Supplemental Material to: The role of perception in generalization: Commentary on Zaman, Yu, & Verheyan (2023)

Jessica C. Lee[1]

René Schlegelmilch[2]

1. University of Sydney

2. University of Bremen

Supplemental Material to: The role of perception in generalization:
Commentary on Zaman, Yu, & Verheyan (2023)

Re-analysis: Mixture of Response Strategies

One take-away message of our commentary is that it is difficult to investigate the role of perception in generalization. A within-subjects design is needed, and there is always the possibility that early tasks will affect later tasks, regardless of whether they consist of perceptual or learning judgements. However, as stated, we believe that it is worthwhile attempting to dissociate the contributions of perception and the cognitive processes of stimulus-outcome learning between generalization and identification, but that this requires more adequate statistical models. Therefore, we applied our recently proposed method (Schlegelmilch, Wills, & Lee, 2023) to meaningfully disentangle the prevalence and the shape characteristics in generalization and identification gradients for quantitative comparison, to test ZYV's hypotheses more directly. In this re-analysis, we took their data at face value (i.e., ignoring the interpretational issues discussed above). These quantitative measures are outlined below. Importantly, for this, we quantified the initial color ratings via individual regressions (boundary and between color discrimination), and further used the training scores as indicators of learning success, to be correlated with the assessed gradient characteristics.

Color Ratings & Training Performance

First, as predictors for gradient characteristics in the following analyses, we calculated the average rate of responding + for each CS+ and CS− in training (mean of 12 trials each, roughly reflecting learning speed), yielding two measures for each individual. As a measure of color perception, we applied to each individual a simple beta-regression to their blueness and greenness ratings, similar to the method by the original authors (mean-centered stimulus values as predictor; using the betareg R package Grün, Kosmidis, & Zeileis, 2011). The measure of interest is the slope of the color rating gradients used as indicator of color discrimination later. There was no evidence favoring a correlation between the difference in training scores on CS+ and

CS− and the individuals' color slopes (blue: $r = -.04$ $BF_{10} = .25$; green: $r = .1$ $BF_{10} = .4$; R-package BayesFactor, default settings; Morey & Rouder, 2018). Thus, failure of differential learning was not due to failure of color discrimination, and later correlations can be taken as reflecting different processes. The distribution of the CS' training scores can be seen in Figure 2. Both measures correlated with $r = -.35$, $BF_{10} = 459$, indicating that better learning of CS+ came with better learning of CS−, though, not perfectly related.

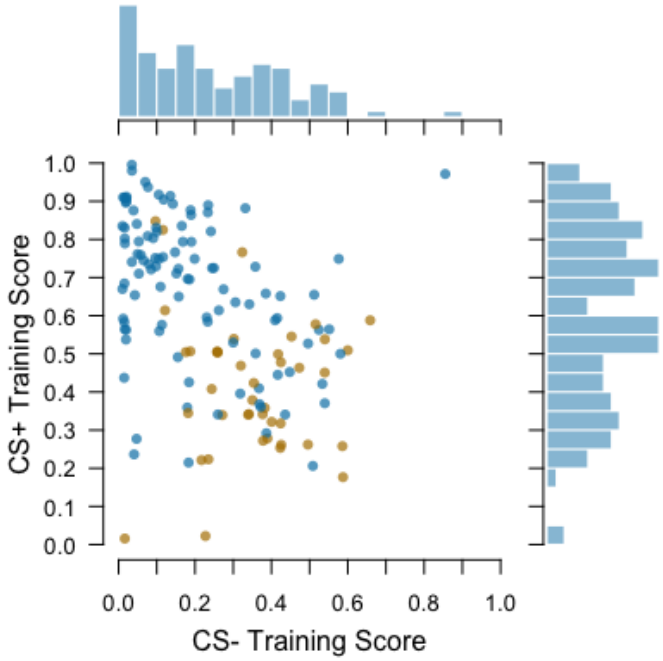


Figure 1. Average Individual Training Performances. Data points represent individual mean + responses (x-axis) to the CS+, orange to the CS−). Bars = corresponding histograms. Orange symbols = participants as originally excluded based on verbal reports.

