A Mixture of Experts in Associative Generalization

Jessica C. Lee
Peter F. Lovibond
Brett K. Hayes
School of Psychology, UNSW Sydney, Australia

Stephan Lewandowsky
School of Psychological Science, University of Bristol, UK

Abstract

After learning that one stimulus predicts an outcome (e.g., an aqua-colored rectangle leads to shock) and a very similar stimulus predicts no outcome (e.g., a slightly greener rectangle leads to no shock), some participants generalize the predictive relationship on the basis of physical similarity to the predictive stimulus, while others generalize on the basis of the relational difference between the two stimuli (e.g., "higher likelihood of shock for bluer stimuli"). To date, these individual differences in generalization rules have remained unexplored in associative learning. Here, we present evidence that a given *individual* simultaneously entertains belief in both "similarity" and "relational" rules, and generalizes using a mixture of these strategies. Using a "mixture of experts" modelling framework constrained by participants self-reported rule beliefs, we show that considering multiple rules predicts generalization gradients better than a single rule, and that generalization behavior is better described as switching between, rather than averaging over, different rules.

Keywords: generalization; associative learning; rules; mixture of experts models; peak shift

Introduction

Pavlov (1927) noted that after conditioning where a conditioned stimulus (CS) was repeatedly paired with an outcome (+), conditioned responses were emitted not only to the trained CS+, but also to other stimuli sharing properties with the CS+. This ability to generalize learning to novel situations and stimuli is fundamental to human and non-human animal behavior. Understanding the theoretical processes behind generalization is therefore important to explain how behaviors can be adaptive or sometimes also maladaptive (overgeneralization of fear or threat; see Lissek, 2012).

In associative learning, a typical generalization experiment consists of an initial training phase where a visual CS+ (e.g., an aqua-colored rectangle presented on a screen) is paired with an outcome (e.g., electric shock). Participants are then presented with generalization stimuli (GS) varying along a continuous stimulus dimension (e.g., hue, size) and the amount of responding (e.g., conditioned physiological responses or explicit predictive ratings of the outcome) is measured. Plotting these responses along the dimension produces a generalization *gradient*. The typical form of the gradient after training with a single CS+ is symmetrical, peaked, and roughly Gaussian in shape (Ghirlanda & Enquist, 2003; Shepard, 1987). Gradients with this shape are typically interpreted as generalization based on physical similarity to the trained CS+ (see the "similarity" gradient in Figure 1).

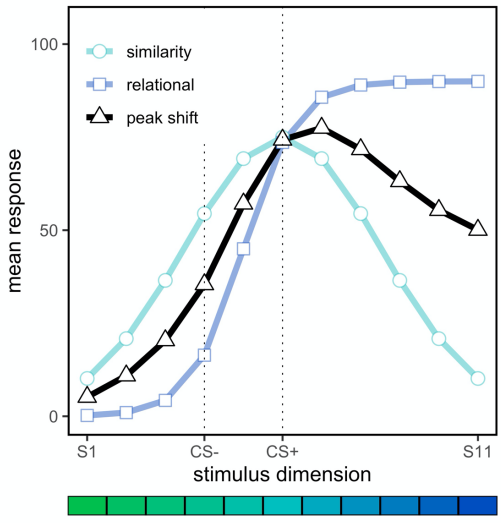

Figure 1: Idealized similarity-based, relational-based, and peak-shifted generalization gradients following differential training with an aqua (bluey-green) CS+ and slightly greener CS-. Note that the peak-shifted gradient is the average of the similarity and relational gradients.

A generalization phenomenon that has received a considerable amount of attention in the conditioning literature is the *peak shift* effect (Hanson, 1959; see Purtle, 1973 for a review). Peak shift describes a phenomenon where responding is maximal at a novel stimulus, rather than at the CS+ encountered during training, which would be expected under similarity-based responding. Peak shift is obtained using differential training with a CS+ and CS- that differ slightly within a dimension (e.g., the CS+ is slightly bluer than the CS-), and can be inferred if the gradient peak shifts to a novel stimulus on the opposite side of the CS+ to the CS- (see the peak shift gradient in Figure 1). Although appearing to be evidence of relational responding, associative models predict peak shift as arising from the interaction between excitation from the CS+ and inhibition from the CS- (e.g., Blough, 1975; Ghirlanda & Enquist, 1998; McLaren &

Mackintosh, 2002). Therefore, demonstrations of peak shift in humans are often explained in associative terms (e.g., Lee & Livesey, 2018; Livesey & McLaren, 2009; Wills & Mackintosh, 1998).

