slopes for each participant. The full model results are shown in Table 9. The overall slope on target velocity band predictor was significantly positive, (β = 0.71, 95% CrI [0.58, 0.84]; pd= 100%), indicating that participants exhibited discrimination between bands. The interaction between slope and condition was not significant, (β = -0.06, 95% CrI [-0.24, 0.13]; pd= 72.67%), suggesting that the two conditions did not differ in their ability to discriminate between bands (see Figure 11).

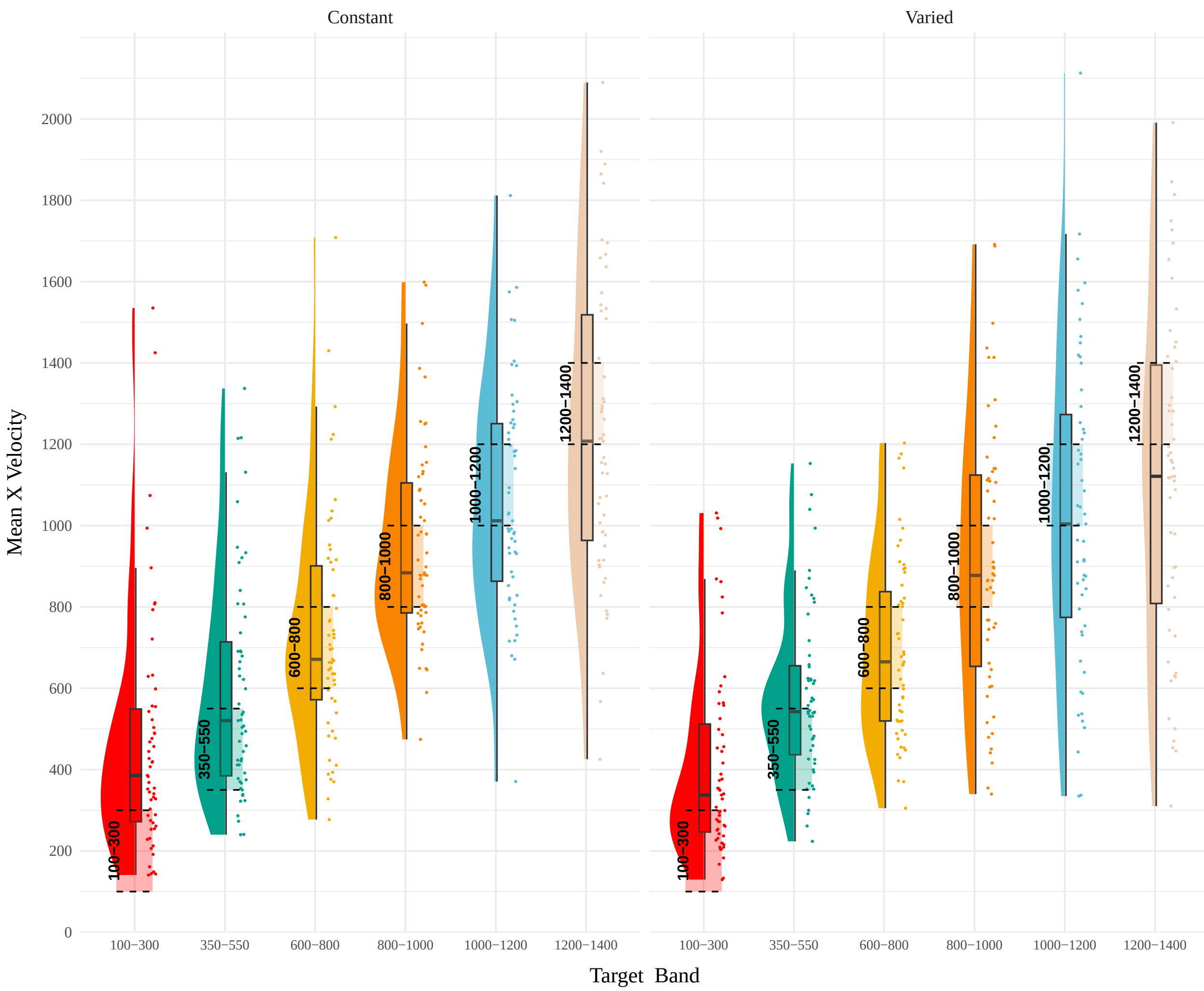
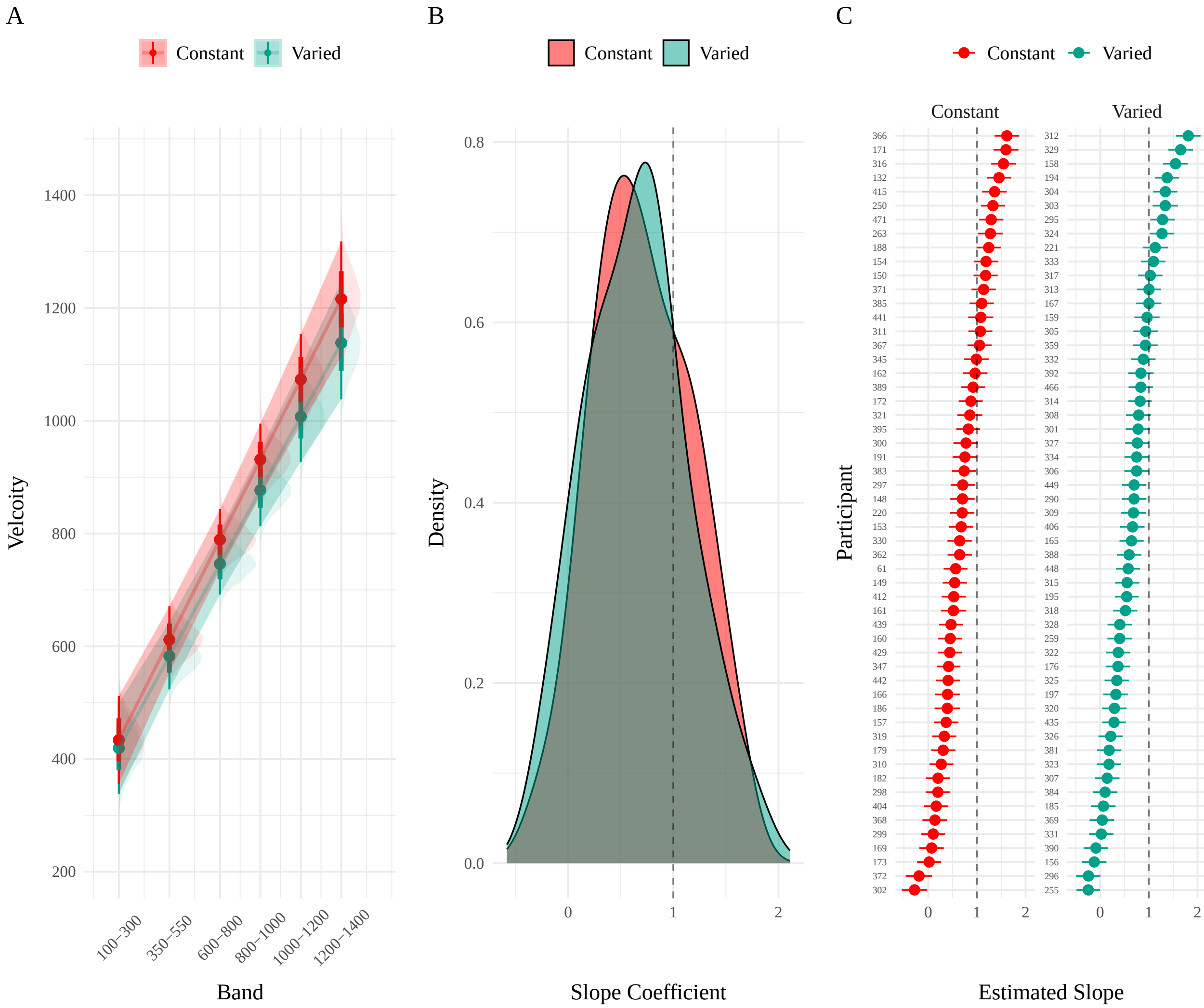


Figure 11: E2 testing x velocities. Translucent bands with dash lines indicate the correct range for each velocity band.

Table 9



(a) 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.

Experiment 2 Summary

Experiment 2 extended the findings of Experiment 1 by examining the effects of training variability on extrapolation performance in a visuomotor function learning task, but with reversed training and testing bands. Similar to Experiment 1, the Varied group exhibited poorer performance during training and testing. However unlike experiment 1, the Varied group did not show a significant difference in discrimination between bands.

Experiment 3

Methods & Procedure

The major adjustment of Experiment 3 is for participants to receive ordinal feedback during training, in contrast to the continuous feedback of the prior experiments. After each training throw, participants are informed whether a throw was too soft, too hard, or correct (i.e. within the target velocity range). All other aspects of the task and design are identical to Experiments 1 and 2. We utilized the order of training and testing bands from both of the prior experiments, thus assigning participants to both an order condition (Original or Reverse) and a training condition (Constant or Varied). Participants were once again recruited from the online Indiana University Introductory

Psychology Course pool. Following exclusions, 195 participants were included in the final analysis, n=51 in the Constant-Original condition, n=59 in the Constant-Reverse condition, n=39 in the Varied-Original condition, and n=46 in the Varied-Reverse condition.