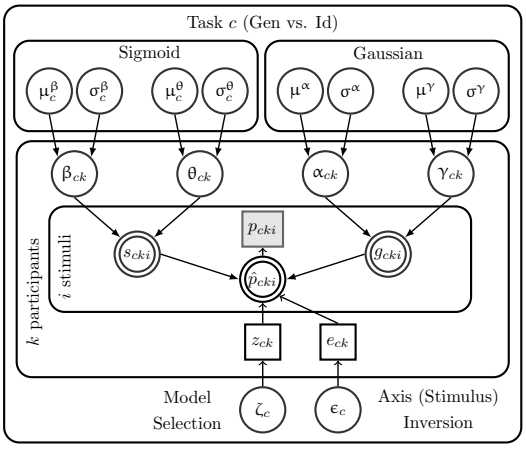
Modeling Generalization & Identification

To adequately quantify the shape characteristics of the generalization and identification gradients we applied the same measurement approach to both tasks. For this we used a method proposed recently (Schlegelmilch et al., 2023), which modifies the Augmented Gaussian model introduced by J. C. Lee, Mills, Hayes, and Livesey (2021),

in a hierarchical Bayesian mixture model. That is, we assume that the individual differences reflect a mix of standard sigmoid functions (rule-like) and Gaussian functions around the CS+ (similarity-like) to account for the variety of response patterns.

First, for each individual k, both Sigmoid and Gaussian predictions are calculated over the stimuli i (i.e., s_{cki} and g_{cki}, respectively). These predictions depend on the priors illustrated in Figure 1. The model selection is based on a transdimensional Markov-chain Monte-Carlo method (M. D. Lee & Wagenmakers, 2014; Sisson, 2005) which evaluates in each of the iterations whether it is more likely that the pattern is purely sigmoidal ($z_{ck} = 1$) or whether a Gaussian ($z_{ck} = 2$) should be augmented left-hand starting with the estimated peak location γ_{ck}. The term 'left-hand' thus refers to the notation in Figure 1, stating 'if $z_{ck} = 2$ for $x_i < \gamma_{ck}$'. Note, the stimulus values x_i are transformed from 1 to 14 to standardized values ($M = 0$, $SD = 1$) for scaling the parameters and 0 reflects the middle of the stimulus space. In other words, γ_{ck} estimates which stimulus in the continuum is the point where the Sigmoid becomes a Gaussian. To avoid, that the estimated peak could be placed at the lower endpoint of the stimulus continuum, rendering the Gaussian slope unidentifiable, we truncated the γ_{ck} estimates to be larger or equal in location to stimulus 3.

The model indicator z_{ck} is sampled from ζ_c, which reflects the overall probability of model assignment in the given task. Thus, the relative number of assignments z_{ck} in all model iterations indicates within-participant model probability, and ζ_c can be used to test for changes in the average assignments in the experimental design in a logistic regression. Here, the model in Figure 1 displays the corresponding logistic function, implementing a model-matrix (MM), reflecting the main effect of task (Gen vs. Id) using contrast coding (taking 1 and -1 values). We then multiplied these values before application, with the effect estimate of task. The intercept of the slope had the prior $\beta_{\text{int}} \sim \text{Gaussian}(5, 1)$ (i.e., assuming a generally higher prevalence of the Sigmoid as found in Schlegelmilch et al., 2023), and the effect of condition with $\beta_{\text{effect}} \sim \text{Gaussian}(0, 1)$. We estimated the threshold intercepts of both Sigmoid and Gaussian hierarchically without model matrix, despite to-be-expected task differences,



Priors

Sigmoid

$\mu_c^\beta \sim \text{Gaussian}_{[.01,15]}(3,1)$

$\mu_c^\theta \sim \text{Gaussian}_{[-1.5,1.5]}(0,1)$

$\sigma_c^\beta \sim \text{Gaussian}_{[0,2.5]}(1,.1)$

$\sigma_c^\theta \sim \text{Gaussian}_{[0,1]}(1,.1)$

Gaussian

$\mu^\alpha \sim \text{Gaussian}_{[.01,15]}(3,1)$

$\mu^\gamma \sim \text{Gaussian}_{[-1.3,.5]}(-0.61,1)$

$\sigma^\alpha \sim \text{Gaussian}_{[0,.5]}(1,.1)$

$\sigma^\gamma \leftarrow .25$

Individual Predictions

$\beta_{ck} \leftarrow \mu_c^\beta + \text{Gaussian}_{[-4,4]}(0, 1/(\sigma_c^\beta)^2)$

$\theta_{ck} \leftarrow \mu_c^\theta + \text{Gaussian}_{[-.5,.5]}(0, 1/(\sigma_c^\theta)^2)$

