Journal of Experimental Psychology: General

ISSN: 0096-3445

2023, Vol. 152, No. 8, 2345–2358
https://doi.org/10.1037/xge0001403

The Idiosyncratic Nature of How Individuals Perceive, Represent, and Remember Their Surroundings and Its Impact on Learning-Based Generalization

Jonas Zaman[1, 2, 3], Kenny Yu[4], and Steven Verheyen[3]

[1] Centre for the Psychology of Learning and Experimental Psychopathology, Faculty of Psychology and Educational Sciences, KU Leuven
[2] Health Psychology, Faculty of Psychology and Educational Sciences, KU Leuven
[3] Department of Psychology, Education and Child Studies, Erasmus University Rotterdam
[4] Quantitative Psychology and Individual Differences, Faculty of Psychology and Educational Sciences, KU Leuven

The current study adopted a multimodal assessment approach to map the idiosyncratic nature of how individuals perceive, represent, and remember their surroundings and to investigate its impact on learning-based generalization. During an online differential conditioning paradigm, participants ($n = 105$) learned the pairing between a blue color patch (CS+) and an outcome (i.e., shock symbol) and the unpairing between a green color patch and the same outcome. After the learning task, the generalization of outcome expectancies was assessed to 14 stimuli spanning the entire blue-green color spectrum. Hereafter, a stimulus identification task assessed the ability to correctly identify the CS+ among this stimulus range. Continuous and binary color category membership judgments of the stimuli were assessed preconditioning. We found that a response model with color perception and identification performance as sole predictors was preferred to contemporary approaches that use stimulus as a predictor. Interestingly, incorporating interindividual differences in color perception, CS identification, and color categories significantly improved the models' ability to account for different generalization patterns. Our findings suggest that insight into the idiosyncratic nature of how individuals perceive, represent, and remember their surroundings provides exciting opportunities to understand post-learning behaviors better.

Public Significance Statement

This study strongly suggests that the pattern of learning-based generalization can best be understood through a combination of stimulus perception and memory for the stimulus during learning. The study highlights the importance of inter-individual differences in perception, memory, and the level of representation, as they accounted for differences in generalized responses between individuals.

Keywords: differential conditioning, generalization, color perception, categorization, identification,

Supplemental materials: https://doi.org/10.1037/xge0001403.supp

Behaviors are not solely based on direct experience but often involve an extrapolating component, using past learning to guide current behaviors. Generalization is an efficient mechanism to maximize learning gains, foster behavioral consistency, and avoid (costly) re-learning. However, whereas insufficient generalization leads to suboptimal use of past knowledge, rigid generalization may result in inflexible and potentially problematic behaviors (Dymond et al., 2015; Vlaeyen & Linton, 2012). To appropriately

This article was published Online First April 13, 2023.

Jonas Zaman https://orcid.org/0000-0002-2218-3018

Jonas Zaman is now at School of Social Sciences, University of Hasselt.

Jonas Zaman is a Postdoctoral Research Fellow of the Research Foundation Flanders (FWO, 12P8619N) and received funding from the Efic-Grünenthal grant (EGG ID 358254826), a Special Research Funds (FWO, 1500620N), and a travel grant (FWO, V433120N). Kenny Yu is supported by an FWO research project (co-PI: JZ, G079520N).

The evaluation, opportunities for promotion, and ability to obtain research funding of all authors are partly dependent on the number of articles they publish.

The authors have no conflict of interest to report.

We preregistered the study on the Open Science Framework (https://osf.io/chu3w) and explicitly mention any deviations. All preprocessed and processed data, processing and analyses code, and the experiment can be found at https://osf.io/d5s2r/. The study and its results have been presented at the 26th Associative Learning Symposium in April 2022.

Jonas Zaman served as lead for formal analysis, funding acquisition, project administration, visualization, and writing–original draft. Kenny Yu served in a supporting role for formal analysis and writing–review and editing. Jonas Zaman and Steven Verheyen contributed equally to supervision. Jonas Zaman and Steven Verheyen contributed to conceptualization, writing–review and editing, and methodology equally.

Correspondence concerning this article should be addressed to Jonas Zaman, Centre for the Psychology of Learning and Experimental Psychopathology, Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102, Bus 3726, 3000 Leuven, Belgium. Email: jonas.zaman@kuleuven.be

determine the applicability of past experiences to the current situation, this process relies on both conceptual and perceptual similarities between situations (Dymond et al., 2015).

Stimulus generalization is a well-established finding both in animals and humans, where a range of novel stimuli come to elicit a previously stimulus-trained response, despite these stimuli never being reinforced (Enquist & Ghirlanda, 2005; Ghirlanda & Enquist, 2003; Honig & Urcuioli, 1981; Mednick & Freedman, 1960). The observation that the strength of the trained response wanes with physical differences inspired theories with a similarity-based process at their core, although theoretical conceptions differed (e.g., Atkinson & Estes, 1963; Blough, 1975; Enquist & Ghirlanda, 2005; Ghirlanda & Enquist, 2007; McLaren & Mackintosh, 2000, 2002; Schlegelmilch et al., 2022; Shepard, 1958, 1987). However, regardless of theoretical differences, parametrization of this latent similarity-based generalization process involves relating responses to a physical stimulus dimension. This custom equates mental representations to a static and veridical reflection of physical reality and stimulus similarity to a physical constant for any given stimulus pair and individual (Zaman, Chalkia, et al., 2021). Consequently, differences in response gradients (along the physical dimension) between individuals or groups are, by default, attributed to the latent generalization process without considering alternative explanations.

Recent research started challenging this custom for various reasons (Struyf et al., 2015; Zaman, Chalkia, et al., 2021). In the context of generalization, determining similarity entails comparing the memory of the training stimulus (CS+) with the perception of the current test stimulus. As stimulus perception and memory are variable and often idiosyncratic processes (Hoskin et al., 2019; Zenses et al., 2021), generalization models should account for intra- and interindividual differences in either aspect, as they can substantially influence generalized responding (see Struyf et al., 2015; Zaman, Chalkia, et al., 2021). An increasing amount of recent studies have indeed demonstrated the profound impact of these aspects on generalized responding (Holt et al., 2014; Norbury et al., 2018; Struyf et al., 2017; Zaman, Ceulemans, et al., 2019; Zaman & Lee, 2020; Zaman, Struyf, et al., 2019, 2020; Zenses et al., 2021). It follows that inferences regarding the extent of a latent generalization process, when made without the theoretical and (data) analytical incorporation of differences in perception and memory, are at risk of being incorrect. That is, variations in response gradients do not necessarily reflect differences in the generalization process but may also result from differences in stimulus perception or CS memory (e.g., Zaman, Ceulemans, et al., 2019; Zaman, Struyf, et al., 2019; Zenses et al., 2021).

The bulk of generalization research analyzes response gradients (Enquist & Ghirlanda, 2005; Mednick & Freedman, 1960; Vanbrabant et al., 2015) with stimulus as a categorical predictor (i.e., repeated measures analysis of variance (rmANOVAs) or Generalized Linear Models). This considers stimuli as distinct categories, enabling model flexibility but at the cost of many parameters (# parameters = # stimuli − 1). Other approaches include a higher-order stimulus polynomial (Vanbrabant et al., 2015) to account for the nonlinearity of typical response gradients (bell-shaped or S-shaped, depending on the conditioning paradigm). The disadvantage of both approaches is the complexity of interpreting parameter estimates, derivation of a generalization index, and linking them to theoretical constructs. More recently (and more proximal to theoretical conceptualizations), generative modeling approaches used augmented Gaussian distributions with the standard deviation parametrizing the extent of generalization (J. C. Lee et al., 2021) or a combination of (more abstract) response rules and a similarity-based generalization process (Schlegelmilch et al., 2022) to model response gradients. These approaches, however, still assume physical stimulus features as explanatory variables, ignoring (potential) differences in perception and memory and their impact on generalized responses. Therefore, this study's first aim is to investigate to what extent differences in generalization gradients along a blue-green color spectrum can be attributed to differences in color perception and stimulus identification.

Multiple manners exist to represent a blue-green color dimension: as a continuous hue dimension or as color categories (blue vs. green) with different psychometric curves for both aspects (Decock & Douven, 2014; Douven et al., 2017; Jraissati & Douven, 2017). Lovibond et al. (2020) studied outcome expectancies along a green-blue spectrum after learning to associate a specific color (e.g., blue-ish, conditioned stimulus, CS+) to an outcome (e.g., shock symbol, unconditioned stimulus, US), and in case of differential learning, at the same time the dissociation between another color (i.e., green-ish, CS−) and the US. They found that, in certain subgroups, outcome expectancies plateaued from a specific hue (despite the absence of a ceiling effect). This may suggest using a response strategy based on the conceptual organization of stimuli (e.g., blue = expect a shock, green = no shock) rather than a similarity-based generalization process. Nevertheless, whether and how interindividual differences in the level of representation (continuous features vs. semantic categories) determine generalization response patterns remains an unexplored question. The finding that acquired fears spread to conceptually related stimuli, such as items that belong to the same category (e.g., tools or animals; Dunsmoor & Murphy, 2015; Wong & Beckers, 2021), indicates that generalization is not bound to physical/perceptual similarity. Furthermore, the structure and positioning of items within a category are well-studied in the semantic memory literature, with ample evidence showing that the organization of categories is not universal (Douven, 2016; Verheyen & Égré, 2018; Verheyen & Storms, 2013). There are large interindividual differences regarding the perceived category membership of items (Hampton, 1995; McCloskey & Glucksberg, 1978; Verheyen et al., 2010; Verheyen, White, & Égré, 2019), the degree to which an item is considered typical for a category (Barsalou, 1987, 1989; Hampton & Passanisi, 2016), and the dimensions used to decide upon category membership (Verheyen & Storms, 2013; White et al., 2018). Therefore, this study aims to assess (interindividual) color categories and to investigate their additive effect on generalization gradients along a blue-green color dimension.