Results

Table 10: Experiment 3 - End of training performance. The Intercept represents the average of the baseline (constant condition), and the conditVaried coefficient reflects the difference between the constant and varied groups. A larger positive coefficient indicates a greater deviation (lower accuracy) for the varied group.

Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	121.86	109.24	134.60	1.00
conditVaried	64.93	36.99	90.80	1.00
bandOrderReverse	1.11	-16.02	18.16	0.55
conditVaried:bandOrderReverse	-77.02	-114.16	-39.61	1.00

Training. Figure 13 displays the average deviations from the target band across training blocks, and Table 10 shows the results of the Bayesian regression model predicting the deviation from the common band at the end of training (600-800 for reversed order, and 800-1000 for original order conditions). The main effect of training condition is significant, with the varied condition showing larger deviations (β = 64.93, 95% CrI [36.99, 90.8]; pd = 100%). The main effect of band order is not significant β = 1.11, 95% CrI [-16.02, 18.16]; pd = 55.4%, however the interaction between training condition and band order is significant, with the varied condition showing greater accuracy in the reverse order condition (β = -77.02, 95% CrI [-114.16, -39.61]; pd = 100%).

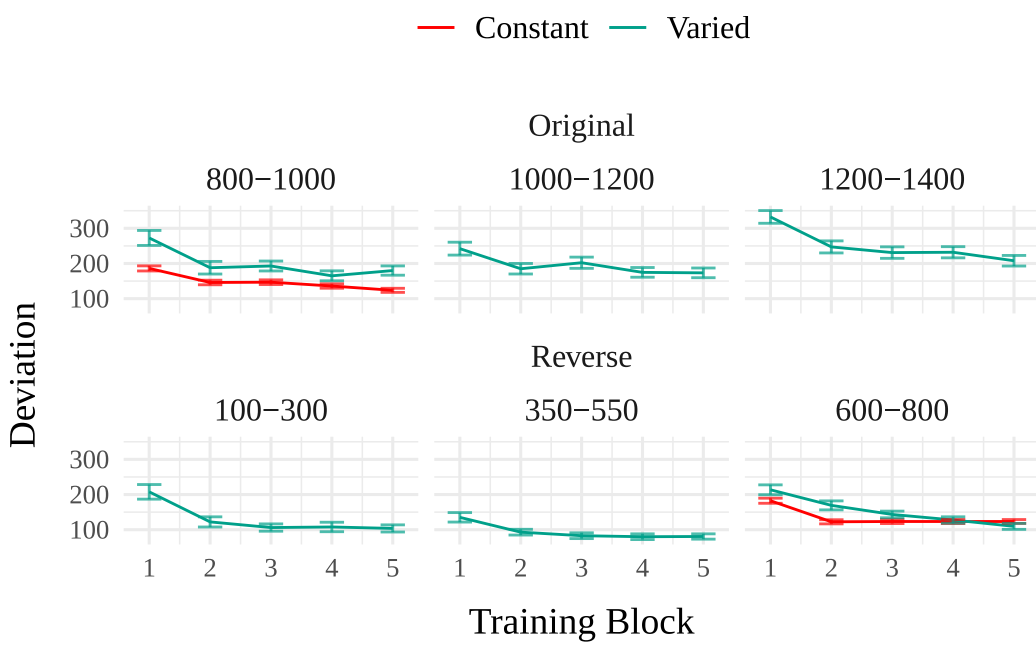


Figure 13: E3. Deviations from target band during testing without feedback stage.

Table 11: Experiment 3 testing accuracy. Main effects of condition and band type (training vs. extrapolation), and the interaction between the two factors. Larger coefficient estimates indicate larger deviations from the baselines (constant training; trained bands & original order) - and a positive interaction coefficient indicates disproportionate deviation for the varied condition on the extrapolation bands

Term	Estimate	95% CrI Lower	95% CrI Upper	pd
Intercept	288.65	199.45	374.07	1.00
conditVaried	-40.19	-104.68	23.13	0.89
bandTypeExtrapolation	-23.35	-57.28	10.35	0.92
bandOrderReverse	-73.72	-136.69	-11.07	0.99
conditVaried:bandTypeExtrapolation	52.66	14.16	90.23	1.00
conditVaried:bandOrderReverse	-37.48	-123.28	49.37	0.80
bandTypeExtrapolation:bandOrderReverse	80.69	30.01	130.93	1.00
conditVaried:bandTypeExtrapolation:bandOrderReverse	30.42	-21.00	81.65	0.87

Testing Accuracy. Table 11 presents the results of the Bayesian mixed efects model predicting absolute deviation from the target band during the testing stage. There was no significant main effect of training condition,β = -40.19, 95% CrI [-104.68, 23.13]; pd = 89.31%, or band type,β = -23.35, 95% CrI [-57.28, 10.35]; pd = 91.52%. However the effect of band order was significant, with the reverse order condition showing lower deviations, β = -73.72, 95% CrI [-136.69, -11.07]; pd = 98.89%. The interaction between training condition and band type was also significant β = 52.66, 95% CrI [14.16, 90.23]; pd = 99.59%, with the varied condition showing disproprionately large deviations on the extrapolation bands compared to the constant group. There was also a significant interaction between band type and band order, β = 80.69, 95% CrI [30.01, 130.93]; pd = 99.89%, such that the reverse order condition showed larger deviations on the extrapolation bands. No other interactions were significant.

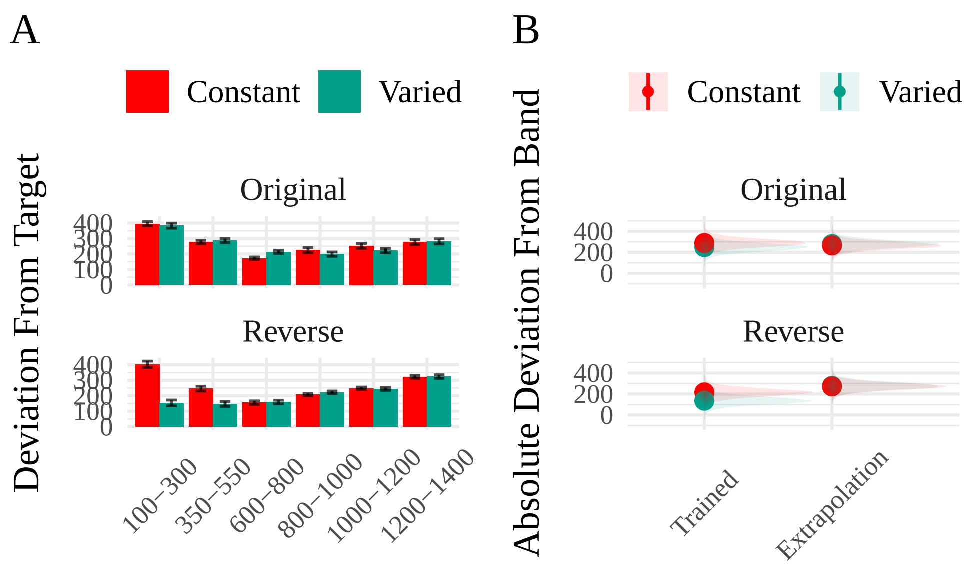


Figure 14: Experiment 3 Testing Accuracy. A) Deviations from target band during testing without feedback stage. B) Conditional effect of condition (Constant vs. Varied) and testing band type (training vs. extrapolation) on testing accuracy. Error bars represent 95% confidence intervals.

Table 12: 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	601.83	504.75	699.42	1.00
conditVaried	12.18	-134.94	162.78	0.56
bandOrderReverse	13.03	-123.89	144.67	0.58
Band	0.49	0.36	0.62	1.00
conditVaried:bandOrderReverse	-338.15	-541.44	-132.58	1.00
conditVaried:Band	-0.04	-0.23	0.15	0.67
bandOrderReverse:bandInt	-0.10	-0.27	0.08	0.86
conditVaried:bandOrderReverse:bandInt	0.42	0.17	0.70	1.00

Testing Discrimination. The full results of the discrimination model are presented in Table 11. For the purposes of assessing group differences in discrimination, only the coefficients including the band variable are of interest. The baseline effect of band represents the slope cofficient for the constant training - original order condition, this effect was significant β = 0.49, 95% CrI [0.36, 0.62]; pd = 100%. Neither of the two way interactions reached significance, β = -0.04, 95% CrI [-0.23, 0.15]; pd = 66.63%, β = -0.1, 95% CrI [-0.27, 0.08]; pd = 86.35%. However, the three way interaction between training condition, band order, and target band was significant, β = 0.42, 95% CrI [0.17, 0.7]; pd = 99.96% - indicating that the varied condition showed a greater slope coefficient on the reverse order bands, compared to the constant condition - this is clearly shown in Figure 15, where the steepness of the best fitting line for the varied-reversed condition is noticably steeper than the other conditions.

