AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: Learning, Memory, and Cognition

© 2022 American Psychological Association
ISSN: 0278-7393

2023, Vol. 49, No. 6, 1004–1017
https://doi.org/10.1037/xlm0001153

Individual Differences in Stimulus Identification, Rule Induction, and Generalization of Learning

Jonas Zaman[1, 2], Kenny Yu[3], and Jessica C. Lee[4]

[1] Health Psychology, Faculty of Psychology and Educational Sciences, KU Leuven

[2] Centre for the Psychology of Learning and Experimental Psychopathology, Faculty of Psychology and Educational Sciences, KU Leuven

[3] Quantitative Psychology and Individual Differences, Faculty of Psychology and Educational Sciences, KU Leuven

[4] School of Psychology, University of New South Wales



In the field of stimulus generalization, an old yet unresolved discussion pertains to what extent stimulus misidentifications contribute to the pattern of conditioned responding. In this article, we perform cluster analysis on six datasets (four published datasets and two unpublished datasets, included N = 950) to examine the relationship between interindividual differences in (a) stimulus identification, (b) patterns of generalized responding, and (c) verbalized generalization rules. The datasets were obtained from online predictive learning tasks where participants learned associations between colored cues and the presence or absence of a hypothetical outcome. In these datasets, stimulus identification and expectancy ratings were assessed in separate phases to a range of colors varying between blue-green. Using cluster analyses on performance during stimulus identification, we identified different subgroups of participants (good vs. bad identifiers). In all six datasets, we found a close relationship between the pattern of stimulus identification and the shape of the expectancy gradient across the test dimension between the identified subgroups. Furthermore, participants classified as *good identifiers* were more likely to report a similarity generalization rule than a relational or linear rule, suggesting that individual differences in stimulus identification are related to individual differences in generalization rules. These findings suggest that greater consideration should be given to interindividual variability in stimulus identification, inductive rules, and their relationship in explaining patterns of generalized responses.

Keywords: generalization, stimulus identification, rules, conditioning, individual differences

Supplemental materials: https://doi.org/10.1037/xlm0001153.supp

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

The behaviors of many organisms reflect a transfer of learning as they act similarly across various novel situations (e.g., Enquist & Ghirlanda, 2005; Ghirlanda & Enquist, 2003; Honig & Urcuioli, 1981; Mednick & Freedman, 1960). Stimulus generalization is when novel, unreinforced test stimuli come to elicit a learned response (i.e., a conditioned response, CR) that was acquired to another stimulus (i.e., the conditioned stimulus, CS+) by virtue of its pairing with a motivationally significant stimulus (i.e., the unconditioned stimulus, US). Typically, the averaged response strength (or probability) decreases as a function of physical distance to the trained CS+, producing a symmetrical, bell-shaped generalization gradient across the test dimension (e.g., hertz, hue,

This article was published Online First August 18, 2022.

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

Jessica C. Lee https://orcid.org/0000-0003-4253-2008

Jonas Zaman served as lead for writing, formal analysis and visualizations. Kenny Yu contributed equally to writing–review and editing and served in a supporting role for analysis. Jessica C. Lee contributed equally to writing–review and editing.

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) and a Special Research Funds (FWO, 1500620N). Kenny Yu is supported by an FWO research project (co-PI: Jonas Zaman, G079520N). Jessica C. Lee received salary support from a Discovery grant (DP190103738) and is funded by a Discovery Early Career Research Award (DE210100292), both from the Australian Research Council.

Link to data repository: https://osf.io/g8bjy/. Link to analyses code repository: https://osf.io/g8bjy/. Institutional URL: https://ppw.kuleuven.be/ogp.