Recent human research reported the use of various response rules during a generalization phase (J. C. Lee et al., 2018; Lovibond et al., 2020; Wong & Lovibond, 2017). For instance, while some participants generalized along a green-blue stimulus dimension based on the similarity (i.e., *similarity* rule) of the test stimulus to the reinforced stimulus (CS+), others reported adhering to a more abstract rule (i.e., "the bluer the stimulus, the more I expect a shock," called the *linear* rule). While this work has demonstrated the profound impact of response rules, it failed to identify what caused participants to develop different types of response rules despite similar learning experiences. A recent study found latent group differences in the ability to correctly identify the CS+ (Zaman, Yu, & Lee, 2022). Through cluster analysis, participants allocated to a group with good CS identification performance had a much higher probability of reporting a similarity-based response rule. However, as CS identification performance depends on perceptual and memory

processes, it remains unclear whether differences in attention during learning, perceptual sensitivity, or memory accuracy determine which response rule one adopts. Therefore, the studies' final aim was twofold: to replicate the effects of response rules on generalization and to investigate differences in learning, color perception, categorization, and CS identification between the rule subgroups.

In an online differential conditioning paradigm, participants learned the pairing between a blue color patch (CS+) and an outcome (i.e., shock symbol) and the unpairing between a green color patch (CS−) and the same outcome. Color perception and color categories were assessed before this predictive learning task. After the learning task, the generalization of outcome expectancies was assessed to 14 stimuli spanning the entire blue-green color spectrum (S1–S14). Hereafter, verbalized response rules were assessed. The experiment ended with an identification task in which participants had to identify the CS+ among the test range (Figure 1).

Method and Materials

Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. We preregistered the study on the Open Science Framework (https://osf.io/chu3w) and explicitly mentioned any deviations. All preprocessed and processed data, processing and analyses code, and the experiment can be found at https://osf.io/d5s2r/ (Zaman, Yu, & Verheyen, 2022).

Participants

Data was collected online via Prolific, with the experiment run on Pavlovia (https://pavlovia.org/). The study lasted ±25 min, and participants received 5.28£. We conducted a power analysis in simR (Green & Macleod, 2016) based on half of the observed effect sizes from two pilot studies ($n_1 = 44$, $n_2 = 32$, see the online supplemental materials). At $n = 100$, the lower bound of the 95% CI of the simulated power was above a power of .95. This number was incremented to $n = 150$ to account for exclusions and the unbalanced rule distribution between the linear and similar blue rules (based on the pilot data). Inclusion criteria on Prolific were age >18, native English speaking, normal or corrected-to-normal vision, and a prolific approval rate >90%.[1] We excluded participants that indicated at the end of the experiment that they did not respond seriously ($n = 3$) or that selected an incongruent learning rule (one of the green rules, $n = 14$, see below). Deviating from the preregistration, we additionally excluded those who selected the "no rule" option ($n = 26$) as there was no evidence of differential learning in this group (see the online supplemental materials). Two participants did not complete the experiment. This led to a final N of 105, with 32 (30.1%) participants who selected the linear (called linBlue) response rule and 73 (69.9%) who selected the similarity (called simBlue) response rule. The study received approval from the Ethics review Committee of the Erasmus University Rotterdam (Approval No: 20-063).

Stimuli

To generate test stimuli, we used CIELUV (Malacara, 2002). This is a three-dimensional space designed by color scientists with the aim to represent colors that are perceived as equally different by equally distanced points in that space (as opposed to RGB space). Blue and green color prototypes and test stimuli were based on Douven (2019). The respective CIELUV coordinates of the blue and green color prototypes are (26.45, −10.21, −104.25) and (63.18, −59.74, 61.39). In addition to these two colors (S1 and S14), 12 colors gradually varying from blue to green were created by positioning 12 equally-spaced points on the line connecting both prototypes in this three-dimensional space. Although an important improvement in color representation, CIELUV only approximates uniformity. Moreover, because of the online format of the current study, participants used their own display monitors. Consequently, monitor calibration was not possible, which most likely affected the correctness of the approximations of the true CIELUV coordinates. Because of these limitations, the stimuli may only be approximately equidistant. The color patches had a size of 512 × 512 pixels, similar to Douven et al. (2017), and were created in MS Paint. S4 and S12 served as CS+ and CS−, respectively.

Protocol

Color Evaluation Phase

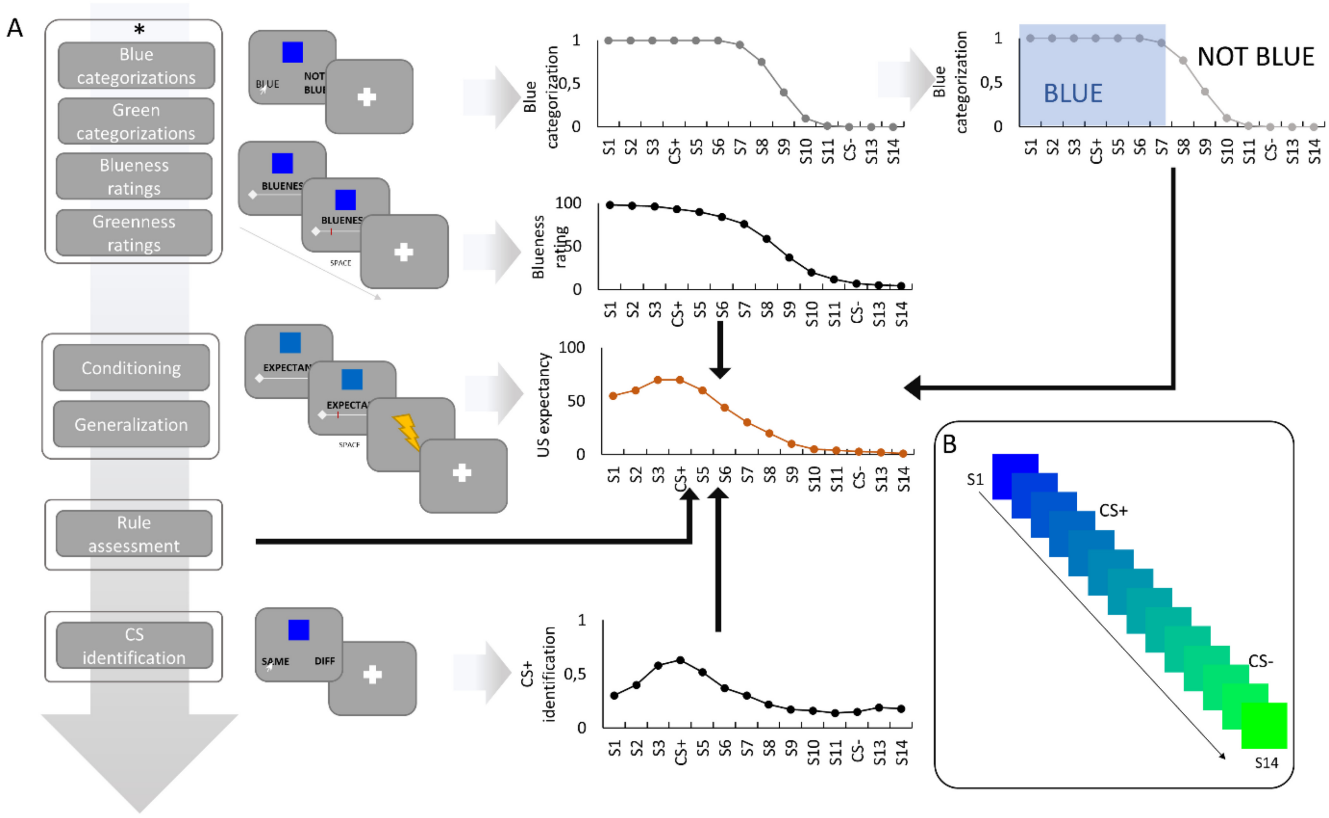
The experimental procedure commenced with four different color evaluation tasks (color [blue or green] × type of task [categorization or rating]). During these tasks, all color patches (S1:S14) were repeatedly presented and rated on the degree of blueness (or greenness) using a visual analog scale (VAS) (e.g., blueVAS: 0 = not blue, 100 = blue) or categorized on color membership (e.g., not blue/blue). All color patches were presented twice for the rating task and eight times for the categorization task. This total of 28 or 112 trials was equally distributed across two or eight blocks, with a random presentation order within blocks. A trial commenced with the presentation of a color patch with underneath a color VAS or two response options (e.g., not blue/blue). Participants had to confirm their VAS response using the space button, after which the trial ended. For the categorization tasks, trials immediately ended upon clicking one of the response options. The presentation duration of the color patch was set to a minimum of 2 s and limited to a maximum of 6 s. A 1 s fixation-cross separated subsequent trials. The order of the color evaluation tasks was random.