A recent study from our lab (Lee et al., 2018), however, suggests a different explanation - that peak shift might instead be explained by individual differences in self-generated hypotheses about how the learned association generalizes along the dimension (i.e., participants' generalization rules). Note that we use the term "rule" to describe these explicit generalization strategies, but are agnostic as to whether they are generated by different mechanisms (see Pothos, 2005). Lee et al. showed that after differential training, there were two major rules reported by participants. Some participants reported generalizing on the basis of the physical similarity to the CS+ (similarity subgroup) whereas others reported generalizing on the basis of the relational difference between the CS+ and CS- (e.g., "bluer than"; relational subgroup). The form of the generalization gradients was consistent with these self-reported rules, with a monotonically increasing linear/sigmoidal gradient shown in the relational subgroup, and a peaked, symmetrical gradient shown in the similarity subgroup (see Figure 1 for idealized gradients). Critically, these distinct gradients formed a peak-shifted gradient when averaged (Figure 1), demonstrating that at the aggregate, a peak shift effect in humans can be explained through mixtures of generalization rules between participants.

The mixture-of-rules explanation presented by Lee et al. (2018) assumes that each participant derives and uses a single rule. This account can explain a peak-shifted gradient displayed at the aggregate-level, but it cannot account for peak shift displayed at the level of a subgroup where all participants have reported the same rule. Indeed, in both experiments of Lee et al., the gradient in the similarity subgroup was numerically (but not significantly) peak-shifted, suggesting that a similar mixture of similarity and relational learning might be occurring *within* a subgroup or even within an *individual*. This idea is supported by evidence that individuals can show both similarity- and rule-based generalization when learning patterning discriminations (Shanks & Darby, 1998), categories (Nosofsky & Palmeri, 1998; Little & McDaniel, 2015; Thibault et al., 2018), and continuous functions (DeLosh et al., 1997). The aim of the current study was to test the premise that participants use multiple rules in associative generalization and extend the mixture-of-rules explanation of generalization proposed by Lee et al. (2018) to a formal model. Specifically, we asked:

1. Do participants entertain belief in multiple generalization rules during differential training?
2. Does consideration of multiple rules predict generalization better than a single rule?
3. How do the rules combine to determine generalization performance?

To answer these questions, we report data from an experiment and compare three "mixture of experts" models – a single expert model, an averaging model, and a choice model – in their ability to predict the empirical gradients.

Experiment

In this experiment, we presented participants with standard differential training using the same stimulus dimension (colored shapes varying between green-blue) as Lee et al. (2018).

Method

Participants 100 Mechanical Turk workers (34 female, *M* age = 34.2, *SD* age = 9.7) participated in exchange for payment (USD$2 for a 12 minute task). Workers had to have completed 500 Human Intelligence Tasks and have an approval rate > 90% in order to be eligible for the task.

Procedure Participants completed an online computer task where their task was to predict whether a machine would give a hypothetical "Mr. X" electric shocks (no actual shocks were administered). Participants were told that different symbols would appear on the machine, and that they should use those symbols to predict whether a shock would be delivered.

The stimuli were 11 colored rectangles (S1-S11) varying between green and blue (.4 and .6 hue on the HSB scale, keeping saturation and brightness constant at 1 and .75 respectively, see Figure 2).

Figure 2: Stimuli presented in the experiment for a participant allocated to the green-blue counterbalancing condition.

All participants received differential training where the CS+ was the midpoint (S6) on the green-blue dimension and the CS- (S4) was either slightly greener or slightly bluer than the CS+ (counterbalanced; see Figure 2). The CS+ was followed by the outcome 75% of the time and the CS- was never followed by the outcome. Participants received 12 CS+ and 12 CS- trials randomized in 3 blocks of 4 trials of each. The first CS+ trial of each block was always followed by the outcome.