The data are available at https://osf.io/g8bjy/

The experiment materials are available at https://osf.io/xnva4/

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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

and size). Early theories of generalization were formulated based on animal research and attempted to explain such response gradients through a similarity-based generalization process, with similarity operationalized as the amount of common activated units (Atkinson & Estes, 1963; Blough, 1975; Enquist & Ghirlanda, 2005; Ghirlanda & Enquist, 2007; Mclaren & Mackintosh, 2000, 2002) or distance between stimuli in an abstract psychological space (Shepard, 1958, 1987). While aggregate response patterns across a wide range of experiments in both animals and humans follow the shape predicted by a similarity-based generalization process (Ghirlanda & Enquist, 2003; Mednick & Freedman, 1960), this is far less the case for individuals, who exhibit a variety of response patterns (Lee et al., 2021; Zaman, Chalkia, et al., 2021). Identifying the causes responsible for this interindividual variability remains a challenge for many associative generalization models (but see Schlegelmilch et al., 2021). As overgeneralization of fear has been implicated in many forms of psychopathology (Dunsmoor & Paz, 2015; Kindt, 2014; Lissek & Grillon, 2010; Vlaeyen & Linton, 2012), the identification of the various underlying mechanisms yielding interindividual differences have important clinical, as well as theoretical implications.

There are multiple ways in which stimulus generalization might occur with potential interindividual differences (Zaman, Chalkia, et al., 2021). A classic debate in the stimulus generalization literature pertains to whether generalization reflects the organism failing to discriminate between the test stimulus and the training stimulus perceptually, or whether it occurs despite the animal recognizing the test stimulus to be different (e.g., Ennis, 1988; Guttman & Kalish, 1956; Kalish, 1958; Lashley & Wade, 1946; Mackintosh, 1974; Mostofsky, 1965; Prokasy & Ball, 1963; Shepard, 1987). In other words, do people generalize conditioned responses from the CS to the test stimulus while recognizing that the test stimulus is different (e.g., Hull, 1943; Shepard, 1987), or did the test stimulus merely elicit the conditioned response because it was misidentified as the training stimulus (Lashley & Wade, 1946)? In this context, recent research has started to focus on the importance of stimulus perception to better understand both inter and intraindividual response variations during stimulus generalization (Struyf et al., 2015; Zaman, Chalkia, et al., 2021). These authors found that during the assessment of stimulus generalization, participants misidentified many test stimuli as the CS+, with large interindividual differences in the overall extent and the distribution of misidentifications across the stimulus spectrum. Using cluster analyses, they found subgroups of *good* and *bad stimulus identifiers* with distinct generalization patterns between those subgroups (Zaman, Ceulemans, et al., 2019; Zaman, Struyf, et al., 2019; Zaman, Struyf, et al., 2021). Differences in stimulus identification contributed in multiple manners to interindividual response patterns during generalization testing. Conditioned responses were strong when test stimuli were misidentified as the CS+ but attenuated on trials when test stimuli were identified as different from the CS+ (even when the actual CS+ was presented). Furthermore, interindividual differences in stimulus identification led to different learning experiences; thereby, affecting the shape of the response gradient. Thus, by using cluster analysis, these studies were able to show qualitative differences in the pattern of participants' identification judgements and account for interindividual differences in response patterns during generalization testing.

Inspired by recent research in humans where response gradients in specific individuals deviated from a similarity-based response pattern, various authors have proposed the involvement of an inductive or cognitive process (Lee et al., 2018; Lovibond et al., 2020; Schlegelmilch et al., 2021; Wong & Lovibond, 2017). For instance, while some participants generalize along a green-blue stimulus dimension based on similarity of the test stimulus to the CS+, others reported adhering to a more abstract rule (i.e., the bluer the stimulus, the more I expect a shock). These findings are reminiscent of the distinction between strategies based on exemplar similarity versus abstract rules made in category learning (e.g., Erickson & Kruschke, 1998; Little & McDaniel, 2015; Nosofsky, 1986), nonlinear compound discriminations (Shanks & Darby, 1998), multiple cue judgment (e.g., Hoffmann et al., 2016), and function learning (e.g., DeLosh et al., 1997), suggesting that these individual differences in response rules are general across a variety of inductive tasks. Currently, the sources of these individual differences in inductive rules in the context of associative learning are unknown (but see Schlegelmilch et al., 2021).

Although the inductive and perceptual components of generalization have historically been separate, they may be related via the individual differences in each component. In support, evidence from related fields suggests a relationship between perception and higher-order cognition. Goldstone and Barsalou (1998) argue that perception and conception, traditionally considered to be opposed as "bottom-up" and "top-down" processes, respectively, are intimately linked. Their argument is based on three lines of evidence: (1) abstract concepts often have their origins in perception, (2) perception and conception share similar functional properties, and (3) perception can change as a result of categorization. For example, Goldstone (1994) showed that after performing a classification task, participants showed increased ability to discriminate between stimuli that were classified in different categories (acquired distinctiveness; Honey & Hall, 1989), demonstrating that categorization can affect perception.