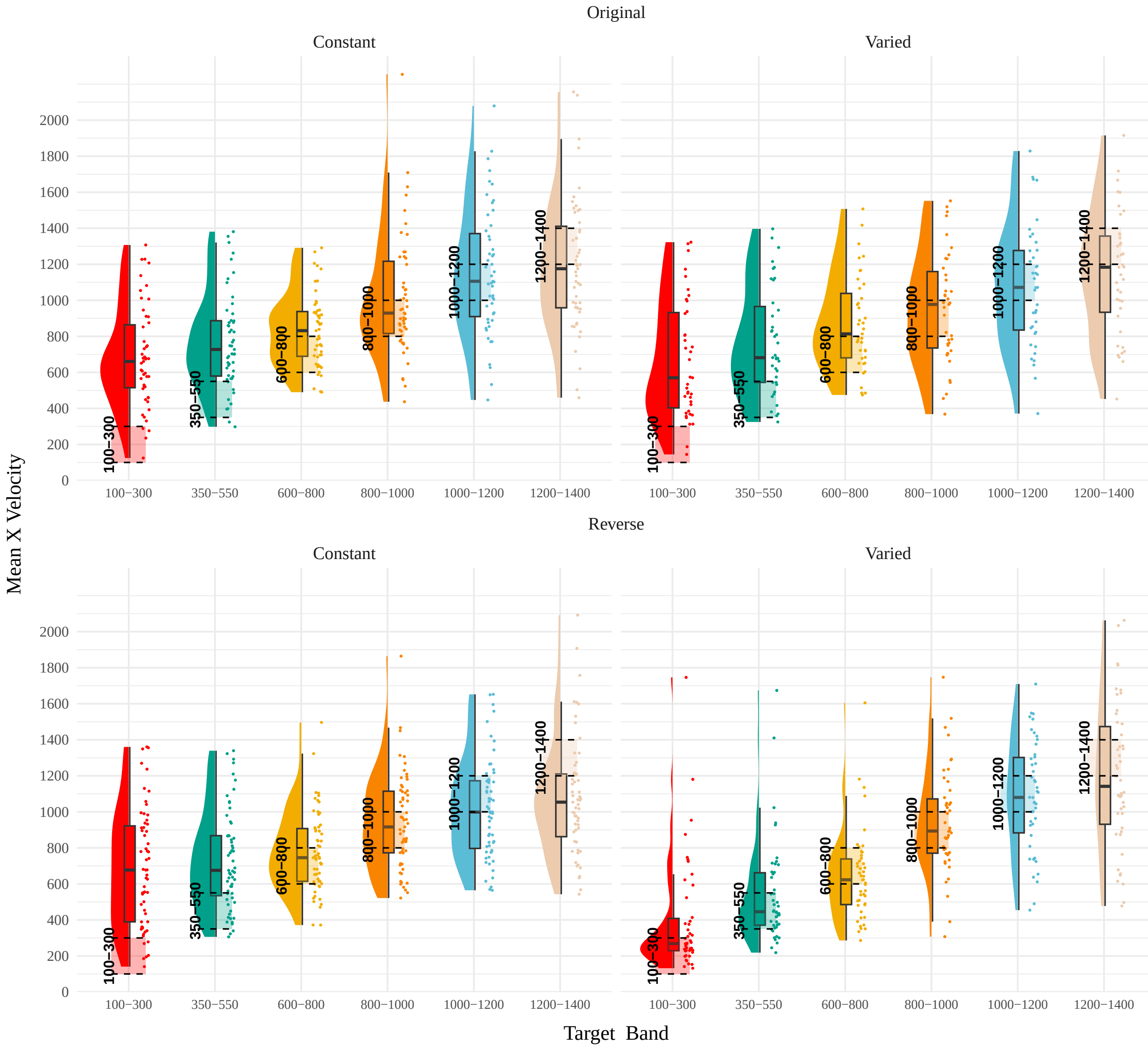


Figure 15: e3 testing x velocities. Translucent bands with dash lines indicate the correct range for each velocity band.

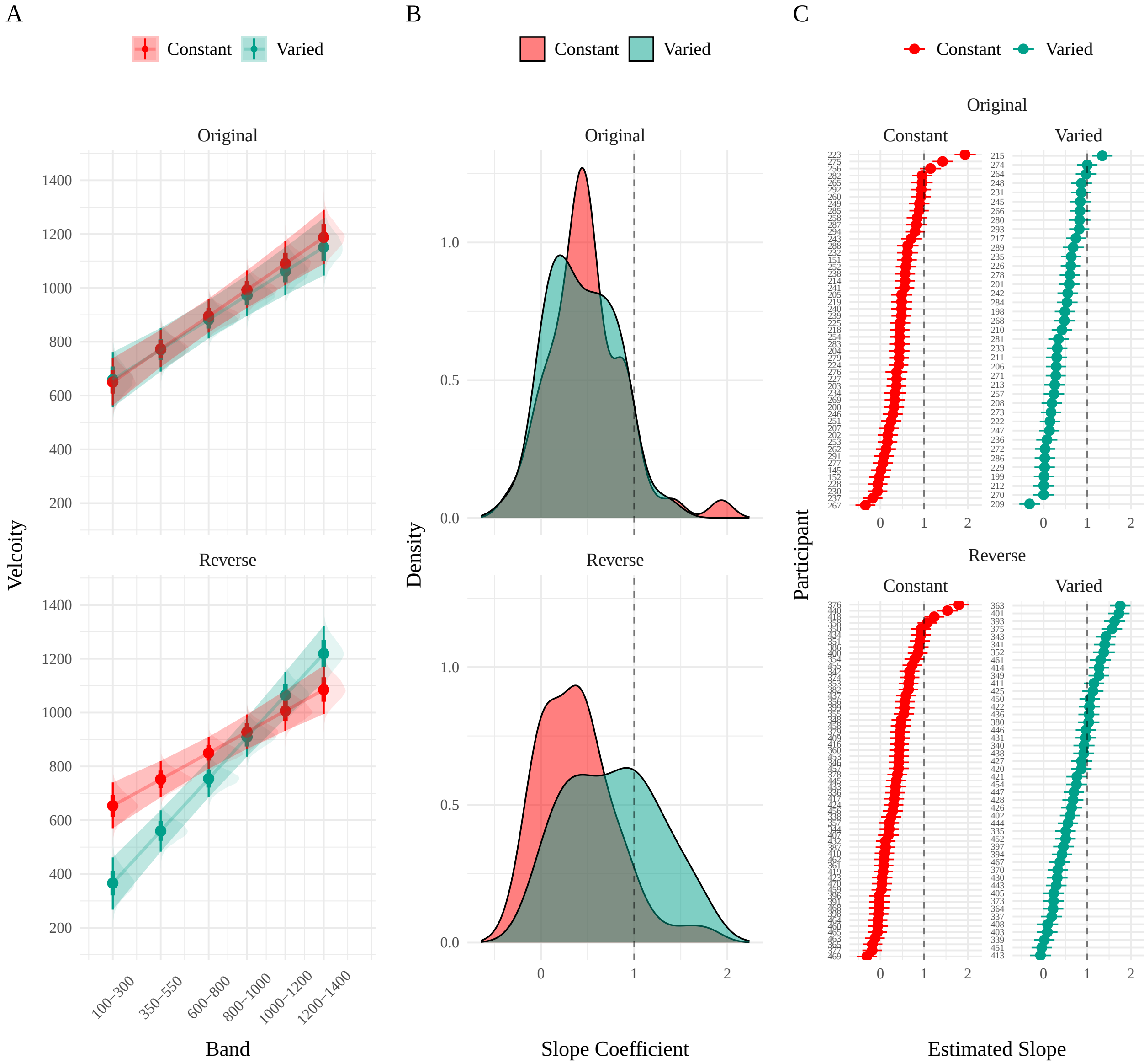


Figure 16: Conditional effect of training condition and Band. Ribbons indicate 95% HDI. The steepness of the lines serves as an indicator of how well participants discriminated between velocity bands.