On each trial, participants were presented with a symbol in the middle of the screen and asked to make a prediction about the outcome for Mr. X. They were told to press the L key if they predicted a shock, or press the A key if they predicted no shock. After making a response, participants were shown feedback about the actual outcome (shock or no shock) for 5s alongside the stimulus, followed by a blank 2s inter-trial-interval (ITI).

After the training phase, participants were told that for the following phase they would not be receiving feedback about the shock outcome. Each of the 11 test stimuli were presented once, in randomized order. On each trial, participants were presented with the symbol and asked "What is the likelihood of this symbol leading to a SHOCK?". Participants made a rating on a visual analogue scale ranging from "Definitely NO SHOCK" to "Definitely SHOCK". Participants clicked

"Continue" once they were finished making their rating. All test ratings ranged between 0-100.

To assess the rules participants generated, they were first presented with a 3 alternative forced-choice (3AFC) question and asked to select the option they thought was most true**.** Participants chose from three options: a similarity rule ("The more SIMILAR the symbol to an AQUA (greeny-blue) color, the HIGHER the likelihood of shock"), and two relational rules ("The GREENER/BLUER the symbol, the HIGHER the likelihood of shock"). One of the relational rules was consistent, and the other was inconsistent, with the training contingencies. We have previously shown that these 3 options are sufficient to capture the range of rules reported in a free-response question after differential training (Lee et al., 2018).

Participants then rated their degree of belief in the same three rules. The rules appeared on the same screen with a rating scale below each ranging from "Definitely FALSE" to "Definitely TRUE" (0-100). Each rating was independent of the others. Participants were required to make all three ratings before continuing.

Results

Participants were excluded from analysis if they did not pass the training criterion (accuracy > 50% in last block of training) or if they indicated that they were colorblind. We also excluded participants who chose the relational rule stated in the opposite direction to the training contingencies. After exclusions, 78 participants remained.

Figure 3a shows acquisition over training trials, and Figure 3b shows the overall generalization gradient. Note that the mean gradient is slightly peak-shifted, with the highest predictions of the outcome for the stimulus adjacent to the CS+. For brevity, we will not report any analyses for the training data nor the group-level gradient.

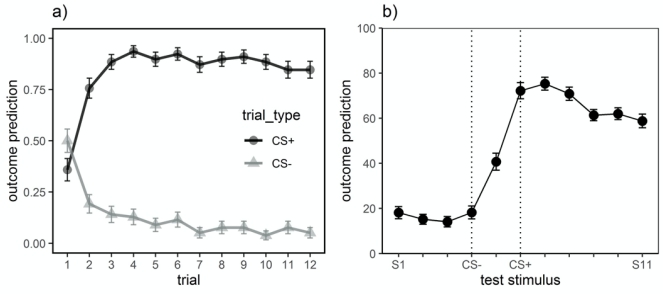

Figure 3: a) Acquisition over training trials, b) Mean generalization gradient

Figure 4 shows the joint distribution of rule beliefs scaled to range between 0-1. Note that we coded relational belief as the belief rating given to the rule that was in the consistent direction (greener/bluer) with each participants' training contingencies. From the figure, it is clear that the majority of participants have a moderate-to-high degree of belief in both similarity and relational rules, and that generally, participants gave higher belief ratings to the rule that they selected in the forced-choice question. Despite a negative correlation between relational and similarity beliefs, r = -.473, $t(76)$ = 4.68, p < .001, it seems that many participants hold some level of belief in relational *and* similarity generalization rules following differential training.

Participants' responses on the initial forced-choice question were consistent with their subsequent belief ratings. Participants who chose the similarity option gave higher belief ratings to the similarity (M=84.8, SE=2.8) than to the relational (M=49.0, SE=3.7) and inconsistent relational (M=24.0, SE=4.1) rules, and those who chose the relational option gave higher ratings to the relational (M=79.5, SE=3.7) than to the similarity (M=50.2, SE=5.4) and inconsistent relational (M=23.8, SE=4.7) rules.