Conditioning Phase

A standard differential conditioning paradigm was adopted, consisting of 12 CS+ (S4) presentations and 12 CS− (S12) presentations. The CS+ trials were followed by the US (a symbol of an electrical shock) in 75% of the trials (nine trials). CS− trials were never followed by the US. Trials were presented in random order. The trial structure was identical to that of the color rating tasks apart from the US expectancy VAS (1 = no shock, 10 = shock) at the bottom of the screen, through which participants indicated whether they expected the electrical shock symbol to follow. On every trial, a color patch with the US expectancy VAS underneath was presented for a minimum of 2 s and a maximum of 6 s depending on the participant's response, as they had to confirm their VAS rating using the space button, after which the trial ended. A 1 s fixation-cross separated subsequent trials.

[1] In the preregistration, 80% instead of 90% was mentioned.

Figure 1
Study Overview



Note. (A) Overview of the protocol and illustration of some predictors used to model US expectancy gradients. (B) Overview of the test dimension. See the online article for the color version of this figure.

Generalization Phase

This task immediately followed the conditioning phase, and the trial structure was identical to that during the conditioning phase. The only difference was the range of presented stimuli. The CS+ and CS− were each presented 8 times ($8 \times 2 = 16$) additionally to 2 presentations per test stimulus ($2 \times 12 = 24$). The same reinforcement rate for the CS+ was adopted during the conditioning phase. The 40 trials were equally distributed across two blocks, with a random presentation order within blocks.

Rule Assessment

After the generalization task, similar to Lovibond et al. (2020), participants were instructed to select the most appropriate response rule from a list of five different options: (a) The bluer the stimulus was, the more likely for the shock symbol to appear (called linBlue rule); (b) The greener the stimulus was, the more likely for the shock symbol to appear (called linGreen rule); (c) The more the stimulus resembled a specific color blue, the more likely for the shock symbol to appear (called simBlue rule), (d) The more the stimulus resembled a specific color green, the more likely for the shock symbol to appear (called simGreen rule); We added a fifth response option: (e) None of these rules applied. Hereafter, confidence ratings for Options 1–4 were obtained in a random order using a VAS (0 = no confidence, 100 = high confidence).

Identification Phase

Before this phase, participants were instructed that this task aimed to identify whether the presented color patch was identical to the color patch predictive of the US (i.e., CS+). Eight repetitions per color patch were presented ($8 \times 14 = 112$), equally distributed across eight blocks, with a random presentation order within blocks. The words SAME and DIFFERENT (as the CS+) were depicted underneath the presented color patches. Participants responded by mouse clicking, after which the trial immediately ended, without a limit on presentation duration.

Analyses

All analyses were performed in R version 4.0.1. (R Core Team, 2013). As a manipulation check, we tested for differential learning during the conditioning phase with the following Generalized Linear Model (GLM) using the lme4 package v1.1.23 (Bates et al., 2015): US expectancy (trial level) as a dependent variable and Trial[2] (continuous predictor: 1:12), CS (categorical predictor: CS+ vs. CS−), and Trial × CS interaction as fixed effects, and a subject-dependent intercept as a random effect. Omnibus tests of main

[2] Relative trial number per CS type was used.

effects were obtained using analysis of variance from the car package v3.0.8 (Fox & Weisberg, 2019). We conducted posthoc contrasts using the models' estimated means with the emmeans package v.1.5.0 (Lenth, 2022).

For the generalization phase, US expectancy gradients (averaged over trials) across the test dimension were analyzed with different GLMs varying in the way they operationalize the stimulus dimension. An overview of the different operationalizations can be found in Table 1. In a first step, we investigated which continuous dimension (or combination) best fitted the data. In a second step, we investigated whether conceptual structures embedded within the test range further contributed to the pattern of responding via the inclusion of binary predictors (e.g., color category membership, color prototype). In the last step, the use of idiosyncratic compared to sample-averaged predictors was assessed (e.g., individual-averaged blueness ratings and CS+ identifications). We used the Bayesian Information Criterion (BIC) for model selection, with lower values indicating a better fit. The BIC is often used to compare non-nested models and penalizes the number of parameters to avoid overfitting. A ΔBIC of > 10 is considered strong evidence in favor of the model with the lowest BIC (Raftery, 1995). In the case of a ΔBIC of <10, we preferred the most parsimonious model.

For the CS+ identification data, analogous to Zaman, Ceulemans, et al. (2019), we explored whether meaningful clusters could be identified for the CS+ identification gradients using the kmeans algorithm from the R stats package v3.6.2 (R Core Team, 2013). To this end, participants are positioned within a 14-dimensional space with coordinates on each dimension corresponding to CS+ identification response probabilities for a given stimulus. The algorithm repeatedly divides the whole sample of participants into an incrementing number of clusters so that the distance between participants and cluster centroids is minimized until the maximum number of clusters is reached. We set the maximum number of clusters to 10. Each participant was allocated to the cluster for which their squared Euclidean distance to the cluster centroid was minimal. The centroids were computed as the mean score profile per cluster. To avoid ending in a local optimum, we ran the analysis 10,000 times, each time using a different random initialization of the centroid matrix (Hofmans et al., 2015). We used the screen ratio test and the silhouette width (Rousseeuw, 1987), with values of the latter $>.25$ indicative of meaningful clustering, to decide if any meaningful clustering occurred and, if so, on the optimal number of clusters (see the online supplemental materials). Additional analyses on the effect of CS identification clusters on US expectancy ratings can be found in the online supplemental materials.

A final analysis focused on the difference between rule types (linBlue vs. simBlue) for the various measures. For the conditioning data, this was done by adding Rule type and its interactions to the aforementioned model. For the generalization data, differences in US expectancy gradients between Rule types were in a first step assessed using a GLM with Stimulus as a categorical predictor (S1:S14), Rule type, and their interaction (as was done in Lovibond et al., 2020). As a sensitivity analysis, we next investigated whether the inclusion of Rule type within the final idiosyncratic model of the previous analyses further improved model fit based on the BIC. For the CS+ identification data, a GLM with Stimulus as a categorical predictor (S1:S14), Rule type, and their interaction was fitted. In addition, we investigated whether there was a relationship between rule type and the outcome of the cluster analysis (that grouped participants based on their CS identification data). Bayes factors (BF) and Cramer's V effect sizes for the contingency tables were obtained with BayesFactor (Morey & Rouder, 2021) and the effectsize package (Ben-Shachar et al., 2020) in R. For the color evaluation data (i.e., categorizations and ratings) (r_{ij}), a logistic function was fitted with the individual shape parameters (κ, β) being drawn from a group distribution with its mean (μ), depending on Rule (G_j) type in JAGS.[3] Parameter estimation was performed using Markov Chain Monte Carlo (MCMC) with the Gibbs sampling method (Casella & George, 1992) through JAGS (Depaoli et al., 2016) in R (R Core Team, 2013). We ran three MCMC chains per model, with 30,000 iterations, of which the first 1,000 were discarded (i.e., burn-in). Samples were thinned by a factor of 2 for each chain (i.e., only each second sample was retained), resulting in 14,500 samples for each parameter. The sample convergence was assessed through the computation of the R-hat (Gelman–Rubic) statistic based on the three sample chains. All R-hats were below 1.01, indicating that MCMC convergence was reached. Visual inspection of the chains confirmed this as they all resembled "a fat hairy caterpillar" without any trend. We calculated Bayes Factors (BF) based on the change in the probability that the standardized difference (δ) between the means of these group distributions was equal to 0 under the prior compared to its probability under the posterior distribution (M. D. Lee & Wagenmakers, 2013). As a rule of thumb, a BF larger >10 is considered strong evidence that the standardized difference differed from 0, while a $BF < .10$ indicated strong evidence for the H_0 (i.e., $\delta = 0$) (Jeffreys, 1961). A graphical model representation is presented in Figure 2. For the analysis of rule differences in confidence ratings, see the online supplemental materials.

Results

Manipulation Check

The overall patterns of US expectancy ratings reflect differential learning as US expectancy ratings decreased for the CS− and increased for the CS+ throughout trials; Trial, $F(1,2{,}412) = 80.18$, $p < .001$; CS, $F(1, 2{,}412) = 41.68$, $p < .001$; CS × Trial, $F(1, 2{,}412) = 210.40$, $p < .001$.

Comparing Various Operationalizations of the Test Dimension

In Table 2, various ways of operationalizing the test dimension were compared using the BIC with lower values indicating a better fit. This is not an exhaustive list, as we only compared different theoretical plausible combinations of predictors and the most common parametrizations of the physical stimulus dimension (e.g., stimulus as a categorical predictor, stimulus as a linear predictor). We did not test combinations of predictors that would be theoretically

[3] This deviates from the preregistration. This was done due to fitting problems with the quickpsy package and additionally has the advantage that uncertainty regarding parameter estimates are taken into account when calculating BFs.