Experiment 3 Summary

Computational Model

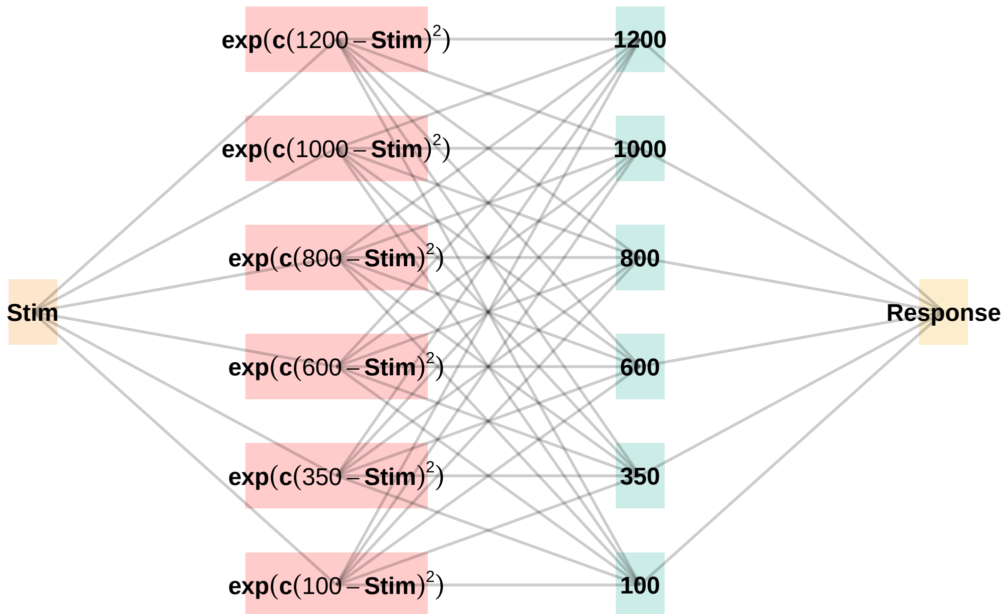


Figure 17: The Associative Learning Model (ALM). The diagram illustrates the basic structure of the ALM model as used in the present work. Input nodes are activated as a function of their similarity to the lower-boundary of the target band. The generalization parameter, c, determines the degree to which nearby input nodes are activated. The output nodes are activated as a function of the weighted sum of the input nodes - weights are updated via the delta rule.

Modeling Approach

The modeling goal is to implement 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. EXAM differs from ALM only in its response rule, as it includes a linear extrapolation mechanism for generating novel responses. 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. EXAM retrieves the two nearest training inputs, and the ALM responses associated with those inputs, and computes the slope between these two points. The slope is then used to extrapolate the response to the novel test stimulus. Because EXAM requires at least two input-output pairs to generate a response, additional assumptions were required in order for it to generate resposnes for the constant group. We assumed that participants come to the task with prior knowledge of the origin point (0,0), which can serve as a reference point necessary for the model to generate responses for the constant group. This assumption is motivated by previous function learning research (Brown & Lacroix (2017)), which through a series of manipulations of the y intercept of the underlying function, found that participants consistently demonstrated knowledge of, or a bias towards, the origin point (see Kwantes & Neal (2006) for additional evidence of such a bias in function learning tasks).

See Table 13 for a full specification of the equations that define ALM and EXAM, and Figure 17 for a visual representation of the ALM model.

Table 13: 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	$P[Y_j \mid X] = \frac{O_j(X)}{\sum_{k=1}^{M} O_k(X)}$	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	$P[X_i \mid X] = \frac{a_i(X)}{\sum_{k=1}^{M} a_k(X)}$	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	$E[Y \mid X_i] = m(X_i) + S \cdot [X - X_i]$	ALM response $m(X_i)$ adjusted by slope.

Model Fitting

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.

< fa regular lightbulb > Approximate Bayesian Computation

To estimate the parameters of ALM and EXAM, 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 models).

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 between simulated and observed data. For the present work, we follow the guidelines from previously 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 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 1 and Equation 2). For the models fit to both training and testing data, discrepancies were computed for both stages, and then averaged together.

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

$$\begin{aligned} discrep_{Train,constant}(Data_{sim}, Data_{obs}) &= \frac{1}{N_{blocks}} \sum_{j=1}^{N_{blocks}} |V_{obs,constant}^{(j)} - V_{sim,constant}^{(j)}| \\ discrep_{Train,varied}(Data_{sim}, Data_{obs}) &= \frac{1}{N_{blocks} \times 3} \sum_{j=1}^{N_{blocks}} \sum_{k=1}^{3} |V_{obs,varied}^{(j,k)} - V_{sim,varied}^{(j,k)}| \end{aligned} \tag{2}$$

The final component of our ABC implementation is the determination of an 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 simply 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 Patterns

Table 14: Models errors predicting empirical data - aggregated over all participants, posterior parameter values, and velocity bands. Note that Fit Method refers to the subset of the data that the model was trained on, while Task Stage refers to the subset of the data that the model was evaluated on.

		ALM		EXAM	
Task Stage	Fit Method	Constant	Varied	Constant	Varied
Test	Fit to Test Data	199.93	103.36	104.01	85.68
Test	Fit to Test & Training Data	216.97	170.28	127.94	144.86
Test	Fit to Training Data	467.73	291.38	273.30	297.91
Train	Fit to Test Data	297.82	2,016.01	53.90	184.00
Train	Fit to Test & Training Data	57.40	132.32	42.92	127.90
Train	Fit to Training Data	51.77	103.48	51.43	107.03

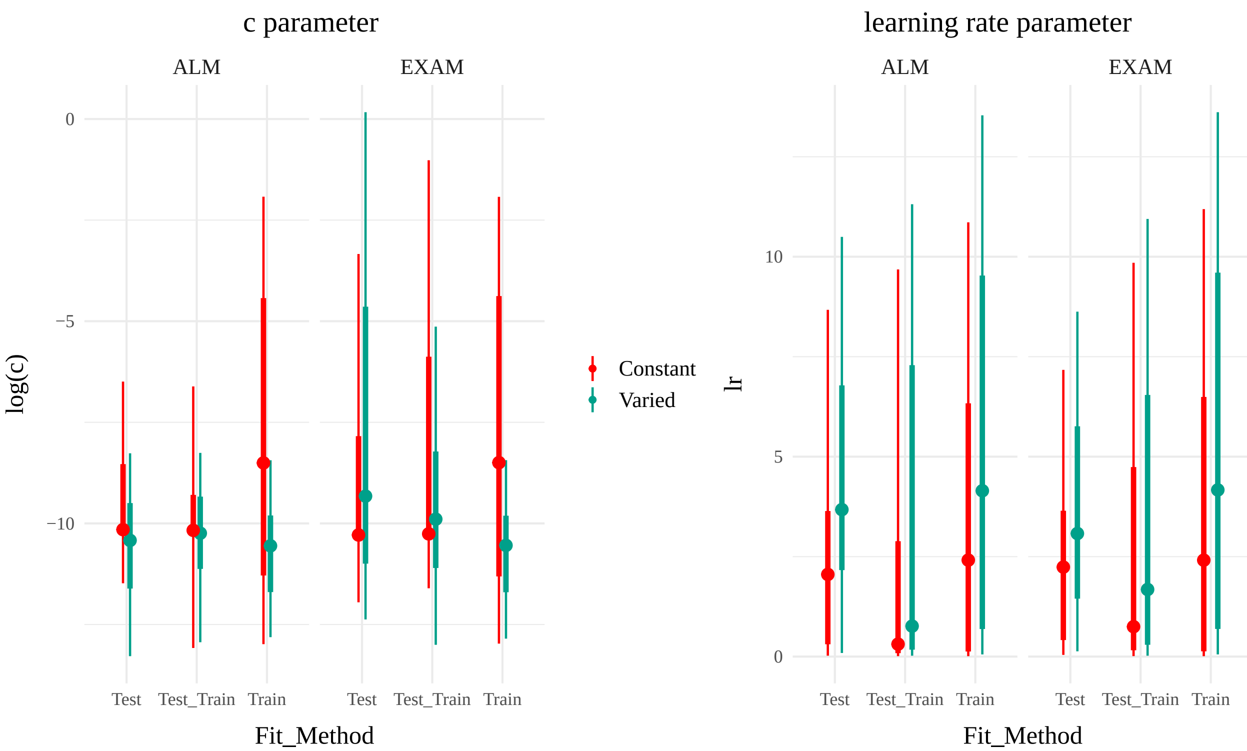


Figure 18: Posterior Distributions of c and lr parameters. Points represent median values, thicker intervals represent 66% credible intervals and thin intervals represent 95% credible intervals around the median. Note that the y axes of the plots for the c parameter are scaled logarithmically.