$s_{cki} \leftarrow (e^{\beta_{ck}\cdot(\theta_{ck}-x_i)})^{-1}$

$\alpha_{ck} \sim \text{Gaussian}(\mu_c^\alpha, 1/(\sigma_c^\alpha)^2)$

$\gamma_{ck} \sim \text{Gaussian}(\mu_c^\gamma, 1/(\sigma_c^\gamma)^2)$

$g_{cki} \leftarrow e^{-(\gamma_{ck}-x_i)^2/2\cdot(1/\alpha_{ck})^2}$

Model Selection

$\epsilon_c \sim \text{Beta}(1,1)$

$\zeta_c \leftarrow (1+e^{\text{MM}})^{-1}$

$o \sim \text{Uniform}(0,10)$

$e_{ck} \sim \text{Categorical}(\epsilon_c)$

$z_{ck} \sim \text{Categorical}(\zeta_c \cdot o)$

$$\hat{p}'_{cki} \leftarrow \begin{cases} s_{cki}, \text{ if } z_{ck} = 1 \\ g_{cki} \cdot \min(s_{ck(i<\gamma_{ck})}), \text{ if } z_{ck} = 2 \text{ for } x_i < \gamma_{ck} \end{cases}$$

$\hat{p}_{cki} \leftarrow \hat{p}'_{cki}$, if $e_{ck} = 1$, else invert i

$p_{cki} \sim \text{Binomial}(\hat{p}_{cki}, 10)$

Figure 2. Graphical Bayesian Model for the Mixture of Gaussian vs. Sigmoid Patterns. The terms MM reflects a Model-Matrix for hypothesis testing (main effect of condition), explained in the text.

as the purpose was to assess the correlation of individual estimates between tasks, which we did in post-hoc analyses, for simplicity.

In order to seamlessly transition between both function's predictions, the Gaussian prediction g_{cki} is transformed via $g_{cki} \cdot \min(s_{ck(i<\gamma_{ck})})$. That is, the standard Gaussian has by definition a peak-height of 1 but the Sigmoid at the estimated transition location is not necessarily 1. The term min() scales the Gaussian curve at its peak down to the Sigmoid height. Specifically, min() reflects that the Sigmoid by definition can only increase with more extreme values ($i < \gamma_{ck}$), and the minimum is the predicted response strength at the transition location, or in other words, the estimated height of the Gaussian peak, if assigned.

This mixture model, allows to assess the sigmoid slope (β) with which the gradient falls off in strength between CS+ and CS−, the sigmoid threshold (θ) measuring whether the point of indifference between CS+ and CS− (50% responding) is generally shifted towards the CS+ or CS−, and importantly, whether the gradient function is rule-like (sigmoidal) or similarity-like (bell-shaped) around the CS+ (z_{ck}). In

the latter, we measure the slope (α) as the decline left-hand to the CS+, if assigned, and the stimulus location (γ) at which the sigmoid turns to a downward trend. This means, β and θ apply to all participants, but α and γ only if a Gaussian was assigned (otherwise sampled from the current posterior). However, as a third model we entered a pure guessing model if both Sigmoid and Gaussian were unlikely, but which was hardly assigned. Finally, in very few cases participants apparently flipped the stimulus scale (or assigned the wrong labels). These cases are estimated via ϵ_{ck} in model selection, which then simply leads to deciding whether the otherwise selected model would be inverted in its predictions along the stimulus axis, that is, to avoid adding separate models for these cases (i.e., same priors as the Sigmoid and Gaussian models). A graphical model is illustrated in Figure 1, together with the parameter priors. The model was implemented using the R-package jags (Plummer et al., 2003; R Core Team, 2018), and converged in 30000 iterations on 3 chains with 5000 iterations adaptation period. The basic results are depicted in Figure 2.

The overall number of participants assigned with confidence to one of the models ($BF_{10} > 3$) was 99 in each task. Within participants who verbally reported similarity-like responding in the questionnaire, the assignments in the generalizations task were 35% Sigmoid and 63% Gaussian, and within the rule-like group, 59% and 28% (contignencyTable: BF_{10}>10), indicating some validity in assessment. As can be seen in Figure 3, the Gaussian model was more often assigned in identification than in generalization with $N = 86$ and $N = 72$, respectively, however, the evidence regarding a difference in assignments on ζ_c between generalization and identification was inconclusive, BF=1.31 (based on comparing the task effect prior to the posterior density ratio in the interval of -.01 and .01). In addition, there was evidence against the hypothesis that color slopes (blue-green dichotomization) differed between the assigned models (all BF's< 1/3 in t-tests, inconclusive for greenness in identification, $BF_{10} = .40$).