Table 1
Different Operationalizations of the Test Dimension

Operationalization	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14
Stimulus	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14
Linear	1	2	3	4	5	6	7	8	9	10	11	12	13	14
CS+ distance	3	2	1	0	1	2	3	4	5	6	7	8	9	10
CS− distance	11	10	9	8	7	6	5	4	3	2	1	0	1	2
Greenness (VAS)	2.88	3.54	4.20	5.36	7.14	10.22	18.41	33.31	53.75	71.78	83.88	89.92	93.54	94.50
Blueness (VAS)	98.00	97.97	96.41	93.15	90.171	84.53	75.97	59.85	37.43	19.99	12.29	7.45	5.11	4.72
CS+ identification	0.45	0.50	0.58	0.63	0.52	0.37	0.30	0.22	0.17	0.16	0.14	0.15	0.19	0.18
Cat. blue	1	1	1	1	1	1	1	0	0	0	0	0	0	0
Cat. green	0	0	0	0	0	0	0	0	0	1	1	1	1	1
Prot. blue	1	1	1	1	1	1	0	0	0	0	0	0	0	0
Prot. green	0	0	0	0	0	0	0	0	0	0	1	1	1	1

Note. cat. = color category with cutoff p(average categorization) $> .8$, prot. = prototype region defined by average color rating >80. CS = conditioned stimulus; VAS = visual analog scale. See the online article for the color version of this table.

incongruent (e.g., modeling generalization as dependent upon physical and perceived stimulus features at the same time). In a first step, we found that a combination of two continuous predictors best predicted the pattern of US expectancy gradients: a blueness dimension (obtained by averaging blueness ratings across the sample) and a CS memory identification dimension (obtained by averaging CS+ identifications across the sample) (Figure 3).[4] Next, based on sample-averaged color category probabilities and blueness ratings, we derived additional binary predictors that coded whether a stimulus belonged on average to, for instance, the category blue or not (called cat. blue, with p(average categorization) $> .80 = 1$ else 0) or was a blue prototype (called prot. Blue, if average blueness $> 80 = 1$ else 0). This was done to investigate whether there was an added effect of conceptual structures embedded within the test dimension (e.g., color category, color prototype) on US expectancy patterns within the best fitting model of step 1 (see Table 2).[5] The inclusion of these conceptual predictors did not lead to model improvement. Finally, the use of idiosyncratic compared to sample-averaged predictors was assessed by comparing the model from Step 1 to a model with blueness and CS+ identifications as predictors, but this time with individual-averaged blueness ratings and CS+ identifications. Using these idiosyncratic predictors substantially improved model fit (BIC reduction of 96). Finally, we reexamined the effect of conceptual structures within this idiosyncratic model by extending it with idiosyncratic category predictors. Results were somewhat mixed. BIC comparison suggested a slight preference for the simpler model (BIC: 13,240 vs. 13,242), while the chi-squared test statistic indicated a significant reduction in model deviance by adding this additional predictor ($\chi^2 = 5.47$, $df = 1$, $p = .0194$). To illustrate differences in model predictions (sample-averaged model vs. idiosyncratic model), we clustered (see next section) participants into subgroups using the predictor with the most interindividual variability (i.e., CS+ identification) (see the online supplemental materials for illustrative examples with blueness ratings). As visible in Figure 4B, the predictions from the idiosyncratic model can capture distinct US expectancy gradients between participants due to their differences in the CS+ identification pattern (see the online supplemental materials for an additional GLM on the effects of cluster on US expectancy ratings).

Cluster Analyses

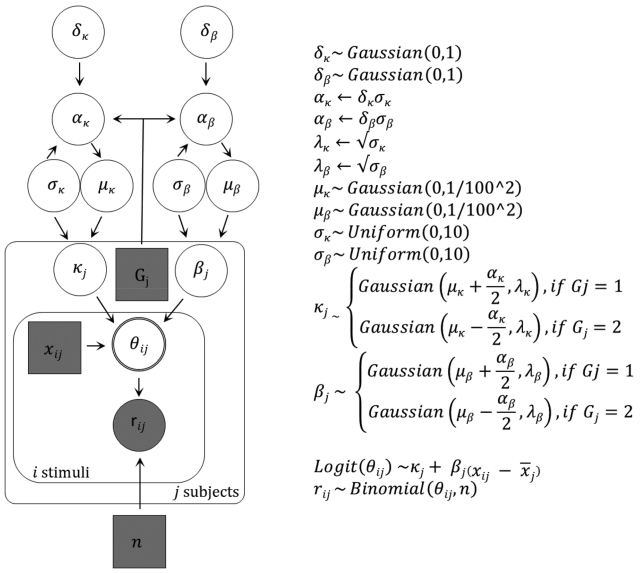
For the CS+ identification data, various cluster solutions reached a silhouette width >0.25. Based on the screen ratio, a four-cluster solution was preferred (see the online supplemental materials)

[4] Note that a model with greenness ratings instead of blueness ratings yielded a very similar BIC of 13496 due to the almost perfect negative correlation between blueness and greenness ratings ($r = -0.92$, $p < .001$).

[5] One may ask whether the color ratings and color categorizations capture truly distinct variables. We consider their difference to be akin to the distinction made in threshold theories of categorization, where a continuous variable like typicality or similarity to a prototype forms the basis for binary category membership decisions (Hampton, 1995; Verheyen et al., 2010). While at the average level, this will lead to two highly correlated variables, at the individual level there is a clear distinction in the nature of the variables, one continuous and the other binary.

Figure 2

Graphical Model for the Analysis of the Color Rating and Categorization Data



Note. Individual averaged categorization probabilities or color ratings (r_{ij}) came from a binomial distribution with a given number of trials (n) and a response probability (θ_{ij}) per individual (j) and stimulus (i). This response probability came from a logistic function with two shape parameters (κ_j, β_j) and the difference between each stimulus (x_{ij}) and the average stimulus value. β_j determines the steepness of the slope with higher absolute values reflecting a steeper increase. κ_j is the stimulus value of the logistic function's midpoint. As the stimulus range was mean-centered, positive values indicate a midpoint located to the left of the mean and negative values a midpoint that is located to the right of the center of the stimulus range when $\beta_j > 0$. The opposite holds when $\beta_j < 0$. Individual estimates for both κ and β came from group distributions with their means (μ) depending on Rule type (G_j) and differing with a value of α between rule subgroups. The standardized difference (δ) is based on α and the variance of the group distributions (σ) for κ and β separately. The prior distribution of each standardized difference was centered around 0 with a precision of 1. Normal distributions are expressed in jags using precision ($1/SD^2$) instead of SD. Squares and circles denote discrete and continuous random variables, respectively. Empty and shaded shapes reflect hidden versus observed values.

(Figure 4A). Twenty-three (22.1%) participants were allocated to cluster 1, 18 (17.3%) to cluster 2, 20 (19.2%) to cluster 3, and 43 (41.4%) to cluster 4.[6]

Difference Between Rule Subgroups

Acquisition

The pattern of US expectancy ratings differed between rule subgroups; Rule effect, $F(1, 472) = 4.48$, $p = .035$; CS × Rule effect, $F(1,2,409) = 2.54$, $p = .11$; Rule × CS × Trial, $F(1,2,409) = 13.62$, $p < .001$. In the *simBlue* and the *linBlue* group, US expectancy ratings increased for the CS+ (with a steeper increase for the simBlue group: $B_{diff} = 2.69$, $SE = 0.76$, $p < .001$) but only decreased for the CS− in the simBlue group (although the difference with the linBlue group failed to reach significance: $B_{diff} = -0.77$, $SE = 0.54$, $p = .15$) (Figure 5A, see the online supplemental materials for posthoc analyses per subgroup).

Generalization Phase

As expected, rule subgroups differed on their US expectancy gradients; Rule effect, $F(1, 105) = 8.38$ $p = .004$; Rule × Stimulus (categorical predictor, S1:S14) effect, $F(13, 1,365) = 2.91$, $p < .001$ (Figure 5B). As a sensitivity analysis, we investigated Rule effects within the sample-averaged predictor model and the final idiosyncratic predictor model. Results were somewhat mixed with BIC's not always indicating a clear preference between models with and without rule, while chi-squared tests consistently favored the models that included rule and its interactions with blueness and CS+ identification (see the online supplemental materials). As results were similar for both models we only report the output of the idiosyncratic model. A peculiar finding was that within the idiosyncratic model, increases in blueness ratings and CS identifications were associated with a reduced increase in US expectancy ratings in the linear compared to the similarity rule subgroup. ($\Delta\beta_{Blueness} = -0.085$, $SE = 0.0358$, $p = .017$; $\Delta\beta CS+_{identification} = -8.80$, $SE = 4.005$, $p = .028$). This contradicts the content of the reported rules and the pattern observed when plotted across the physical dimension, requiring some elaboration. β estimates within the idiosyncratic model indicate the effect on US expectancy for a fixed change of a given predictor while keeping all other predictors in the model constant. Although the increase in US expectancy is lower when CS identifications increase, the stronger increase in CS identification in the linear rule group results in a net steeper US expectancy. In the Figure S7 of the online supplemental materials, rule differences in predicted US expectancy gradients are plotted separately for simulated blueness patterns and CS identifications.

CS+ Identification

As visible in Figure 5C, different patterns in CS+ identification accuracies were observed between rule subgroups; Rule effect, $F(1,103) = 16.97$, $p < .001$; Rule × Stimulus (categorical predictor, S1:S14) effect, $F(13, 1,330) = 2.94$, $p < .001$. These differences were mainly found where groups were expected to differ the most (i.e., S1 and S2). Apart from S1 and S2, differences at S8 were found (see the online supplemental materials). Furthermore, we tested whether there is a relationship between CS+ identification cluster and rule. We found that there was a strong relationship between cluster and rule ($\chi^2 = 306.08$, $df = 3$, $p < .001$; BF > 10.000, $V_c = 0.46$, 95% CI [0.27, 1]; Figure 4C).