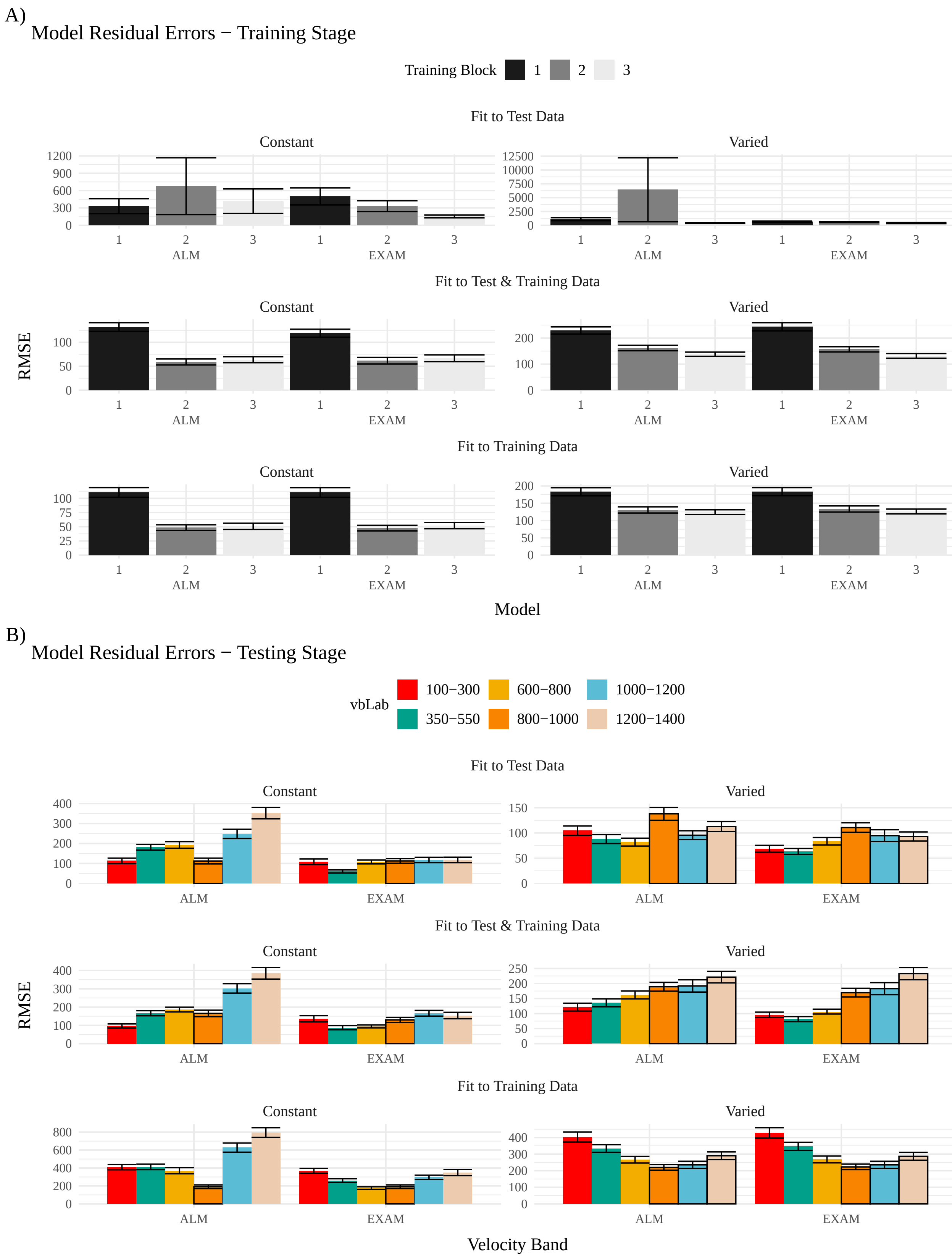


Figure 19: Model residuals for each combination of training condition, fit method, and model. Residuals reflect the difference between observed and predicted values. Lower values indicate better model fit. Note that y axes are scaled differently between facets. A) Residuals predicting each block of the training data. B) Residuals predicting each band during the testing stage. Bolded bars indicate bands that were trained, non-bold bars indicate extrapolation bands.

The posterior distributions of the c and lr parameters are shown Figure 18, and model predictions are shown alongside the empirical data in Figure 20. 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. Aggregated residuals are displayed in Figure 19. The pattern of training stage residual errors are unsurprising across the combinations of models and fitting method . Differences in training performance 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 the model predictions for the testing stage (Figure 19) also show an unsurprising pattern across fitting methods - 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). Although EXAM tends to perform better for both Constant and Varied participants (see also Figure 21), the relative advantage of EXAM is generally larger for the Constant group - a pattern consistent across all three fitting methods. The primary predictive difference between ALM and EXAM is made clear in Figure 20, which directly compares the observed data against the posterior predictive distributions for both models. Regardless of how the models are fit, only EXAM can capture the pattern where participants are able to discriminate all 6 target bands.

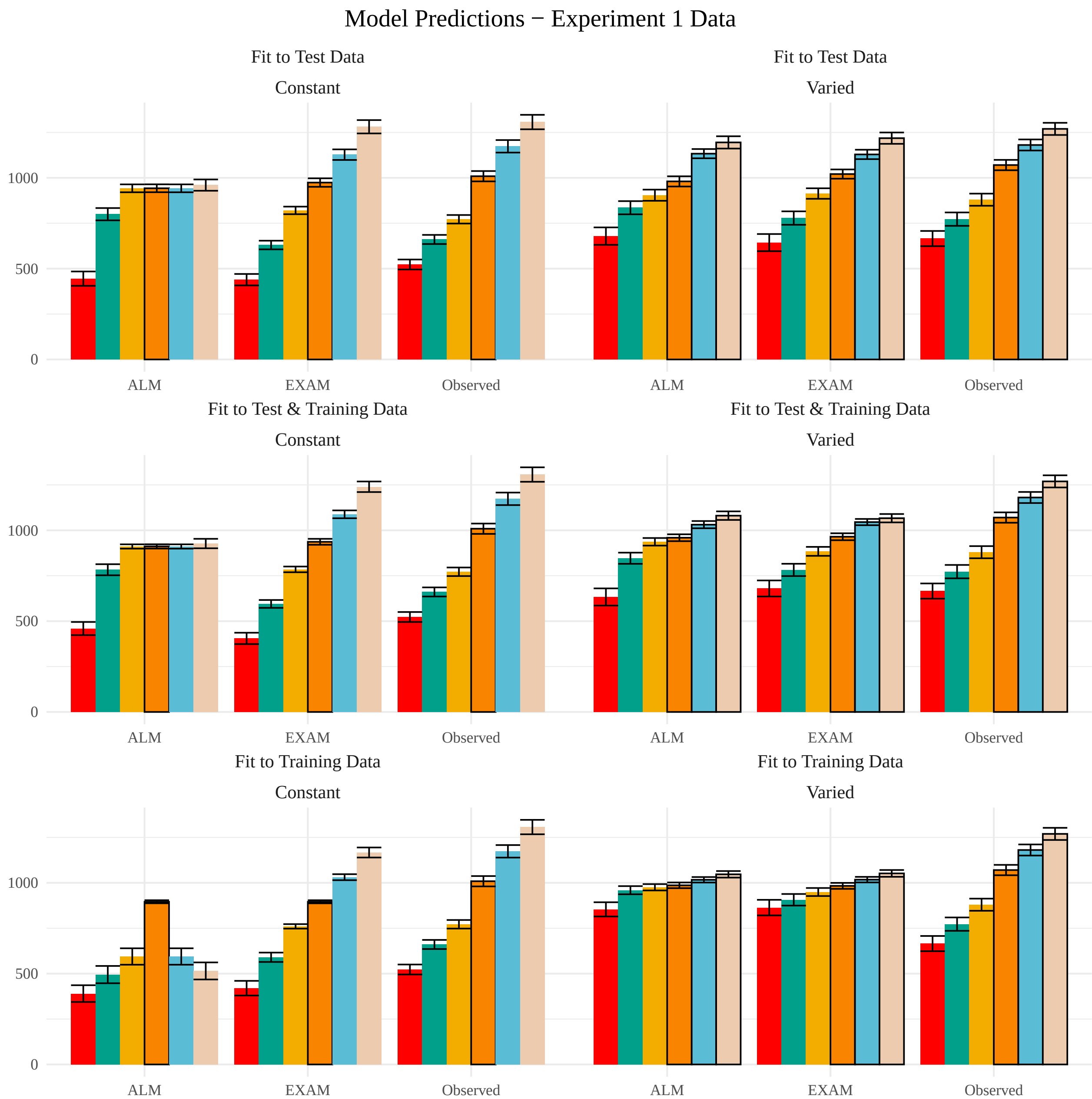


Figure 20: Empirical data and Model predictions for mean velocity across target bands. Fitting methods (Test Only, Test & Train, Train Only) - are separated across rows, and Training Condition (Constant vs. Varied) are separated by columns. Each facet contains the predictions of ALM and EXAM, alongside the observed data.