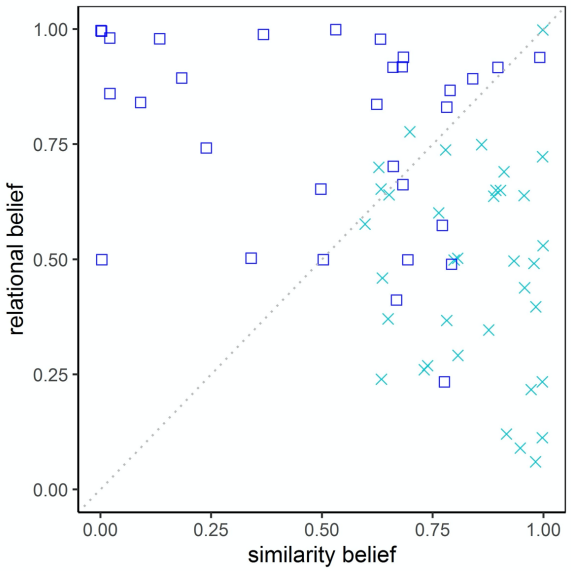

Figure 4: Scatterplot of rule beliefs. Color/shape represent whether participants chose relational (blue squares) or similarity (aqua crosses) in the forced-choice question. The grey dotted line represents equal belief for both rules.

Modelling

The critical questions are whether participants *use* multiple rules in generalization, and if so, how? To answer these questions, we adopted a "mixture-of-experts" modelling framework, similar to the Population of Linear Experts Model (POLE) in function learning (Kalish et al., 2004). This framework conceptualizes generalization as the result of multiple underlying *experts*, which are hypothetical learners generalizing in a specific way. For the current differential training procedure, we assumed the existence of three experts:

1. A similarity expert, that generalizes according to a Gaussian function (with 3 parameters: $mean$, $standard\ deviation$, and $height$), weighted by S
2. A relational expert, that generalizes according to a logistic function (with 3 parameters: $location$, $scale$, and $maximum$), weighted by R

3. A guessing expert, that responds at the midpoint of the scale (50), weighted by G

The inclusion of the guessing expert was to ensure that for participants who gave relational and similarity beliefs in similar ratios, the model treated participants with higher degrees of belief (low guessing parameter) differently to participants with lower degrees of belief (high guessing parameter).

The degree to which each expert influences behavior is determined by the relative weights of S, R, and G. For example, a weight of 1 for R would mean that behavior is driven entirely by the relational expert, and would follow a logistic function. A key point of departure from the POLE model (Kalish et al., 2004) is that we used participants' *empirical* beliefs in each rule (scaled to range between 0-1) to determine the weights of the similarity and relational experts. For the guessing expert, we calculated the weight for each participant using the equation:

$$G = (1 - S)(1 - R)$$

The weights were normalized so that they added to 1 by dividing each weight by the sum of the three weights. The expert weights were calculated for each participant and entered into the models as fixed parameters.

We chose a Gaussian function for the similarity expert as it captures the shape of generalization gradients in animals and humans (Ghirlanda & Enquist, 2003; Shepard, 1987). The Gaussian function has three parameters: the $mean$, which is the location along the dimension where the gradient peaks, the standard deviation (SD), which controls the width of the gradient, and the $height$ of the gradient, which is the height of the peak. The Gaussian function is given by:

$$y = height * e^{\frac{(x-mean)^2}{2SD^2}}$$

We chose a logistic function for the relational expert as many instances of relational-based generalization follow a sigmoidal shape (e.g., Lee et al., 2018; Livesey & McLaren, 2009). The logistic function also has three parameters: the location (loc), which is the location of the midpoint, the $scale$, which controls the steepness of the curve, and the maximum (max), which is the maximum height of the curve. The logistic function is given by:

$$y = \frac{max}{1 + e^{-scale(x-loc)}}$$

A strength of the model is that it estimates the parameters of the underlying Gaussian and logistic functions (mean, SD, height, location, scale, maximum). We implemented the model in a hierarchical Bayesian framework, estimating subject-level and group-level parameters and thus enabling us to capture individual differences in the underlying expert functions.

We compared three different models:

1) A *single expert* model: assumes that participants respond according to a single expert (the similarity or relational experts were given a weight of 1, using responses on the forced-choice question)
2) An *averaging* model: assumes that participants respond according to the weighted average of the relational, similarity, and guessing experts
3) A *choice* model: assumes that responding on each trial is determined by a probabilistic choice (using the weights) between the three experts

We assumed that the 6 key group-level parameters were drawn from Gaussian (normal) distributions (see Table 1). We calibrated the priors by dividing the data into halves (based on odd/even subject numbers) and fitting the first half of the data using non-hierarchical versions of each model with uniform priors for the 6 expert function parameters (mean, standard deviation, height, location, scale, maximum). For each parameter, we used the mean and standard deviation of the posterior distributions as the mean and standard deviations of the Gaussian priors to fit the second half of the data (see Table 1 for the values used for the Gaussian priors).