Furthermore, there is evidence that individual differences in stimulus discriminability predict the type of cognitive strategy used to execute an experimental task. Rouder and Ratcliff (2004) demonstrated that the ease of stimulus discriminability determined which model best accounted for categorization data. When stimuli were difficult to discriminate, responses were best predicted by a rule-based model (e.g., Ashby & Maddox, 1993), while for clear and distinct stimuli, an exemplar account (e.g., Nosofsky, 1986) was favored. In the former, categorizations were based on a great simplified abstraction or rule of the current stimulus set, while for the latter, a detailed representation of the stimulus set was required. In line with these findings, differences in stimulus discriminability were recently proposed to predict the type of generalization rule in context of associative learning (Zaman, Chalkia, et al., 2021). To date, whether there is indeed a relationship between differences in stimulus identification and adopted generalization rules for simple cue-outcome associations remains to be tested.

The current study aimed to test the relationship between individual differences in stimulus identification, generalization gradients, and explicit generalization rules. We achieved this aim by reanalyzing six datasets obtained from predictive learning tasks, four of which were reported in Lovibond et al. (2020) and two of which were unpublished studies carried out by the same authors. The original study examined how individual differences in rules predicted differences in stimulus identification and expectancy gradients. However, it did not consider how individual differences in identification might predict generalization rules and response gradients. To

this end, we conducted a series of cluster analyses to identify distinct subgroups of individuals based on the stimulus identification data and tested how these clusters were related to expectancy gradients and inductive rules.

Method

Six different datasets were analyzed, four of which have been previously published in Lovibond et al. (2020) and were retrieved from Open Science Framework (OSF, https://osf.io/xnva4/). The same authors were willing to share two additional unpublished datasets that followed a similar procedure and were run at the same time as the experiments reported in Lovibond et al. (2020). We only report the main procedural details needed to interpret our analyses below. The full details of the protocol used to generate those datasets (1, 2, 5, and 6) can be found in Lovibond et al. (2020). Table 1 provides an overview of the different datasets and their main methodological differences. We adopted the same exclusion criteria as in the original paper, excluding participants if they reported being colorblind, failed to meet the learning criteria, or failed the instructions check (applicable to datasets 5 and 6 only; see Table 1). The experiment was approved by the University of New South Wales Human Research Ethics Advisory Panel C (file HREAP 2615).

Stimuli

The stimuli consisted of colored rectangles measuring 236 × 118 pixels within a 472 × 472 pixel white canvas (i.e., for a 21.5-in. screen with a 1920 × 1080 pixel resolution, the rectangles measured 5.85 × 2.93 cm). In total, 17 different stimuli were created (see Figure 1A) across the green-blue spectrum. The color changes between stimuli were equally spaced (in steps of .02 between stimuli with an integer number), ranging between hue values of .4 and .6, keeping saturation and brightness constant. Although Lovibond et al. (2020) did not ascertain whether the equal increments between stimuli held in psychological space (as opposed to physical space), the stimulus set closely resembles one of the stimulus sets reported by Douven et al. (2017) to be linear in psychological space. In Experiments 1–4, only the stimuli with an integer number were used (in total 11), whereas, in the remaining experiments, all 17 stimuli were used (i.e., there was finer resolution in the middle of the dimension). Color allocation to the stimulus dimension was counterbalanced (with S1 having a hue value of .4 or .6). S6 always served as the CS+ and was situated at the midpoint of the test dimension. In the differential group, S4 was the CS−. Depending on color counterbalancing, CS− was slightly greener or bluer than the CS+ (Figure 1A).

Procedure

The experiments listed in Table 1 were run online using Amazon's Mechanical Turk (MTurk) platform. All experiments consisted of an initial acquisition phase and then two test phases in counterbalanced order. This means that participants either completed the generalization test first and the identification test second, or vice versa (see Figure 1B). The experiments were identical except for the minor procedural details noted below.