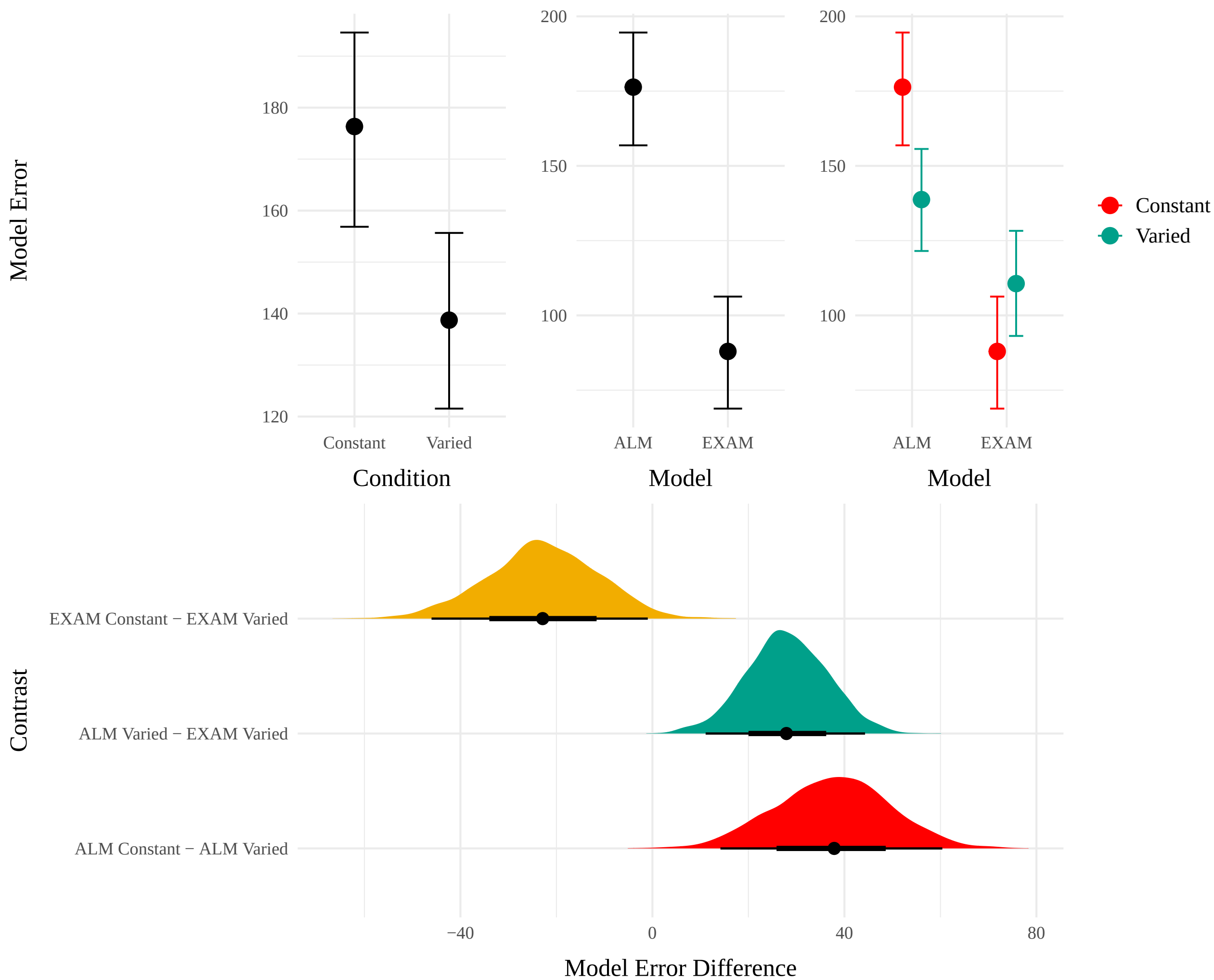


Figure 21

To quantitatively assess whether the differences in performance between models, we fit a bayesian regressions predicting the errors of the posterior predictions of each models as a function of the Model (ALM vs. EXAM) and training condition (Constant vs. Varied).

Model errors were significantly lower for EXAM (β = -37.54, 95% CrI [-60.4, -14.17], pd = 99.85%) than ALM. There was also a significant interaction between Model and Condition (β = 60.42, 95% CrI [36.17, 83.85], pd = 100%), indicating that the advantage of EXAM over ALM was significantly greater for the constant group. To assess whether EXAM predicts constant performance significantly better for Constant than for Varied subjects, we calculated the difference in model error between the Constant and Varied conditions specifically for EXAM. The results indicated that the model error for EXAM was significantly lower in the Constant condition compared to the Varied condition, with a mean difference of -22.879 (95% CrI [-46.016, -0.968], pd = 0.981).

Table 15: Models errors predicting empirical data - aggregated over all participants, posterior parameter values, and velocity bands. Note that Fit Method refers to the subset of the data that the model was trained on, while Task Stage refers to the subset of the data that the model was evaluated on.

	E2				E3			
	ALM		EXAM		ALM		EXAM	
Task Stage	Constant	Varied	Constant	Varied	Constant	Varied	Constant	Varied

Fit to Test Data								
Test	239.7	129.8	99.7	88.2	170.1	106.1	92.3	72.8
Train	53.1	527.1	108.1	169.3	70.9	543.5	157.8	212.7
Fit to Test & Training Data								
Test	266.0	208.2	125.1	126.4	197.7	189.5	130.0	128.5
Train	40.0	35.4	30.4	23.6	49.1	85.6	49.2	78.4
Fit to Training Data								
Test	357.4	295.9	305.1	234.5	415.0	298.8	295.5	243.7
Train	42.5	23.0	43.2	22.6	51.4	63.8	51.8	65.3

Figure 22: Empirical data and Model predictions from Experiment 2 and 3 for the testing stage. Observed data is shown on the right. Bolded bars indicate bands that were trained, non-bold bars indicate extrapolation bands.

Table 16: Results of Bayesian Regression models predicting model error as a function of Model (ALM vs. EXAM), Condition (Constant vs. Varied), and the interaction between Model and Condition. The values represent the estimate coefficient for each term, with 95% credible intervals in brackets. The intercept reflects the baseline of ALM and Constant. The other estimates indicate deviations from the baseline for the EXAM mode and varied condition. Lower values indicate better model fit.

Experiment	Term	Estimate	Credible Interval		pd
			95% CrI Lower	95% CrI Upper	
Experiment 1					
Exp 1	Intercept	176.30	156.86	194.59	1.00
Exp 1	ModelEXAM	−88.44	−104.51	−71.81	1.00
Exp 1	conditVaried	−37.54	−60.40	−14.17	1.00
Exp 1	ModelEXAM:conditVaried	60.42	36.17	83.85	1.00
Experiment 2					
Exp 2	Intercept	245.87	226.18	264.52	1.00
Exp 2	ModelEXAM	−137.73	−160.20	−115.48	1.00
Exp 2	conditVaried	−86.39	−113.52	−59.31	1.00
Exp 2	ModelEXAM:conditVaried	56.87	25.26	88.04	1.00
Experiment 3					
Exp 3	Intercept	164.83	140.05	189.44	1.00
Exp 3	ModelEXAM	−65.66	−85.97	−46.02	1.00
Exp 3	conditVaried	−40.61	−75.90	−3.02	0.98
Exp 3	bandOrderReverse	25.47	−9.34	58.68	0.93
Exp 3	ModelEXAM:conditVaried	41.90	11.20	72.54	0.99
Exp 3	ModelEXAM:bandOrderReverse	−7.32	−34.53	21.05	0.70
Exp 3	conditVaried:bandOrderReverse	30.82	−19.57	83.56	0.88
Exp 3	ModelEXAM:conditVaried:bandOrderReverse	−60.60	−101.80	−18.66	1.00

Model Fits to Experiment 2 and 3. Data from Experiments 2 and 3 were fit to ALM and EXAM in the same manner as Experiment1 . For brevity, we only plot and discuss the results of the "fit to training and testing data" models - results from the other fitting methods can be found in the appendix. The model fitting results for Experiments 2 and 3 closely mirrored those observed in Experiment 1. The Bayesian regression models predicting model error as a function of Model (ALM vs. EXAM), Condition (Constant vs. Varied), and their interaction (see Table 16) revealed a consistent main effect of Model across all three experiments. The negative coefficients for the ModelEXAM term (Exp 2: β = -86.39, 95% CrI -113.52, -59.31, pd = 100%; Exp 3: β = -40.61, 95% CrI -75.9, -3.02, pd = 98.17%) indicate that EXAM outperformed ALM in both experiments. Furthermore, the interaction between Model and Condition was significant in both Experiment 2 (β = 56.87, 95% CrI 25.26, 88.04, pd = 99.98%) and Experiment 3 (β = 41.9, 95% CrI 11.2, 72.54, pd = 99.35%), suggesting that the superiority of EXAM over ALM was more pronounced for the Constant group compared to the Varied group, as was the case in Experiment 1. Recall that Experiment 3 included participants in both the original and reverse order conditions - and that this manipulation interacted with the effect of training condition. We thus also controleld for band order in our Bayesian Regression assessing the relative performance of EXAM and ALM in Experiment 3. There was a significant three way interaction between Model, Training Condition, and Band Order (β = -60.6, 95% CrI -101.8, -18.66, pd = 99.83%), indicating that the relative advantage of EXAM over ALM was only more pronounced in the original order condition, and not the reverse order condition (see Figure 23).