Color Evaluation

Table 3 displays the posterior parameter means and their 95% credible intervals (CIs) from the logistic function for the various color evaluation datasets per rule. For the categorization datasets, we found substantial evidence in the blue dataset and strong evidence in the green dataset in favor of a slope difference (β) between rule subgroups (blue: BF = 7.75, green: BF = 42.38) with steeper slopes in the simBlue compared to the linBlue subgroup. As visible in Figure 5D–E, these differences, however, were very subtle. No evidence was found for differences in the inflection point (κ) between subgroups in either dataset (blue: BF = 0.26, green: BF = 0.42).

[6] For one participant the clustering was not possible due to a lack of identification responses for certain stimuli.

Table 2
Overview of Tested GLMs

Stim.	Lin.	d. CS+	d. CS−	gVAS	bVAS	CS+ id.	c. b	c. g.	p.b.	p.g.	BIC	n.par
x											13,569	16
	x										13,625	4
		x									13,652	4
			x								13,620	4
		x	x								13,535	5
				x							13,598	4
					x						13,558	4
				x	x						13,536	5
						x					13,564	4
		x	x			x					13,505	6
		x				x					13,523	5
			x			x					13,528	5
				x	x	x					13,502	6
					x	**x**					**13,495**	**5**
				x		x					13,496	5
	x					x					13,510	5
					x	x	x				13,502	6
					x	x		x			13,502	6
					x	x			x		13,500	6
					x	x				x	13,502	6

Note. Stim. = stimulus (categorical). Lin. = stimulus as continuous predictor (linear). A second order polynomial of stimulus yielded a worse BIC of 13,632; highligted in bold = preffered model based on BIC comparison; d. CS+ = CS+ distance. d. CS− = CS− distance. gVAS = greenness ratings. bVAS = blueness ratings. CS+ id = CS+ identifications. c.b. = blue category. c.g. = green category. p.b. = blue prototype. p.g. = green prototype. n.par. = number of parameters in the GLM. CS = conditioned stimulus; VAS = visual analog scale; BIC = Bayesian Information Criterion; GLM = Generalized Linear Model.

For both color rating data sets, no evidence was found for subgroup differences in any of the parameters; blue: BF (β) = 0.22, BF (κ) = 0.25; green: BF (β) = 0.39, BF (κ) = 0.27 (Figure 5F, G).

Discussion

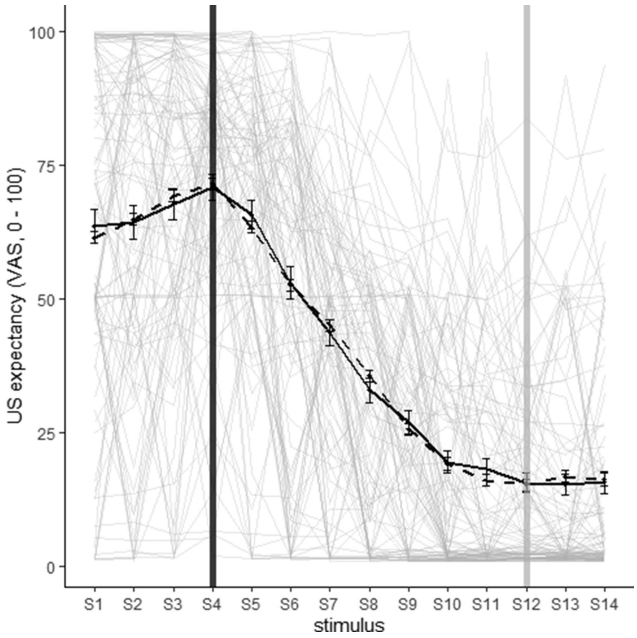
We investigated how differences in generalization gradients across blue-green color patches can be understood from interindividual differences in color perception and categorization, CS+ identification, and response rules. A response model with color perception and CS+ identification performance as sole predictors was preferred to contemporary approaches that use stimulus as a predictor. Interestingly, incorporating interindividual differences in color perception and CS+ identification significantly improved the models' ability to account for different generalization patterns (Figure 4). Additionally, we found some evidence for the involvement of idiosyncratic conceptual color categories and response rules. Our findings suggest that generalization research should adopt models incorporating differences in perception, memory, and category structures between subjects to arrive at more insightful and correct inferences regarding the extent and existence of a latent generalization process.

Learning and generalization models, capable of explaining a wide range of behaviors, have devoted relatively little attention to how mental stimulus representations relate to our physical surroundings, despite their potential to foster insight into the causes of interindividual generalization differences (see Zaman, Chalkia, et al., 2021). Although some theories acknowledge the stochastic nature of mental representations, analytic customs equate the mental and the physical in practice. As a corollary hereof, all behavioral variations are, by default, attributed to differences in a latent generalization process. The alternative possibility that behavioral differences emerge due to processes other than generalization, like differences in perception or memory of the training stimulus, is hardly explored. We investigated to what extent differences in the response pattern during a generalization test can be understood from the actors' perception and memory. We found that the shape of the US expectancy gradient was best predicted by and, on average, very closely approximated by a response model that incorporated color perception and CS+ identification performance.

As we observed interindividual differences in both color perception and CS+ identification performance, with especially distinct patterns between individuals in the latter, we tested their potential to account for differences in generalized responding. The use of idiosyncratic color perception and memory predictors (compared to group-averaged predictors) further improved the model's ability to account for distinct generalization patterns between individuals. As seen in Figure 4B, the model's predictions better approximated US expectancy gradients than a model ignorant of individual differences in perception and memory. To visualize these effects, we clustered participants on their CS identification data (and on blueness ratings, see the online supplemental materials). In Clusters 1 and 2 compared to Cluster 4, higher US expectancies at S1–S3, for instance, stem from an increased probability of being mistaken for the CS+ (S4). Combined with recent work demonstrating the strong impact of misidentifications on both learning and generalization, these findings challenge contemporary conceptions about generalization (Struyf et al., 2017; Zaman, Ceulemans, et al., 2019; Zaman, Struyf, et al., 2019, 2020). In all of these studies, a large extent of the behavioral differences during a generalization test was not due to differences in a latent generalization process. On the contrary, model parameters capturing generalization tendencies (i.e., regression β's) were fixed across subjects. The current study extends these findings by demonstrating that the incorporation of differences in perception and memory can explain differences in response gradients rather

Figure 3

US Expectancy Gradient



Note. Mean observed (solid) and predicted (dashed) US expectancy with sample-averaged blueness and CS+ identification as predictors. Black bar = CS+, grey bar = CS−.

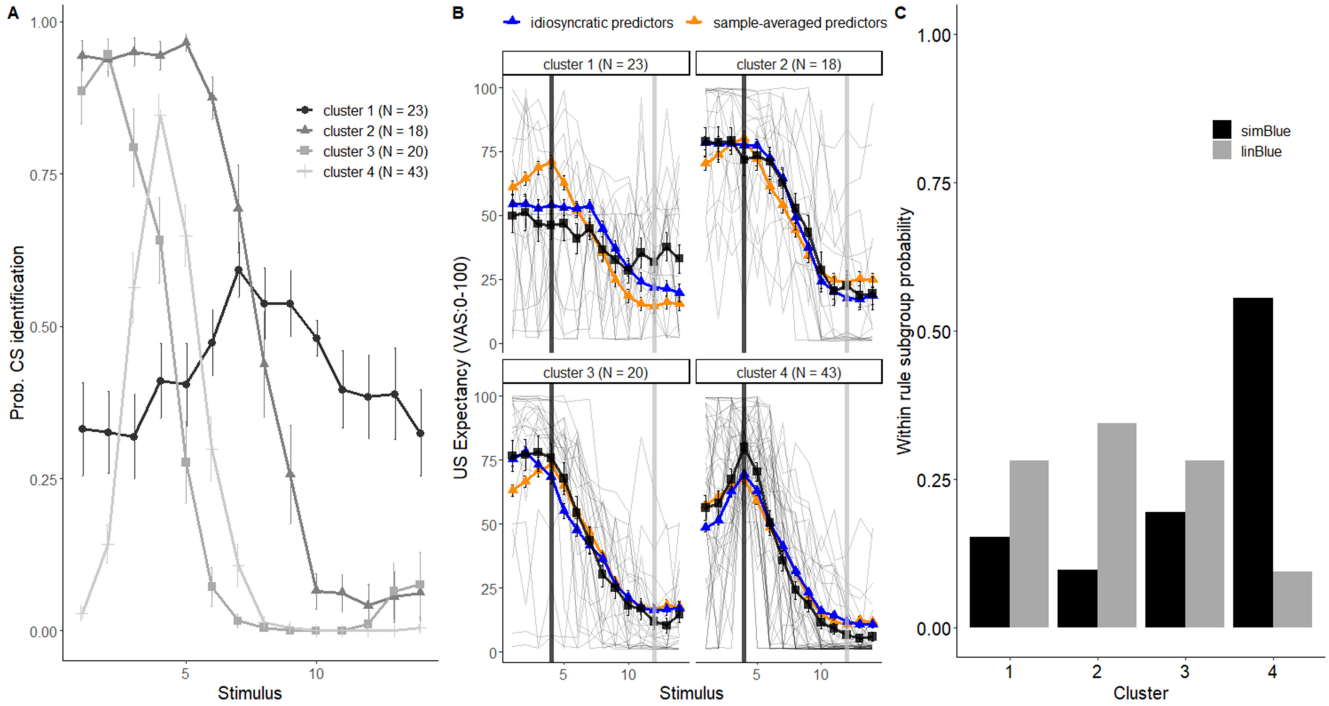
than the default assumption to attribute them to differences in a latent generalization process.