Acquisition

The acquisition phase consisted of a predictive learning task. Participants learned to predict the occurrence of a hypothetical shock (unconditioned stimulus, US) based on symbols consisting of colored shapes (conditioned stimuli, CS). On each trial, a CS was presented along with a prompt asking participants to indicate whether they expected the US by pressing one of two keys ("L" for shock or "A" for no shock). Upon responding, feedback about the correct outcome (either "SHOCK!" or "no shock occurred") was displayed along with the CS for 5 s, after which a blank 2 s intertrial interval (ITI) began (see Figure 1B). During simple conditioning, the CS+ (an aqua rectangle, see Figure 1A) was presented on each of the 12 trials, with the CS+ reinforced at 75%. In addition to these 12 CS+ trials, 12 CS− (see Figure 1A) trials were included (that were never reinforced) in the differential conditioning group. Thus, this phase consisted of 12 or 24 trials for the simple and differential conditioning groups, respectively.

Test

Generalization Task. In the generalization test, participants were instructed to continue to make predictions but that they would no longer receive feedback. On each trial, participants were presented with the test stimulus along with a text prompt ("What is the likelihood of this symbol leading to a shock?"), and a visual analog scale (expVAS) underneath. Instead of making a keypress choice as in training, participants were asked to indicate their expectancy of the shock symbol on the expVAS using the mouse. The scale was labeled "Certain NO SHOCK (0% of shock)" on the left and "Certain SHOCK (100% of shock)" on the right. All

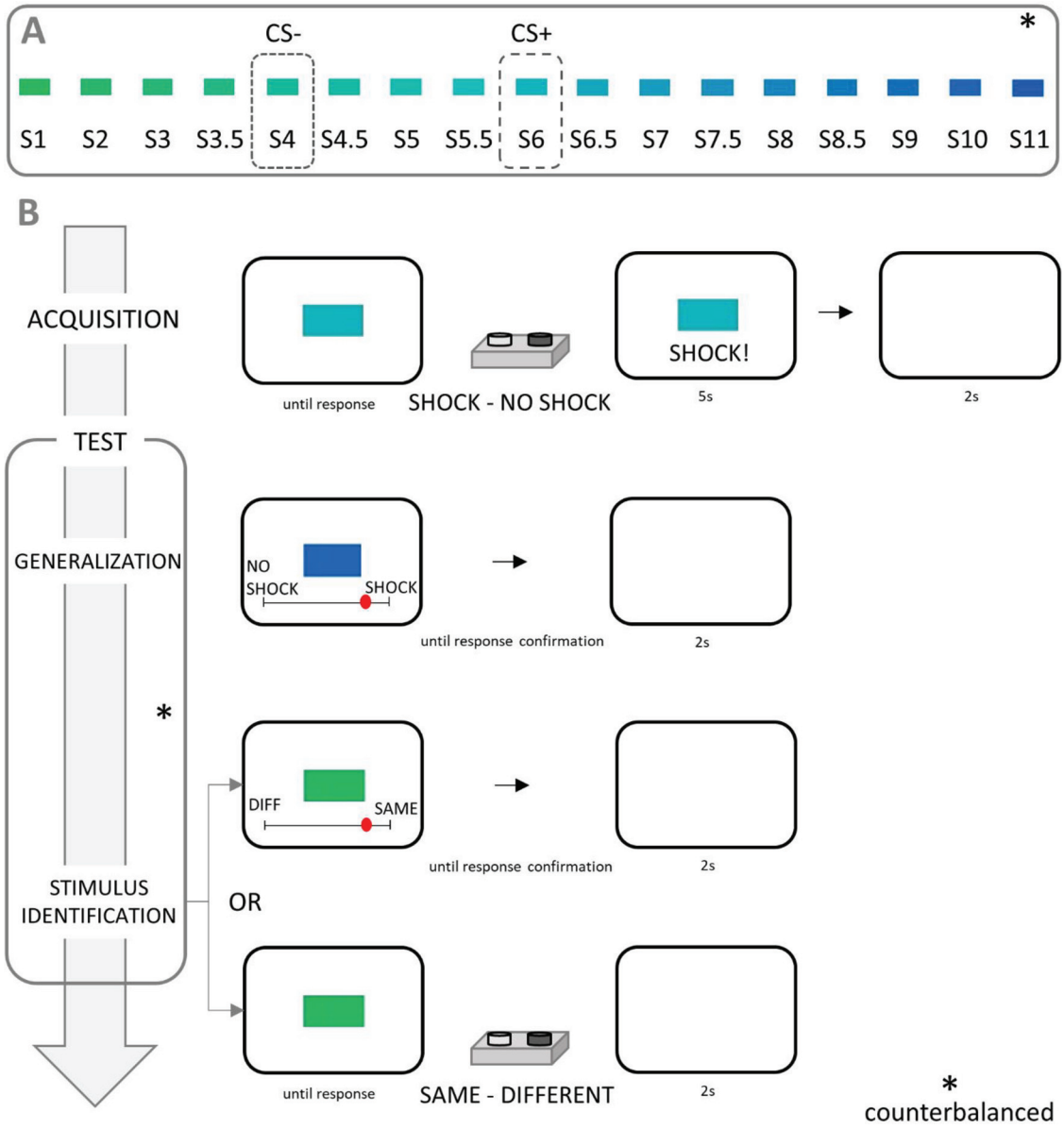
Table 1
Overview of the Different Datasets