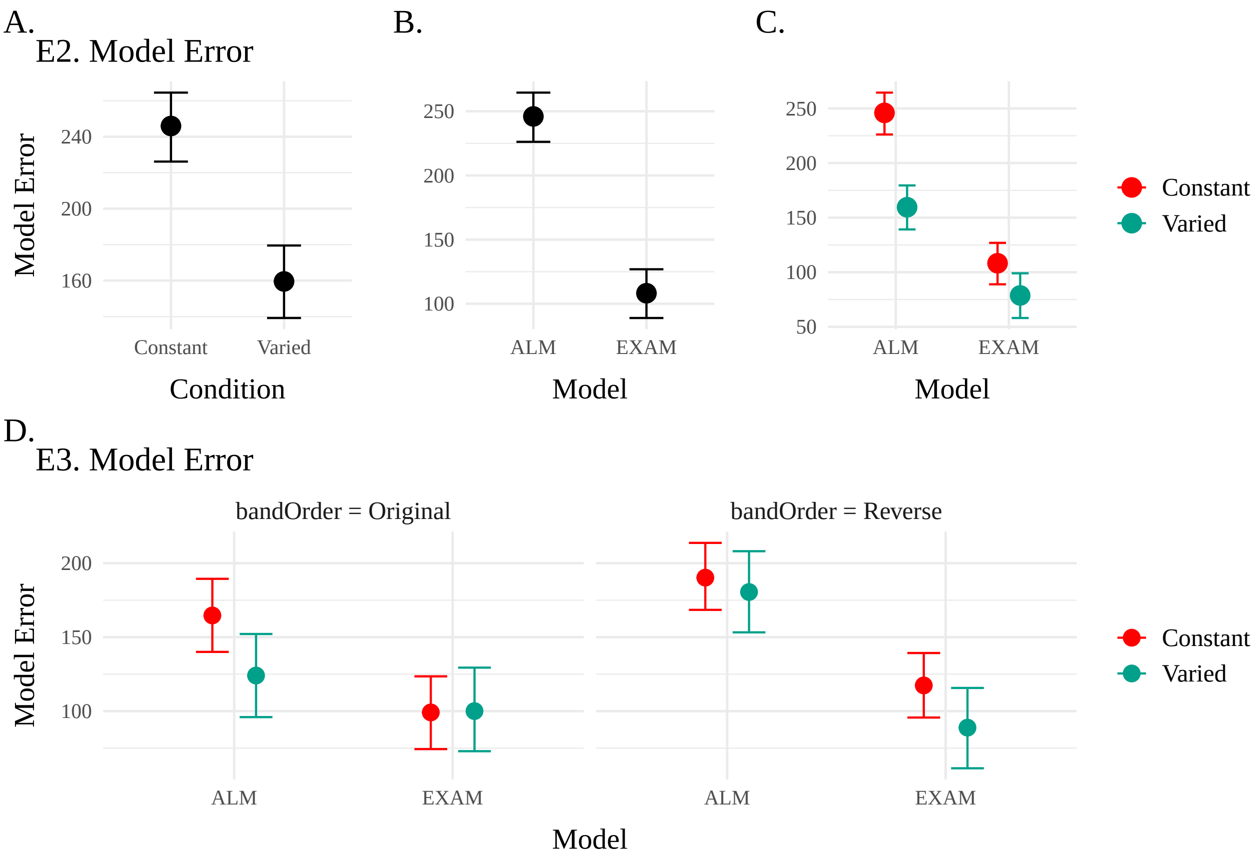


Figure 23: Conditional effects of Model (ALM vs EXAM) and Condition (Constant vs. Varied) on Model Error for Experiment 2 and 3 data. Experiment 3 also includes a control for the order of training vs. testing bands (original order vs. reverse order).

Computational Model Summary. Across the model fits to all three experiments, we found greater support for EXAM over ALM (negative coefficients on the ModelEXAM term in Table 16), and moreover that the constant participants were disproportionately well described by EXAM in comparison to ALM (positive coefficients on ModelEXAM:conditVaried terms in Table 16). This pattern is also clearly depicted in Figure 24, which plots the difference in model errors between ALM and EXAM for each individual participant. Both varied and constant conditions have a greater proportion of subjects better fit by EXAM (positive error differences), with the magnitude of EXAM's advantage visibly greater for the constant group. It also bears mention that numerous participants were better fit by ALM, or did not show a clear preference for either model. A subset of these participants are shown in Figure 25.

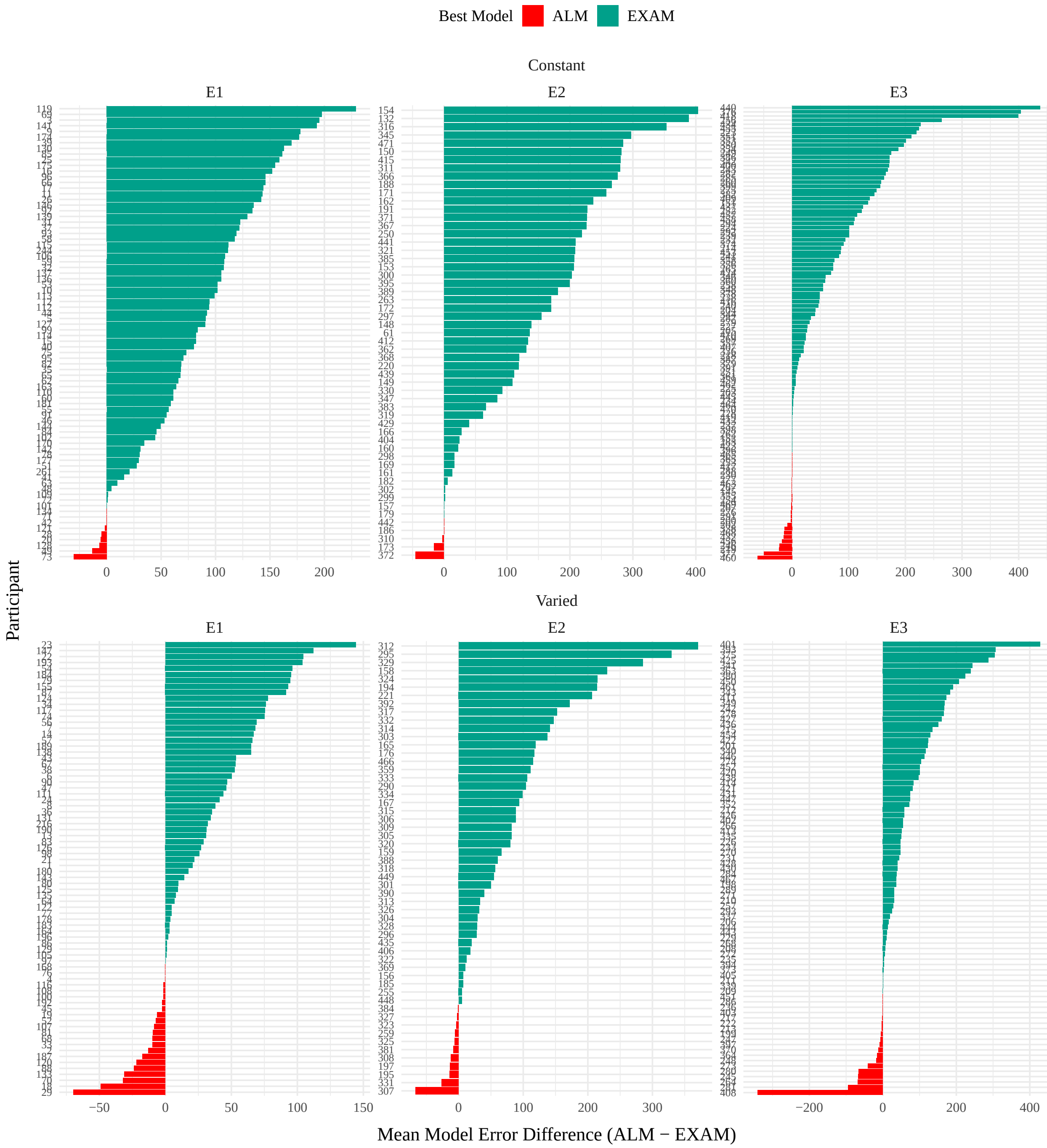


Figure 24: Difference in model errors for each participant, with models fit to both train and test data. Positive values favor EXAM, while negative values favor ALM.

Figure 25: 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. Bolded bars indicate bands that were trained, non-bold bars indicate extrapolation bands.

General Discussion

Experimental Result Summary

Across three experiments, we investigated the impact of training variability on learning and extrapolation in a visuomotor function learning task. In Experiment 1, participants in the varied training condition, who experienced a wider range of velocity bands during training, showed lower accuracy at the end of training compared to those in the constant training condition.

Crucially, during the testing phase, the varied group exhibited significantly larger deviations from the target velocity bands, particularly for the extrapolation bands that were not encountered during training. The varied group also showed less discrimination between velocity bands, as evidenced by shallower slopes when predicting response velocity from target velocity band.

Experiment 2 extended these findings by reversing the order of the training and testing bands. Similar to Experiment 1, the varied group demonstrated poorer performance during both training and testing phases. However, unlike Experiment 1, the varied group did not show a significant difference in discrimination between bands compared to the constant group.

In Experiment 3, we provided only ordinal feedback during training, in contrast to the continuous feedback provided in the previous experiments. Participants were assigned to both an order condition (original or reverse) and a training condition (constant or varied). The varied condition showed larger deviations at the end of training, consistent with the previous experiments. Interestingly, there was a significant interaction between training condition and band order, with the varied condition showing greater accuracy in the reverse order condition. During testing, the varied group once again exhibited larger deviations, particularly for the extrapolation bands. The reverse order conditions showed smaller deviations compared to the original order conditions. Discrimination between velocity bands was poorer for the varied group in the original order condition, but not in the reverse order condition.