Another advantage of the adopted approach concerns the issue of parameter comparability between subjects. Traditional analyses of response gradients using rmANOVAs do not provide a straightforward index that parametrizes generalization tendencies. More recent modeling approaches overcome this by fitting Gaussian distributions with the standard deviation capturing generalization tendencies (J. C. Lee et al., 2021; Schlegelmilch et al., 2022). However, these approaches still assume that perception and memory are accurate representations of reality (i.e., veridicality assumption), because they use the physical stimulus dimension as the independent variable (see Zaman, Chalkia, et al., 2021). By not accounting for differences in memory or perception, comparing parameter estimates or generalization indices between individuals or groups amounts to comparing apples and oranges. By expressing generalization tendencies as the expected response increment when color perception or the chance of CS+ identification varies, an index emerges that is comparable between people despite having a different perceptual sensitivity or CS+ memory accuracy. These findings are promising as they demonstrate the added value of a multimodal assessment approach and open the avenue for a field where knowledge from perception, memory, and learning research is combined in models that enable researchers better to understand the various mechanisms and their interactions at play.

We found inconsistent evidence on the additional effect of conceptual structures (i.e., prototype regions or color categories) on US expectancy gradients. There may be multiple reasons for this: (a) The bulk of studies demonstrating conceptual generalization (e.g., Dunsmoor et al., 2011; Dunsmoor & Murphy, 2014, 2015; Wong & Beckers, 2021) paired multiple exemplars to the US during conditioning rather than one exemplar (many vs. one CS+). However, a recent preregistered report found some conceptual generalization evidence after only using one exemplar during training (Mertens et al., 2021). (b) It is possible that our CS+ was not sufficiently typical, as conceptual generalization from typical to atypical category exemplars is easier than vice versa (Dunsmoor & Murphy, 2014). However, the high blueness ratings with little variation at the CS+ make this an unlikely explanation. Interestingly, the CS+ identification gradients within specific clusters suggest a pattern of response probabilities that closely parallel color category membership (see cluster 2, Figure 4A, B) with high and equal identification probabilities for S1-S5. Future research should investigate to which extent these findings reflect the broadening of the CS+ representation from a single or multiple exemplars to the entire category. Furthermore, in cluster 1 (Figure 4A), identification responses did not peak at the CS+ but at a more prototypical color, also suggesting a shift in the representation of the CS+ away from the CS− analogous to the notion of category caricatures that arise in the context of contrasting categories (Ameel & Storms, 2006; Davis & Love, 2010). (c) The most likely explanation for our findings regarding conceptual structures is that in our study, blueness ratings were very high, with little variation between subjects for most stimuli belonging to the blue category. Hence, including the categorical blue predictor on top of average blueness ratings may lead to too little differential information (i.e., collinearity problem). A (multidimensional) stimulus set where category structures and perceptual dimensions are less intertwined would be an exciting avenue to explore this research question further. As in taxonomically organized categories, perceptual and conceptual dimensions only come apart at levels situated at the so-called superordinate level, this may require moving up in the conceptual hierarchy to break the link between percept and concept. For instance, superordinate category members do not necessarily have the same shape (e.g., animals), unlike exemplars of basic-level (e.g., dogs) or subordinate level categories (e.g., Pekinese; Rosch et al., 1976) do.

In addition to the collinearity between the physical and the conceptual representations of the color patches, the stimuli in the current study also differed along a single dimension only. They constituted approximately equidistant color patches between a blue and a green prototype. As a result, any interindividual differences in the stimuli's conceptual representation could only be a matter of degree (i.e., the extent to which different color patches are considered instances of blue) and not a matter of criteria, where different participants consider different dimensions as relevant or weigh several underlying dimensions differently to establish category membership (Verheyen, Droeshout, & Storms, 2019; Verheyen & Storms, 2018; Verheyen, White, & Égré, 2019). Moving to more abstract, multidimensional stimuli, as suggested earlier, will allow one to study whether degree and/or criteria differences in conceptual representations affect generalization. In addition, individual conceptual differences will then likely be more in play than for color categories, for which it is generally assumed that their prototypes are largely socially shared (e.g., Douven, 2016; Douven et al., 2017). Although the choice to use a color dimension was made in order to build further upon earlier work (i.e., Douven et al., 2017; Lovibond et al., 2020), this specific stimulus set may pose an underestimation as to the involvement of interindividual differences. Relative little variation in color perception (compared to CS

Figure 4
Cluster Analysis



Note. (A) CS+ identification response probabilities per cluster. (B) The black line represent the mean US expectancy gradients (black), the orange line (light grey) predictions using sample-averaged predictors and the blue line predictions with the idiosyncratic predictors (dark grey) per cluster. (C) Within rule type probabilities for each of the different clusters. Black bar = CS+, grey bar = CS−. See the online article for the color version of this figure.

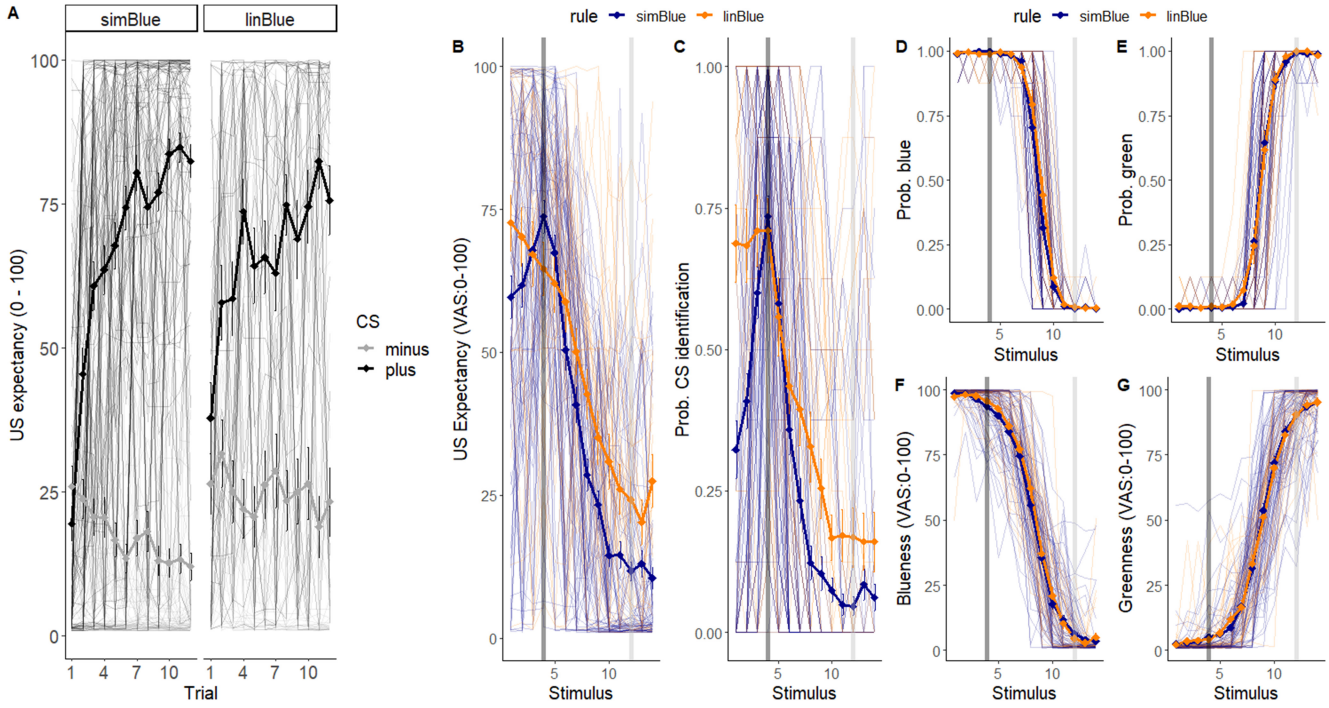
identification) was observed. Given the strong ties of color perception to our perceptual organs, its roots in the physical world, and cross-culturally stability, other stimulus dimensions may be more suited to study individual differences in perception. With less artificial stimuli, subjective appraisals of prototypicality and category membership are also expected to increase due to pronounced inter-individual differences in prior experiences (Verheyen & Égré, 2018; Verheyen & Storms, 2018) and personal interests (Verheyen et al., 2015, 2018).