Dataset	Experiment in Lovibond et al. (2020)	Conditioning	Number of test stimuli	Identification test	Initial *N*	Color blind	Nonlearners	Instr. check	Final sample
1	1	Simple	11	2AFC (CS+)	179	6	7	—	166
2		Differential	11	2AFC (CS+)	221	10	30	—	181
3	—	Simple	11	idVAS (CS+)	188	7	12	—	169
4		Differential	11	idVAS (CS+)	211	6	35	—	170
5	2	Differential	17	2AFC (CS+)	212	15	39	5	153
6		Differential	17	2AFC (CS−)	185	10	27	37	111

Note. The table does not include information about the generalization test as it was identical in all six datasets. 2AFC = two-alternative forced-choice task where they had to identify the stimulus as the CS+ (CS−) or as a different stimulus, idVAS = similarity visual analog scale used to rate the stimulus' similarity to the CS+, Instr. check = instructions check at the end of the experiment.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Figure 1
Overview of the Protocol



Note. (A) In Experiments 1–4, only the stimuli with an integer number were used (11 total), whereas, in the remaining experiments, all stimuli shown (17 total) were used. (B) Procedure of the online experiment and trial structure within each phase. During acquisition, participants had to learn to predict the occurrence of a hypothetical shock from a strange machine based on colored symbols (Lovibond et al., 2020). Note that the ordering of the generalization and identification tests and color allocation to the stimuli (whether the CS was slightly bluer or greener than the CS+) was counterbalanced. Depending on the dataset, either an alternative forced-choice (datasets 1, 2, 5, and 6) or a visual analog scale (datasets 3 and 4) was used during the stimulus identification task. CS = conditioned stimuli; S = stimulus; diff = different. See the online article for the color version of this figure.

11 (datasets 1–4) or 17 (datasets 5–6) stimuli, including the CSs, were presented once in randomized order. After participants confirmed each VAS rating by clicking the "Continue" button, a 2 s blank ITI started before the next trial. All expectancy ratings ranged between 0 and 100.

Stimulus Identification Task. Here, participants were asked to identify each test stimulus as either being the same or different to the exact CS+ (datasets 1–5) or CS- presented in training (dataset 6). In datasets 1, 2, 5, and 6, this was done through a two-alternative forced-choice (2AFC) task where participants were asked to press the "S" key for same or the "D" key for different. The only difference between datasets 5 and 6 was whether participants were asked to identify the test stimuli as being the same as the CS+ (dataset 5) or the CS− (dataset 6). In datasets 3 and 4, participants

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

were asked to make their identification rating using a visual analog scale (idVAS) ranging from "Definitely different" to "Definitely the same" (see Figure 1B). Similar to the generalization test, all 11 (datasets 1-4) or 17 (dataset 5-6) stimuli, including the CSs, were presented once in a different randomized order to the generalization test. Once participants responded either to the 2AFC task via a button press, or clicked on a button to confirm their idVAS rating, the blank 2 s ITI started, after which the next trial began (see Figure 1B). Like the generalization test, participants had unlimited time to respond, and there was no feedback. The order of stimulus presentation was randomized independently from the order in the generalization test. Identification responses were coded as 0 (different) or 1 (same) for the 2AFC task (datasets 1, 2, 5, and 6) or ranged between 0 and 100 if the idVAS was used (datasets 3 and 4).