All three of our experiments yielded evidence that varied training conditions produced less learning by the end of training, a pattern consistent with much of the previous research on the influence of training variability (Catalano & Kleiner, 1984; Soderstrom & Bjork, 2015; Wrisberg et al., 1987). The sole exception to this pattern was the reverse order condition in Experiment 3, where the varied group was not significantly worse than the constant group. Neither the varied condition trained with the same reverse-order items in Experiment 2, nor the original-order varied condition trained with ordinal feedback in Experiment 3 were able to match the performance of their complementary constant groups by the end of training, suggesting that the relative success of the ordinal-reverse ordered varied group cannot be attributed to item or feedback effects alone.

Our findings also diverge from the two previous studies to cleanly manipulate the variability of training items in a function learning task (DeLosh et al., 1997; van Dam & Ernst, 2015), although the varied training condition of van Dam & Ernst (2015) also exhibited less learning, neither of these previous studies observed any difference between training conditions in extrapolation to novel items. Like DeLosh et al. (1997) , our participants exhibited above chance extrapolation/discrimination of novel items, however they observed no difference between any of their three training conditions. A noteworthy difference difference between our studies is that DeLosh et al. (1997) trained participants with either 8, 20, or 50 unique items (all receiving the same total number of training trials). These larger sets of unique items, combined with the fact that participants achieved near ceiling level performance by the end of training - may have made it more difficult to observe any between-group differences of training variation in their study. van Dam & Ernst (2015) 's variability manipulation was more similar to our own, as they trained participants with either 2 or 5 unique items. However, although the mapping between their input stimuli and motor responses was technically linear, the input dimension was more complex than our own, as it was defined by the degree of "spikiness" of the input shape. This entirely arbitrary mapping also would have preculded any sense of a "0" point, which may partially explain why neither of their training conditions were able to extrapolate linearly in the manner observed in the current study or in DeLosh et al. (1997).

Modeling Summary EXAM is the best model for both groups, but EXAM does relatively good at accounting for the constant group. May have seemed counterintuitive, if one assumed that multiple, varied, examples were necessary to extract a rule. But, EXAM is not a conventional rule model - it doesn't require explictly abstract of a rule, but rather the rule-based response occurs during retrieval. The constant groups formation of a single, accurate, input-output association, in combination with the usefulness of the zero point, may have been sufficient for EXAM, and the constant group, to perform well.

One concern may have been that the assumption of participants making use of the zero point turned the extrapolation problem into an interpolation problem - however this concern is ameliorated by the consistency of the results across both the original and reverse order conditions.

- why does Constant do better
  - kind of task that permits for prior knowledge about 0
  - learning - end of training
- what does it suggest that the constant group was disproportionately well explained by the EXAM model?
-

Limitations - amount of training - constant group always having more experience at nearest position

- constant group having more extrapolation items
- only training and extrapolation - no interpolation items - so harder to make claims about extrapolation specifically, as opposed to generalization in general.
- no mechanism to account for sequence effects in models
- more rigorous model comparison to account for exams greater complexity
- knowledge of function

Comparison to Project 1

Differences between the tasks

There are a number of differences between Project 1's Hit The Target (HTT), and Project 2's Hit The Wall (HTW) tasks.

- Task Space Complexity: In HTW, the task space is also almost perfectly smooth, at least for the continuous feedback subjects, if they throw 100 units too hard, they'll be told that they were 100 units too hard. Whereas in HTT, it was possible to produce xy velocity combinations that were technically closer to the empirical solution space than other throws, but which resulted in worse feedback due to striking the barrier.

- Perceptual Distinctiveness: HTT offers perceptually distinct varied conditions that directly relate to the task's demands, which may increase the sallience between training positions encounted by the varied group. In contrast, HTW's varied conditions differ only in the numerical values displayed, lacking the same level of perceptual differentiation. Conversely in HTW, the only difference between conditions for the varied group are the numbers displayed at the top of the screen which indicate the current target band(e.g. 800-1000, or 1000-1200)

- In HTW, our primary testing stage of interest has no feedback, whereas in HTT testing always included feedback (the intermittent testing in HTT expt 1 being the only exception). Of course, we do collect testing with feedback data at the end of HTW, but we haven't focused on that data at all in our modelling work thus far. It's also interesting to recall that the gap between varied and constant in HTW does seem to close substantially in the testing-with-feedback stage. The difference between no-feedback and feedback testing might be relevant if the benefits of variation have anything to do with improving subsequent learning (as opposed to subsequent immediate performance), OR if the benefits of constant training rely on having the most useful anchor, having the most useful anchor might be a lot less helpful if you're getting feedback from novel positions and can thus immediately begin to form position-specific anchors for the novelties, rather than relying on a training anchor.

- HTW and HTT both have a similar amount of training trials (~200), and thus the constant groups acquire a similar amount of experience with their single position/velocity in both experiments. However, the varied conditions in both HTT experiments train on 2 positions, whereas the varied group in HTW trains on 3 velocity bands. This means that in HTT the varied group gets half as much experience on any one position as the constant group, and in HTW they only get 1/3 as much experience in any one position. There are likely myriad ways in which this might impact the success of the varied group regardless of how you think the benefits of variation might be occurring, e.g. maybe they also need to develop a coherent anchor, maybe they need more experience in order to extract a function, or more experience in order to properly learn to tune their c parameter.

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

Berniker, M., Mirzaei, H., & Kording, K. P. (2014). The effects of training breadth on motor generalization. *Journal of Neurophysiology*, *112*(11), 2791–2798. https://doi.org/10.1152/jn.00615.2013

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

Braithwaite, D. W., & Goldstone, R. L. (2015). Effects of Variation and Prior Knowledge on Abstract Concept Learning. *Cognition and Instruction*, *33*(3), 226–256. https://doi.org/10.1080/07370008.2015.1067215

Braun, D. A., Aertsen, A., Wolpert, D. M., & Mehring, C. (2009). Motor Task Variation Induces Structural Learning. *Current Biology*, *19*(4), 352–357. https://doi.org/10.1016/j.cub.2009.01.036

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

Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language*, *126*, 104352. https://doi.org/10.1016/j.jml.2022.104352

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

Catalano, J. F., & Kleiner, B. M. (1984). Distant Transfer in Coincident Timing as a Function of Variability of Practice. *Perceptual and Motor Skills*, *58*(3), 851–856. https://doi.org/10.2466/pms.1984.58.3.851

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

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory & Cognition*, *29*(8), 1165–1175. https://doi.org/10.3758/BF03206386

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

Guo, J.-P., Yang, L.-Y., & Ding, Y. (2014). Effects of example variability and prior knowledge in how students learn to solve equations. *European Journal of Psychology of Education*, *29*(1), 21–42. https://www.jstor.org/stable/43551124

Hu, M., & Nosofsky, R. M. (2024). High-variability training does not enhance generalization in the prototype-distortion paradigm. *Memory & Cognition*, 1–16. https://doi.org/10.3758/s13421-023-01516-1

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

Kwantes, P. J., & Neal, A. (2006). Why people underestimate y when extrapolating in linear functions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1019–1030. https://doi.org/10.1037/0278-7393.32.5.1019

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

Perry, L. K., Samuelson, L. K., Malloy, L. M., & Schiffer, R. N. (2010). Learn Locally, Think Globally: Exemplar Variability Supports Higher-Order Generalization and Word Learning. *Psychological Science*, *21*(12), 1894–1902. https://doi.org/10.1177/0956797610389189

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363. https://doi.org/10.1037/h0025953

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*, S1364661322000651. https://doi.org/10.1016/j.tics.2022.03.007

Roller, C. A., Cohen, H. S., Kimball, K. T., & Bloomberg, J. J. (2001). Variable practice with lenses improves visuomotor plasticity. *Cognitive Brain Research*, *12*(2), 341–352. https://doi.org/10.1016/S0926-6410(01)00077-5

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

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review*, *82*(4), 225–260. https://doi.org/10.1037/h0076770

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

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176–199. https://doi.org/10.1177/1745691615569000

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

van Dam, L. C. J., & Ernst, M. O. (2015). Mapping Shape to Visuomotor Mapping: Learning and Generalisation of Sensorimotor Behaviour Based on Contextual Information. *PLOS Computational Biology*, *11*(3), e1004172. https://doi.org/10.1371/journal.pcbi.1004172

Van Rossum, J. H. A. (1990). Schmidt's schema theory: The empirical base of the variability of practice hypothesis. *Human Movement Science*, *9*(3-5), 387–435. https://doi.org/10.1016/0167-9457(90)90010-B

Wrisberg, C. A., Winter, T. P., & Kuhlman, J. S. (1987). The Variability of Practice Hypothesis: Further Tests and Methodological Discussion. *Research Quarterly for Exercise and Sport*, *58*(4), 369–374. https://doi.org/10.1080/02701367.1987.10608114