The next aim of the study was to investigate possible causes for why participants adopted different response rules. One hypothesis was that a priori differences in perception determine which rule one adopts (Zaman, Yu, & Lee, 2022). We did not find support for this idea as subgroups did not differ on psychometric shape parameters for either color perception dataset. We did observe differences between subgroups with a higher slope in the similarity rule group than in the linear rule group for the color categorization data. However, an inspection of the data revealed that the magnitude of these differences was so small that it seems unlikely that these differences may explain why participants adopted different response rules. Thus, overall, both groups entered the conditioning phase with very minor or no different views on the test dimension. Next, we assessed differences in learning. We found that the linear rule compared to the similarity rule demonstrated somewhat poorer learning. US expectancy ratings increased less steeply for the CS+ across trials and did not decrease for the CS− in the linear rule group (although the group difference was not statistically significant for the CS− between subgroups). This pattern fits with an alternative hypothesis that attentional differences during learning determine response rules. Poorer attention during learning may cause people to adopt a response rule requiring less precise encoding of the stimulus features (in our case the linear rule; Zaman, Yu, & Lee, 2022). In accordance with this was the pattern of CS+ identification between subgroups, which suggests less precise encoding or retrieval of CS characteristics from memory in the linear subgroup. Furthermore, poorer differentiation in confidence ratings between the various rule options within this subgroup compared to the similarity subgroup again points to poorer learning or memory (see the online supplemental materials). It remains unclear to which extent differences in CS identification really reflect differences in attention during learning or reflect more general memory differences.

Finally, compared to previous work (Lovibond et al., 2020), where US expectancy ratings in the linBlue subgroup plateaued from a specific point despite further increases in color, we did not observe such a pattern. On the contrary, in our study US expectancy ratings kept increasing as the stimuli became more prototypical blue, making response patterns more in line with the adopted rule (i.e., "the bluer the stimulus, the more likely the shock to come"). One explanation for these different response patterns may be the use of different stimulus sets. In Lovibond et al. (2020), test stimuli were created by varying hues while keeping saturation and brightness constant. While these stimuli are equispaced along the physical hue scale, they need not be equidistant in perceptual space, potentially resulting in poorer discriminability between stimuli at specific

Figure 5

Rule Comparisons



Note. (A) Mean US expectancy ratings per rule subgroup across conditioning trials for CS+ trials (black) and CS- trials (grey), (B) US expectancy ratings across the test dimension per subgroup during the generalization phase. (C) Probabilities of stimuli being identified as the CS+ during the identification phase. Averaged categorization (D–E) and color ratings (F–G) per rule type for the different color tasks. Blue (dark grey) = simBlue subgroup, orange (light grey) = linBlue subgroup, The thin lines represent individual averages and the error bars denote standard errors of the mean. See the online article for the color version of this figure.

points, which would foster similar levels of responding to those stimuli. In the current study, as stimuli were created based on their positions in CIELUV space (Malacara, 2002)—a space developed by color scientists so that distances in this 3D space better approximate perceptual distances—stimuli should be perceptually equidistant. Furthermore, a recent re-analysis of Lovibond et al. (2020) did find a close relationship between identification performance and generalization patterns (Zaman, Yu, & Lee, 2022). However, the few stimulus repetitions in the original study (1 trial per stimulus) made it impossible to obtain CS identification gradients on the individual level, allowing the authors to demonstrate the relationship between CS identification and generalization only on the subgroup level. This study nicely extended these findings by demonstrating the close relationship between CS identification and generalization on the individual level. Another explanation for the inconsistent findings between ours and previous work is that here the positioning of the CS near the end of the stimulus spectrum left little room for the emergence of clear differences, as typically rule differences are expected at test stimuli located on the non-CS− side of the CS+.

Some limitations should be acknowledged: First, the assessment of color ratings and categories prior to learning and generalization testing may have probed participants regarding these aspects, thereby potentially affecting subsequent behaviors. For instance, this lack of counterbalancing may have primed participants to organize the stimulus set as color categories and generalize accordingly. Yet, if true, we argue that more profound effects of color category on generalization behavior would be expected than those observed. Future studies are needed to scrutinize to what extent such priming effects may have influenced our findings. Another limitation concerns the point of assessment of the CS identifications, which in our study was after generalization testing. This exposure to multiple perceptually similar stimuli prior to the identification task may have affected participants' performance by blurring or biasing memory. However, in a recent study (Zaman, Yu, & Lee, 2022), we found relatively similar identification gradients

Table 3

Posterior Parameters Means

Parameter	linBlue par [95% CI]	simBlue par [95% CI]
	Blue categorization	
β	−51.20 [45.94, 56.60]	−59.19 [55.28, 63.74]
κ	−2.55 [1.88, 3.24]	−2.33 [1.88, 2.79]
	Green categorization	
β	52.43 [47.02, 57.97]	63.14 [58.63, 67.77]
κ	−2.28 [−3.02, −1.54]	−2.78 [−3.31, −2.26]
	Blueness ratings	
β	−23.19 [19.94, 26.41]	−24.01 [21.83, 26.17]
κ	−0.83 [0.52, 1.12]	−0.69 [0.49, 0.89]
	Greenness ratings	
β	21.02 [17.87, 24.19]	23.17 [21.05, 25.29]
κ	−1.05 [−1.33, −0.78]	−1.17 [−1.35, −0.99]

when identification testing preceded rather than followed generalization testing (unpublished analysis).

In conclusion, the current study adopted a multimodal assessment approach where other aspects were assessed than solely learning-based generalization. The assessment of color perception, conceptual structures, and CS identification enabled us to map inter-individual differences. We found that individuals differed on many of these aspects, with these differences substantially contributing to different response gradients during generalization testing. Our findings suggest that generalization research should depart from the custom of interpreting and analyzing generalized responses along the manipulated physical stimulus dimension. Rather it should incorporate the idiosyncratic nature of how individuals perceive, represent, and remember our surroundings and how this can impact post-learning behaviors in multiple manners.

Constraints on Generality

We expect our results to generalize to other lexicalized color categories and believe the results will be reproducible with adult participants from other crowdsourcing platforms or with student participants tested in a laboratory setting. We lack evidence showing that the results will generalize to other categories as it is yet unclear what the nature and extent are of the individual differences in perception, representation, memory, and rule use in other domains. As mentioned in the limitations section, it also remains to be seen what effect counterbalancing the position of the color and identification tasks has on the results. We have no reason to believe that the results depend on other characteristics of the materials, participants, or context.

References

Ameel, E., & Storms, G. (2006). From prototypes to caricatures: Geometrical models for concept typicality. *Journal of Memory and Language*, *55*(3), 402–421. https://doi.org/10.1016/j.jml.2006.05.005

Atkinson, R. C., & Estes, W. K. (1963). Stimulus sampling theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (3rd ed., pp. 121–268). Wiley.

Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101–140). Cambridge University Press.

Barsalou, L. W. (1989). Intraconcept similarity and its implications for interconcept similarity. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 76–121). Cambridge University Press. https://doi.org/10.1017/CBO9780511529863.006

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). Effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), Article 2815. https://doi.org/10.21105/joss.02815

Blough, D. S. (1975). Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*, *1*(1), 3–21. https://doi.org/10.1037/0097-7403.1.1.3

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167–174. https://doi.org/10.1080/00031305.1992.10475878

Davis, T., & Love, B. C. (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science*, *21*(2), 234–242. https://doi.org/10.1177/0956797609357712

Decock, L., & Douven, I. (2014). What is graded membership? *Noûs*, *48*(4), 653–682. https://doi.org/10.1111/nous.12003

Depaoli, S., Clifton, J. P., & Cobb, P. R. (2016). Just another Gibbs sampler (JAGS) flexible software for MCMC implementation. *Journal of Educational and Behavioral Statistics*, *41*(6), 628–649. https://doi.org/10.3102/1076998616664876

Douven, I. (2016). Vagueness, graded membership, and conceptual spaces. *Cognition*, *151*, 80–95. https://doi.org/10.1016/j.cognition.2016.03.007

Douven, I. (2019). Putting prototypes in place. *Cognition*, *193*, Article 104007. https://doi.org/10.1016/j.cognition.2019.104007

Douven, I., Wenmackers, S., Aissati, Y., & Decock, L. (2017). Measuring graded membership: The case of color. *Cognitive Science*, *41*(3), 686–722. https://doi.org/10.1111/cogs.12359

Dunsmoor, J. E., & Murphy, G. L. (2014). Stimulus typicality determines how broadly fear is generalized. *Psychological Science*, *25*(9), 1816–1821. https://doi.org/10.1177/0956797614535401

Dunsmoor, J. E., & Murphy, G. L. (2015). Categories, concepts, and conditioning: How humans generalize fear. *Trends in Cognitive Sciences*, *19*(2), 73–77. https://doi.org/10.1016/j.tics.2014.12.003

Dunsmoor, J. E., White, A. J., & LaBar, K. S. (2011). Conceptual similarity promotes generalization of higher order fear learning. *Learning & Memory*, *18*(3), 156–160. https://doi.org/10.1101/lm.2016411

Dymond, S., Dunsmoor, J. E., Vervliet, B., Roche, B., & Hermans, D. (2015). Fear generalization in humans: Systematic review and implications for anxiety disorder research. *Behavior Therapy*, *46*(5), 561–582. https://doi.org/10.1016/j.beth.2014.10.001

Enquist, M., & Ghirlanda, S. (2005). *Neural networks and animal behavior*. Princeton University Press. https://doi.org/10.1515/9781400850785

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*(1), 15–36. https://doi.org/10.1006/anbe.2003.2174

Ghirlanda, S., & Enquist, M. (2007). How training and testing histories affect generalization: A test of simple neural networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1479), 449–454. https://doi.org/10.1098/rstb.2006.1972

Green, P., & Macleod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Hampton, J. A. (1995). Testing the prototype theory of concepts. *Journal of Memory and Language*, *34*(5), 686–708. https://doi.org/10.1006/jmla.1995.1031