Rule Assessment

Rule assessment was conducted at the end of the experiment for datasets 3 and 4. In the other datasets, rule assessment was conducted immediately after the generalization task (and potentially before the identification task, depending on the allocated test order). Participants were asked a 3AFC question: "*During Phase ONE (where there was feedback), which of the following options did you think was most true concerning the likelihood of a shock occurring?*" They selected from a similarity option ("The more similar the symbol to an aqua [green-blue] color, the higher the likelihood of shock"), and two linear options ("*The greener/bluer the symbol, the higher the likelihood of shock*"). In the simple conditioning group, both linear options were equally consistent with the training contingencies since participants only ever saw one CS+. Therefore, the selection of either option was classified as simply "linear." However, for those who received differential conditioning, one of the linear options was consistent (classified as linear) with their experienced training contingencies, while the other was inconsistent (classified as "linear incon"). The different rule options in the 3AFC question were based on previous research by the same authors that showed a close correspondence between self-reported generalization rules and the shape of the generalization gradients with the blue-green stimuli used here (Lee et al., 2018). In that study, the two major rule subgroups (linear and similarity) emerged based on extensive coding of a large number of open-ended responses. While the forced-choice format of this question means that some participants may end up guessing, this method was sufficiently sensitive to detect differences in generalization between linear and similarity subgroups in the original study (Lovibond et al., 2020), replicating Lee et al. (2018) who used a combination of open-ended and forced-choice rule assessment. It should be noted that the method of rule assessment did not allow Lovibond et al. (2020) to assess when participants formed rules (i.e., during training or test). It is possible that despite the no-feedback procedure of testing being adequate in preventing extinction (see Lee et al., 2022), participants may still learn about the relational features of the stimuli and continue to learn or update rules at test (see Livesey & McLaren, 2009).

Exclusion Criteria

To meet the learning criterion, participants had to predict the shock outcome correctly on at least two of the last four CS+ trials during acquisition. In addition, participants in the differential conditioning group had to predict no shock correctly on at least three of the last four CS− trials. In datasets 5 and 6, subjects were additionally asked to report the stimulus they were asked to identify in the Identification test at the end of the experiment. Failure to report the correct stimulus led to exclusion.

Analyses

All datasets and the full reproducible code can be found at the OSF repository (osf.io/g8bjy/). The analyses were not preregistered as they were reanalyses of a previously published paper.

The data of the test phases consisted of stimulus identification (binary responses or continuous ratings) and US expectancy ratings, assessed in separate blocks. We investigated the relationship between stimulus identification and US expectancy gradients using a similar technique to Zaman, Ceulemans, et al. (2019). This technique involves using cluster analysis to identify distinct subgroups based on the performance during stimulus identification and then examining how the subgroups differ in their generalization gradients. Lovibond et al. (2020) reported analyses of the identification gradients using hits, false alarms, and a symmetry index. In contrast, we decided to cluster individuals using the identification data across the test dimension. This method does not divide the identification data into three separate indices (calculated across subjects), but uses the individual identification data across the entire test dimension (including the CS+) at once to find subgroups. Another advantage is that information regarding interindividual differences located at different points on the stimulus spectrum is not lost (i.e., participants with identical false alarm rates stemming from misidentifications located at different stimuli).

All analyses were performed in R 4.1 (2020-06-06; R Core Team, 2013). To perform the cluster analysis, we used the k-means algorithm (from the stats package v3.6.2; R Core Team, 2013), which repeatedly divides the whole sample of participants based on their stimulus identification data into an incrementing number of clusters 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 its 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). The optimal number of clusters was based on the largest relative decrease in the sum of squared Euclidean distances, the automated scree test (Ceulemans & Kiers, 2006; Wilderjans et al., 2013), and the silhouette width (Rousseeuw, 1987). In all datasets, two subgroups emerged except for datasets 5 and 6, where a three-cluster solution was preferred.