Crucially, Figure 3 shows that the generalization gradients were considerably steeper/narrower in the identification compared to the generalization task, clearly

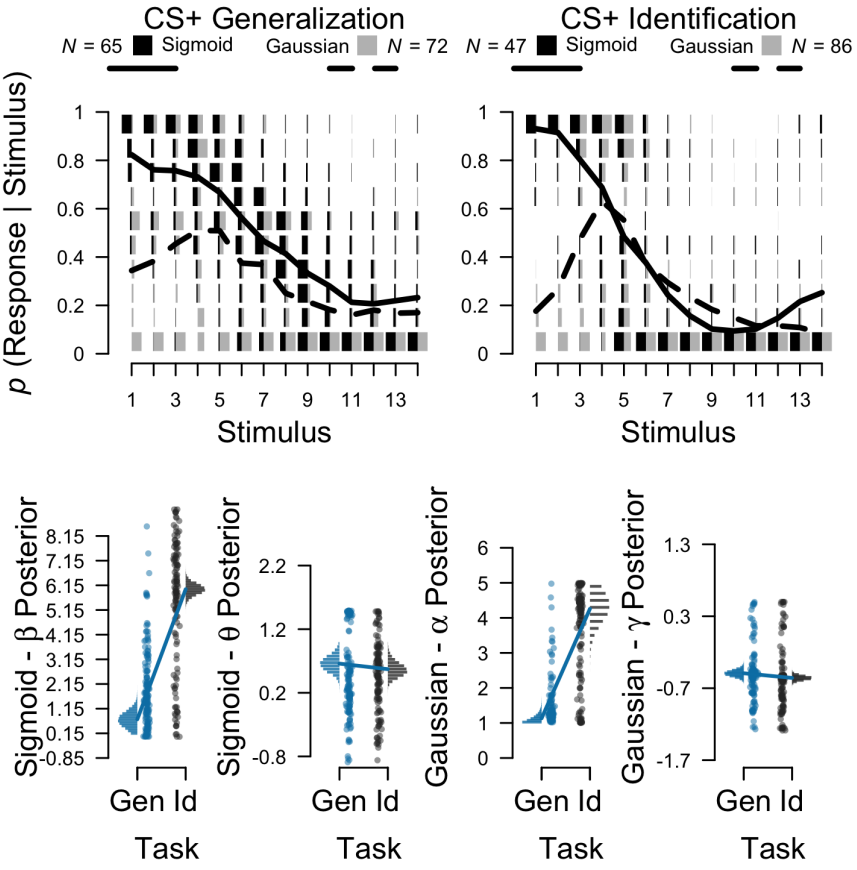


Figure 3. Bayesian Model Selection Results. Top: Model assignments in Generalization (left; y-axis + response in %) and Identification (right; y-axis % 'same as CS+' responses). Bottom: Corresponding parameter estimates in each task. Lines = population means, histograms = hyper posteriors. Small circles = individual maximum likelihood posterior estimates. See text.

showing systematic differences. That is, the comparison on the individual parameter estimates of the steepness of both gradient functions, β (right-hand slope) and α (left-hand slope), reveal the typically observed narrowing in identification relative to generalization (BF_{10}=infinite). Thus, there is a clear change in cognitive processing, that must not be neglected when drawing conclusions regarding the contribution of individual differences in identification to overall variance in generalization.

To test whether the individual gradient characteristics nonetheless correlated between tasks, we derived each individual's slope and threshold parameters (maximum likelihood estimates). The only assessed variable for which there was evidence for a

Table 1

Gradient Parameter Correlations with Training Performance and Color Ratings

	Sigmoid		Gaussian	
	β	θ	α	γ
Generalization				
CS+ Score	0.38** (-.45**)	-0.35** (-0.37**)	–	-.0.26** (-0.40**)
CS− Score	-0.55** (-.55**)	-0.21 (–)	-0.11 (–)	–
Blue Slope	–	–	–	–
Green Slope	0.14(–)	–	–	–
Identification				
CS+ Score	0.16 (0.23*)	0.15 (–)	0.10 (0.21*)	-0.15 (-0.19*)
CS− Score	-0.23* (-0.19*)	-0.29** (–)	-0.18 (–)	0.22** (–)
Blue Slope	-0.16 (–)	-0.18 (–)	-0.10 (–)	0.13 (0.24*)
Green Slope	0.11 (–)	0.30** (0.19*)	0.23* (0.24*)	-0.22* (-0.26**)