We used the "rstan" package (Stan development team, 2018) to fit the models and computed WAICs (Widely Available or Watanabe-Akaike Information Criterion; Watanabe, 2010) using the "loo" package (Vehtari et al., 2017) to perform model comparison. WAIC is a Bayesian measure of predictive accuracy that accounts for model complexity, and is preferable to other criteria as it considers the whole posterior (Gelman et al., 2014). This was important as the choice model has more flexibility than the other two models. Since the majority of participants reported some degree of belief in both relational and similarity rules, the choice model can sample from three different experts (albeit probabilistically, and constrained by the empirical weights) and therefore potentially provide a closer fit to the data.

Table 1: Group-level parameters for the similarity (Gaussian: Mean, SD, Height) and relational (logistic: location, scale, maximum) expert functions. The prior values were obtained from fitting the first half of the data with uniform priors, and the posterior values were obtained from fitting the second half of the data using the calibrated priors.

Model	Prior						Posterior					
	M	SD	Ht	Loc	Scale	Max	M	SD	Ht	Loc	Scale	Max
Single expert	.19	.29	74.7	-.11	7.1	81.5	.19	.29	75.4	-.11	7.3	81.7
Averaging	.13	.35	58.7	-.08	38.8	88.2	.13	.34	63.6	-.08	46.8	89.6
Choice	.16	.29	67.1	-.10	34.2	81.7	.16	.29	69.1	-.10	47.8	83.5

Figure 5 shows the posterior predictives for each of the three models for a selected participant who responded "relational" in the forced-choice question and reported .66 belief for the similarity rule and .92 belief for the relational rule. It is clear that the choice model provides the best fit to the data, presumably for the reason stated above (greater flexibility). The averaging model does a moderately good job, but it is apparent that the single expert model is constrained to generalize exclusively according to a Gaussian or logistic function (logistic in this case), and thus provides the worst fit of the three candidate models.

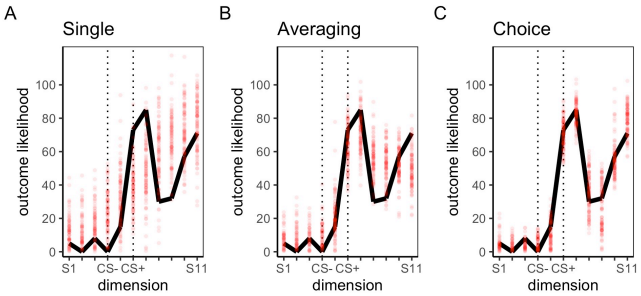

Figure 5: An empirical gradient (solid black line) and posterior predictives (overlaid red points) from each model. This participant was assigned weights S=.41, R=.57, G=.02.

Using the calibrated priors, the best model (lowest WAIC) was the choice model, followed by the averaging model, and then the single expert model (see Table 2). These results suggest that the choice model best fit the individual gradients, and that consideration of multiple rules better accounted for generalization performance than consideration of a single rule, even after accounting for differences in model complexity.

Table 2: WAIC and standard error for each model using calibrated priors.

Model	WAIC	SE
Single expert	4064.5	127.0
Averaging	3833.3	77.2
Choice	2859.4	227.8

General Discussion

Following differential training with an aqua rectangle as the CS+ and a slightly greener or bluer rectangle as the CS-, participants reported simultaneous belief in *both* similarity and relational rules. That is, they believed that the likelihood of the outcome increased according to the physical similarity to the CS+ (similarity rule), but also according to the relational difference between the CS+ and CS- (greener/bluer than). Although beliefs along these dimensions were negatively correlated, participants did seem to believe in two rules that were mutually contradictory for parts of the stimulus space.