Next, we compared US expectancy gradients between the identified subgroups using mixed models (using the lme4 package v1.1.23; Bates et al., 2015), which provide powerful and flexible approaches to analyze repeated measures data (Blackwell et al., 2006). We tested the following model in all datasets, with cluster (two or three levels), stimulus (11 or 17 levels), and their interaction

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

as categorical predictors.[1] US expectancy was the dependent variable. In all models, the random effect consisted of a subject-dependent intercept to account for repeated measures. Omnibus tests of main effects were obtained using anova (from the car package v3.8; Fox & Weisberg, 2019). We conducted post hoc contrasts using the models' estimated means with the emmeans package (v.1.5.0; Lenth, 2020). We reran these analyses for each task order separately (identification test first or generalization test first). This did not alter the overall findings (see online supplemental materials). Furthermore, we also ran an additional analysis with a stricter learning criterion that neither altered the found results (see online supplemental materials).

Finally, we tested whether the probability of reporting each rule was dependent upon stimulus identification using the identified clusters as a predictor. Depending on the number of levels of the dependent variable rule (two or three), we used either bi- or multinomial logistic regression using multinom (from the nnet package v.7.3.14; Venables & Ripley, 2002). Logistic regression expresses the logarithm of the odds (or logit-transformed probability) of a categorical dependent variable as a linear relationship with the predictor variables. In these models, the estimated beta for the intercept (β_i) is the log odds that expresses the probability that a certain outcome occurred relative to the probability of another outcome (that serves as the reference category). In all models, the linear rule served as the reference category. Beta estimates for the predictor cluster are log odds-ratios that convey to what extent the odds change if one belongs to Cluster 2 compared with Cluster 1. In other words, the estimated parameter for the cluster (β_c) indicates whether the odds for Cluster 2 are higher ($\beta > 0$), similar ($\beta = 0$), or lower ($\beta < 0$) than the odds for Cluster 1 (or vice versa depending on which subgroup is used as a baseline). Follow-up log odd tests per cluster are reported in the online supplemental materials and are in line with the reported cluster differences. Finally, we obtained Bayes factors (BF) and Cramer's *V* effect sizes for the contingency tables using the BayesFactor (Morey et al., 2018) and effectsize packages (Ben-Shachar et al., 2020) in R. Unless otherwise stated default parameter values were used, with a Poisson sampling distribution for the Bayes factors. As a rule of thumb, a BF of > 10 is considered strong evidence in favor of the alternative model compared with the null model. A BF of $< .1$ indicates strong evidence in favor of the null model. A BF between $.3 < BF < 3$ is considered weak or inconclusive evidence (Jeffreys, 1961).

To foreshadow our results, we report the analyses for the datasets grouped by conditioning protocol because many of the effects were found consistently in all datasets. First, we report the datasets that used simple conditioning (CS+ only). Next, we report the results from the differential conditioning (CS+ and CS−) datasets. The dataset comprising the CS− identification responses is reported separately at the end because participants were asked to identify the CS+ (not the CS−) in the identification test for all other datasets.

Results

Simple Conditioning - Datasets 1 and 3

Stimulus Identification

Despite differences in whether stimulus identification was assessed using binary (dataset 1) or continuous (dataset 3) measures, it is clear that the gradients for both *good* and *bad* clusters were very similar across datasets (Figure 2A and 2D). In dataset 1, 47 participants (28.31%) were allocated to a cluster representing poor performance on the stimulus identification task (i.e., bad *identifiers*). Here, response probabilities hovered around the chance level of .5 for the entire range of test stimuli with many mistakes for both the CS+ and all the test stimuli (Figure 2A). For the remaining 119 participants (71.69%), the allocated cluster indicated good identification accuracy (i.e., good *identifiers*). For these participants, response probabilities peaked at the location of the CS+ (i.e., correct responses) and declined steeply around the CS+ (Figure 2A). Response probabilities were around .5 for the adjacent test stimuli (S5 and S7), and much lower (i.e., fewer misidentifications) for more distant test stimuli (S1:S4 and S8:S11).