Note. * indicates = $BF_{10} > 3$, ** = $BF_{10} > 10$, no sign = inconclusive, – = $BF_{10} < 1/3$. Numbers (or signs) in brackets reflect estimates (or evidence) without applying the non-preregistered participant exclusions, if unequal to the data-set with exclusions. CS+ Score: training + responses on CS+ (higher = more accurate); CS− Score: training + responses on CS− (lower = more accurate); Blue and Green Slopes: Rate of change in corresponding color ratings along stimulus space (higher = stronger dichotomization); β: higher values = stronger between CS discrimination; θ: boundary (50% responding) shift towards blue < 0 < shift towards green; α: higher values = stronger decline in + responses left to CS+; γ: peak of Gaussian = towards blue < 0 < towards green.

between-task correlation was β, with $r = .35$, $BF_{10} > 100$, indeed, indicating some inter-individual stability regarding the decline of response strength between CS+ and CS−. However, as highlighted in the commentary, this can not be well interpreted due to carry-over effects, but further shows no evidence favoring a correlation with the color discrimination at the boundary (see below). There were three weaker correlations, however, which were between-variable and task, and thus difficult to interpret and not reported.

More intriguingly, the correlations between the model measures, training scores and color ratings (see Table 1) leave no doubt that the CS+ and/or the CS− training performances explained variance in both tasks, while color ratings only predicted variance in the identification task. First, during generalization, the higher the accuracy

in prior training on CS− the more precise the generalization of (or discrimination from) CS+, with $r = -.55$. This was also the case in identification but to a lesser degree, which seems unsurprising due to continued reinforcement during generalization which should further alter learned stimulus-outcome associations. The same but weaker trends were present regarding prior CS+ training accuracy. This means, that + outcome generalization involves a relative comparison on how much both outcomes were associated to the stimuli, beyond perception or CS+ memory.

Importantly, the pattern changed in the identification task in a dissociable manner. Here, indeed, the greenness-color slopes correlated with how strongly the gradients were shifted and with the left-hand slopes around the CS+ if a Gaussian was assigned, thus, indicating some influence on identification. However, the former seems difficult to interpret, while the latter seems unsurprising, indicating that the same-different identification judgments became more precise (left-hand) with more differentiating color ratings (e.g., either due to screen settings in the online experiment, or perception). Most importantly, it seems that the color ratings, indeed, *could* be predictive, but there was overall evidence *against* a similar role in generalization (except one weak inconclusive correlation on the green slope and β), contradicting the authors' original conclusions regarding the influence of individual differences in perception on generalization.

Summary

In summary, we believe that the issues raised in the commentary as well as the insights using adequate descriptive modeling methods (Schlegelmilch et al., 2023) undermine the validity of ZYV conclusions. In our view, the ZYV's arguments and methods are not convincing enough to recommend discarding existing methods and established theories. In contrast, our present reanalysis reveals a role of training performance of both CS+ and CS− on the subsequent generalization gradients. This not only turns the more pressing question to *why* some participants fail to uncover the rather simple stimulus-outcome discrimination, but further implies that participants do

consider absence of outcomes (i.e., CS− associations) for making inferences, which is not an uncommon assumption, but which ZYV hardly took into account. However, while color ratings did neither contribute to explaining variance in the discrimination-training performance nor the generalization gradients' characteristics, they seemed nonetheless related to the gradients in the same-different identification task. This result seems reassuring and highlights that processes involved in associative learning and perception could be disentangled, which might help arriving at more reliable insights in future studies.

References

Grün, B., Kosmidis, I., & Zeileis, A. (2011). *Extended beta regression in r: Shaken, stirred, mixed, and partitioned* (Tech. Rep.). Working Papers in Economics and Statistics.

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.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Morey, R. D., & Rouder, J. N. (2018). Bayesfactor: Computation of bayes factors for common designs [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.12-4.2)

Plummer, M., et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, pp. 1–10).

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Schlegelmilch, R., Wills, A. J., & Lee, J. C. (2023, Sep). *Is generalization general? a comparison of predictive and category learning and the influence of response format and framing.* PsyArXiv. Retrieved from `psyarxiv.com/eakwz` doi: 10.31234/osf.io/eakwz

Sisson, S. A. (2005). Transdimensional markov chains: A decade of progress and future perspectives. *Journal of the American Statistical Association*, *100*(471), 1077–1089. doi: 10.1198/016214505000000664