To test whether participants used both of these rules when generalizing, we compared three different mixture of experts models in their ability to fit the individual gradients. The first was a single expert model, which assumed that each participant only used a single rule (and therefore only one expert was active for a given participant). A further two models considered the possibility that a given subject would be using multiple rules–one model assumed that performance was the weighted average between experts (averaging model) and the other assumed that performance was the result of a probabilistic choice between experts on each trial.

After accounting for model complexity, both models that considered multiple rules (averaging and choice) were more accurate in predicting individual generalization gradients than a model that assumed each participant only used a single rule. Importantly, we also found that a model where generalization was determined by a probabilistic choice between the experts was preferred over a model where generalization resulted from the weighted average between experts. This implies that although individuals may favor one rule over another overall, they switch flexibly between rules from trial to trial. To the best of our knowledge, this is the first attempt to model individual differences and multiple rule use in associative generalization.

A mixture of experts

Rule- and similarity-based generalization have generally been treated as theoretically distinct in associative learning (e.g., Shanks & Darby, 1998) and other cognitive domains (e.g., Pothos, 2005). In contrast, our mixture of experts model integrates relational and similarity-based generalization into a single theoretical framework. In this way, it is similar in philosophy to recent hybrid and rational models that assume joint contribution of rules and similarity in other types of learning (category learning: Schlegelmilch et al., 2020, function learning: DeLosh et al., 1997; Griffiths et al., 2008; Lucas et al., 2015).

The modelling results support the idea that individual participants use multiple rules in their generalization. At face value, it may be surprising that individuals entertained belief in mutually contradictory rules. Although counterintuitive, these results are consistent with *knowledge partitioning,* the notion that people are capable of representing pieces of contradictory knowledge in separate parcels (Kalish et al., 2004). When this occurs, studies show that participants alternate in their responding, similar to our participants switching between rules across test trials.

One area in which knowledge partitioning occurs is in function learning, which involves learning the function relating a continuous outcome (y) to a continuous input (x). The major difference between function learning and generalization in associative learning is that in function learning participants are making predictions about a continuous output (y), while in associative learning the outcome is binary (present or absent), but the dependent measures of learning are often continuous (e.g., predictive ratings, physiological response level). Interestingly, when participants observe inputs and outcomes consistent with multiple linear functions, their behavior appears to be

multimodal, suggesting that they sample from multiple possible functions (see also León-Villagrá et al., 2019). This idea is captured by the POLE model (Kalish et al., 2004), which posits multiple linear experts that accrue weights over training. The current model differs from POLE in that we only had 3 experts, the experts generalized in qualitatively different ways, and we weighted the experts empirically (i.e., based on participants' own responses). In this way, our model ties the hypothetical experts to participants' explicit beliefs.

In a similar task, Konolova and Le Mens (2018) administered a feature inference task where participants were asked to predict the level of a hormone (Protropin, y) based on the level of another hormone (Rexin, x). Critically, the levels of Rexin were predictive of category membership (the samples belonged to either rats or mice), and category membership was a reliable indicator of the levels of Protropin. Konolova and Le Mens found that when categorization based on Rexin levels was uncertain, participants showed bimodal responding, producing answers for each category in proportion to the likelihood of each category. Our results are thus consistent with the existing literature on knowledge partitioning in showing that when participants have multiple hypotheses about how a property should generalize, they switch between hypotheses–rather than averaging their implied responses–when making predictive judgements.

Implications for peak shift

Peak shift is known to be parameter-dependent in animals (Purtle, 1973) and is only inconsistently found in humans (e.g., Lee et al., 2018; Lovibond et al., 2020). Although this may be partly due to the statistical methods employed to measure peak shift (see Lee et al., 2021), the current study presents an additional reason for its elusiveness in humans. If participants sample from relational and similarity rules when generalizing, then this might result in very few *individuals* showing a peak shift. Whether a peak shift is detected at the aggregate level will be dependent on the exact shape of the similarity and relational functions, as well as the degree to which participants believe in each rule, which will vary from experiment to experiment. Sampling from multiple rules also provides one reason why generalization gradients in humans exhibit a high degree of variability from individual to individual (Lee et al., 2021) even in subgroups of participants who report the same generalization rule. Similar to our previous work (Lee et al., 2018), the current results highlight how analysis of aggregate gradients can be misleading when participants derive multiple, qualitatively distinct hypotheses about how to generalize.