Hampton, J. A., & Passanisi, A. (2016). When intensions do not map onto extensions: Individual differences in conceptualization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(4), 505–523. https://doi.org/10.1037/xlm0000198

Hofmans, J., Ceulemans, E., Steinley, D., & Van Mechelen, I. (2015). On the added value of bootstrap analysis for K-means clustering. *Journal of Classification*, *32*(2), 268–284. https://doi.org/10.1007/s00357-015-9178-y

Holt, D. J., Boeke, E. A., Wolthusen, R. P. F., Nasr, S., Milad, M. R., & Tootell, R. B. H. H. (2014). A parametric study of fear generalization to faces and non-face objects: Relationship to discrimination thresholds. *Frontiers in Human Neuroscience*, *8*, 1–12. https://doi.org/10.3389/fnhum.2014.00624

Honig, W. K., & Urcuioli, P. J. (1981). The legacy of Guttman and Kalish (1956): Twenty-five years of research on stimulus generalization. *Journal of the Experimental Analysis of Behavior*, *36*(3), 405–445. https://doi.org/10.1901/jeab.1981.36-405

Hoskin, R., Berzuini, C., Acosta-Kane, D., El-Deredy, W., Guo, H., & Talmi, D. (2019). Sensitivity to pain expectations: A Bayesian model of

individual differences. *Cognition*, *182*, 127–139. https://doi.org/10.1016/j.cognition.2018.08.022

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Jraissati, Y., & Douven, I. (2017). Does optimal partitioning of color space account for universal color categorization? *PLoS ONE*, *12*(6), Article e0178083. https://doi.org/10.1371/journal.pone.0178083

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. https://doi.org/10.1037/xlm0000558

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. https://doi.org/10.1177/1747021820949470

Lee, M. D., & Wagenmakers, E. J. (2013). Bayesian Cognitive modeling: A practical course. In *Bayesian cognitive modeling: A practical course* (pp. 1–264). Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lenth, R. (2022). *emmeans: Estimated marginal means, aka least-squares means* (R package version 1.7.4-1).

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. https://doi.org/10.1037/xlm0000779

Malacara, D. (2002). *Color vision and colorimetry: Theory and applications*. SPIE Press.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472. https://doi.org/10.3758/BF03197480

McLaren, I. P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior*, *28*(3), 211–246. https://doi.org/10.3758/BF03200258

McLaren, I. P. L., & Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior*, *30*(3), 177–200. https://doi.org/10.3758/BF03192828

Mednick, S. A., & Freedman, J. L. (1960). Stimulus generalization. *Psychological Bulletin*, *57*(3), 169–200. https://doi.org/10.1037/h0041650

Mertens, G., Bouwman, V., & Engelhard, I. M. (2021). Conceptual fear generalization gradients and their relationship with anxious traits: Results from a Registered Report. *International Journal of Psychophysiology*, *170*, 43–50. https://doi.org/10.1016/j.ijpsycho.2021.09.007

Morey, R., & Rouder, J. (2021). *BayesFactor: Computation of Bayes factors for common designs* (R package version 0.9.12-4.3).

Norbury, A., Robbins, T. W., & Seymour, B. (2018). Value generalization in human avoidance learning. *eLife*, *7*, Article e34779. https://doi.org/10.7554/eLife.34779

Raftery, A. E. (1995). Bayesian Model selection in social research. *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

R Core Team. (2013). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. http://www.r-project.org/

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Schlegelmilch, R., Wills, A. J., & von Helversen, B. (2022). A cognitive category-learning model of rule abstraction, attention learning, and contextual modulation. *Psychological Review*, *129*(6), 1211–1248. https://doi.org/10.1037/rev0000321

Shepard, R. N. (1958). Stimulus and response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review*, *65*(4), 242–256. https://doi.org/10.1037/h0043083

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Struyf, D., Zaman, J., Hermans, D., & Vervliet, B. (2017). Gradients of fear: How perception influences fear generalization. *Behaviour Research and Therapy*, *93*, 116–122. https://doi.org/10.1016/j.brat.2017.04.001

Struyf, D., Zaman, J., Vervliet, B., & Van Diest, I. (2015). Perceptual discrimination in fear generalization: Mechanistic and clinical implications. *Neuroscience and Biobehavioral Reviews*, *59*, 201–207. https://doi.org/10.1016/j.neubiorev.2015.11.004

Vanbrabant, K., Boddez, Y., Verduyn, P., Mestdagh, M., Hermans, D., & Raes, F. (2015). A new approach for modeling generalization gradients: A case for hierarchical models. *Frontiers in Psychology*, *6*, Article 652. https://doi.org/10.3389/fpsyg.2015.00652

Verheyen, S., Dewil, S., & Égré, P. (2018). Subjectivity in gradable adjectives: The case of tall and heavy. *Mind & Language*, *33*(5), 460–479. https://doi.org/10.1111/mila.12184

Verheyen, S., Droeshout, E., & Storms, G. (2019). Age-related degree and criteria differences in semantic categorization. *Journal of Cognition*, *2*(1), Article 17. https://doi.org/10.5334/joc.74

Verheyen, S., & Égré, P. (2018). Typicality and graded membership in dimensional adjectives. *Cognitive Science*, *42*(7), 2250–2286. https://doi.org/10.1111/cogs.12649

Verheyen, S., Hampton, J. A., & Storms, G. (2010). A probabilistic threshold model: Analyzing semantic categorization data with the Rasch model. *Acta Psychologica*, *135*(2), 216–225. https://doi.org/10.1016/j.actpsy.2010.07.002

Verheyen, S., & Storms, G. (2013). A mixture approach to vagueness and ambiguity. *PLoS ONE*, *8*(5), Article e63507. https://doi.org/10.1371/journal.pone.0063507

Verheyen, S., & Storms, G. (2018). *Education as a source of vagueness in criteria and degree* (pp. 149–167). Springer. https://doi.org/10.1007/978-3-319-77791-7_6

Verheyen, S., Voorspoels, W., & Storms, G. (2015). Inferring choice criteria with mixture IRT models: A demonstration using ad hoc and goal-derived categories. *Judgment and Decision Making*, *10*(1), 97–114. https://doi.org/10.1017/S1930297500003211

Verheyen, S., White, A., & Égré, P. (2019). Revealing criterial vagueness in inconsistencies. *Open Mind*, *3*, 41–51. https://doi.org/10.1162/opmi_a_00025

Vlaeyen, J. W. S., & Linton, S. J. (2012). Fear-avoidance model of chronic musculoskeletal pain: 12 years on. *Pain*, *153*(6), 1144–1147. https://doi.org/10.1016/j.pain.2011.12.009

White, A., Storms, G., Malt, B. C., & Verheyen, S. (2018). Mind the generation gap: Differences between young and old in everyday lexical categories. *Journal of Memory and Language*, *98*, 12–25. https://doi.org/10.1016/j.jml.2017.09.001

Wong, A. H. K., & Beckers, T. (2021). Trait anxiety is associated with reduced typicality asymmetry in fear generalization. *Behaviour Research and Therapy*, *138*, Article 103802. https://doi.org/10.1016/j.brat.2021.103802

Wong, A. H. K., & Lovibond, P. F. (2017). Rule-based generalisation in single-cue and differential fear conditioning in humans. *Biological Psychology*, *129*, 111–120. https://doi.org/10.1016/j.biopsycho.2017.08.056

Zaman, J., Ceulemans, E., Hermans, D., & Beckers, T. (2019). Direct and indirect effects of perception on generalization gradients. *Behaviour Research and Therapy*, *114*, 44–50. https://doi.org/10.1016/j.brat.2019.01.006

Zaman, J., Chalkia, A., Zenses, A.-K. K., Bilgin, A. S., Beckers, T., Vervliet, B., & Boddez, Y. (2021). Perceptual variability: Implications for learning

and generalization. *Psychonomic Bulletin & Review*, *28*(1), 1–19. https://doi.org/10.3758/s13423-020-01780-1

Zaman, J., & Lee, J. C. (2020). *To perceive or not to perceive? Individual differences in perception predict rule induction and response generalization*. PsyArxiv. https://doi.org/10.31234/osf.io/89ncb

Zaman, J., Struyf, D., Ceulemans, E., Beckers, T., & Vervliet, B. (2019). Probing the role of perception in fear generalization. *Scientific Reports*, *9*(1), Article 10026. https://doi.org/10.1038/s41598-019-46176-x

Zaman, J., Struyf, D., Ceulemans, E., Vervliet, B., & Beckers, T. (2020). Perceptual errors are related to shifts in generalization of conditioned responding. *Psychological Research*, *85*, 1801–1813. https://doi.org/10.1007/s00426-020-01345-w

Zaman, J., Yu, K., & Lee, J. C. (2022). Individual differences in stimulus identification, rule induction, and generalization of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0001153

Zaman, J., Yu, K., & Verheyen, S. (2022). *Data from the paper: "The idiosyncratic nature of how individuals perceive, represent, and remember their surroundings and its impact on learning-based generalization."* https://osf.io/d5s2r/

Zenses, A., Lee, J. C., Plaisance, V., & Zaman, J. (2021). Differences in perceptual memory determine generalization patterns. *Behaviour Research and Therapy*, *136*, Article 103777. https://doi.org/10.1016/j.brat.2020.103777

Received June 21, 2022
Revision received November 28, 2022
Accepted February 9, 2023 ■