Conclusion

In this study we have provided evidence that participants derive and use multiple rules when generalizing learned associations, and that participants are more likely to integrate the rules by sampling between them on a particular trial rather than averaging. Generalization in associative learning thus appears to exhibit similar characteristics to that in other cognitive domains. In particular, the results suggest that mixture of experts models and sampling effects are relevant to a wide variety of inductive phenomena. Formal modelling of individual differences and explicit rules and hypotheses may help to better understand associative generalization in humans.

Acknowledgments

This research was funded by two Discovery grants awarded by the Australian Research Council (DE210100292 awarded to JCL, DP190103738 awarded to PFL). SL was supported by a Humboldt Award from the Humboldt Foundation in Germany during part of this work.

References

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The Sine Qua Non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23(4)*, 968-986.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing, 24*, 997-1016.

Griffiths, T. L., Lucas, C. G., Williams, J. J., & Kalish, M. L. (2008). Modeling human function learning with Gaussian processes. *Advances in Neural Information Processing Systems, 21*.

Hanson, H. M. (1959). Effects of discrimination training on stimulus generalization. *Journal of Experimental Psychology, 58(5)*, 321-334.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of Linear Experts: Knowledge partitioning and function learning. *Psychological Review, 111(4)*, 1072-1099.

Konovalova, E. & Le Mens, G. (2018). Feature inference with uncertain categorization: Re-assessing Anderson's rational model. *Psychonomic Bulletin & Review, 25*, 1666-1681.

Lee, J. C., Hayes, B. K., & Lovibond, P. F. (2018). Peak shift and rules in human generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44(12)*, 1955-1970.

Lee, J. C., Lovibond, P. F., Hayes, B. K., & Navarro, D. J. (2019). Negative evidence and inductive reasoning in generalization of associative learning. *Journal of Experimental Psychology: General, 148(2),* 289-303.

Lee, J. C. & Livesey, E. J. (2018). Rule-based generalization and peak shift in the presence of simple relational rules. *PLoS ONE, 13(9):* e0203805.

Lee, J. C., Mills, L., Hayes, B. K., & Livesey, E. J. (2021). Modelling generalisation gradients as augmented Gaussian functions. *Quarterly Journal of Experimental Psychology, 74(1)*, 106-121.

León-Villagrá, P., Klar, V. S., Sanborn, A. N. & Lucas, C. G. (2019). Exploring the representation of linear functions. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, (pp.2112-2118). Cognitive Science Society.

Lissek, S. (2012). Toward an account of clinical anxiety predicated on basic, neutrally mapped mechanisms of Pavlovian fear learning: The case for conditioned overgeneralization. *Depression and Anxiety*, *29*, 257-263.

Little, J. L. & McDaniel, M. A. (2015). Individual differences in category learning: Memorization versus rule abstraction. *Memory & Cognition, 43*, 283-297.

Lovibond, P. F., Lee, J. C., & Hayes, B. K. (2020). Stimulus discriminability and induction as independent components of generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46(6)*, 1106-1120.

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review, 22*, 1193-1215.

Nosofsky, R. M. & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review, 5(3)*, 345-369.

Pothos, E. M. (2005). The rules versus similarity distinction. *Behavioral and Brain Sciences, 28*, 1-49.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Schlegelmilch, R., Wills, A., & von Helversen, B. (2020). A Cognitive Category-Learning Model of Rule Abstraction, Attention Learning, and Contextual Modulation. https://psyarxiv.com/4jukw/

Shanks, D. R. & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 24(4)*, 405-415.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237(4820)*, 1317–1323.

Stan Development Team. (2018). RStan: The R interface to Stan. Retrieved from http://mc-stan.org/

Tenenbaum, J. B. & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*, 629-640.

Thibaut, J., Gelaes, S., & Murphy, G. L. (2018). Does practice in category learning increase rule use or exemplar use – or both? *Memory & Cognition, 46*, 530-543.

Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2019). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.1.0, https://CRAN.R-project.org/package=loo

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research, 11(12),* p. 3571-3594.

Wills, S., & Mackintosh, N. (1998). Peak shift on an artificial dimension. *The Quarterly Journal of Experimental Psychology, 51B(1)*, 1-31.