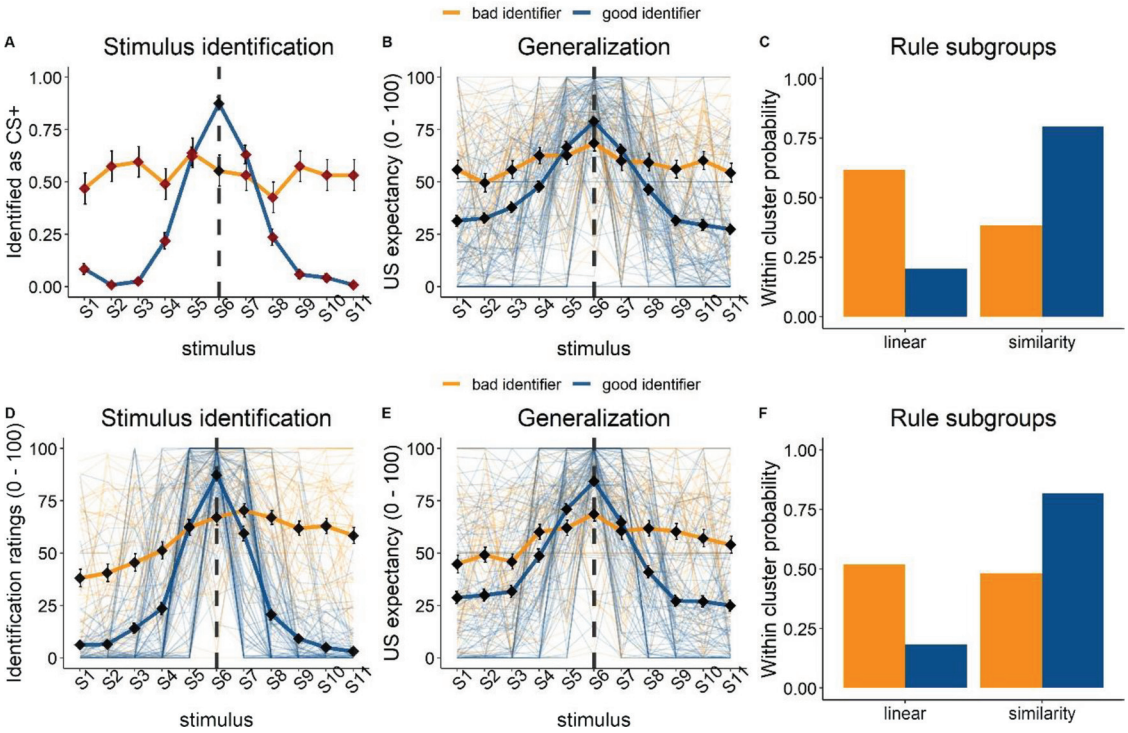
In dataset 3, 54 participants (31.95%) were allocated to a *bad identifiers* cluster with overall lower, flatter, and slightly asymmetrical identification ratings across the stimulus range (Figure 2D). The remaining 115 participants (69.05%, *good identifiers*) that were clustered together had average identification ratings that peaked at the CS+ and steeply decreased as test stimuli became less similar (Figure 2D).

US Expectancy

As visible in Figures 2B and 2E, there were large differences in US expectancy across the stimulus spectrum in both datasets (dataset 1: stimulus effect: $F(10, 1640) = 22.292$, $p < .001$, dataset 3: stimulus effect: $F(10, 1670) = 37.256$, $p < .001$) with distinct gradients between subgroups of participants that were identified as *good* compared with *bad identifiers*. *Bad identifiers* had a flatter US expectancy gradient that was overall located higher compared with the gradient of the *good identifiers*, which was symmetrical with a narrow peak at the CS+ location (dataset 1: cluster effect: $F(1, 164) = 50.559$, $p < .001$; Cluster × Stimulus effect: $F(10, 1640) = 9.154$, $p < .001$; dataset 3: cluster effect: $F(1, 167) = 32.077$, $p < .001$; Cluster × Stimulus effect: $F(10, 1670) = 13.984$, $p < .001$). Exploratory, we tested whether differences in US expectancy between subgroups paralleled differences in stimulus identification across the stimulus range (see Table 2). To perform these analyses, we divided the stimulus dimension into sections where the good identifiers showed higher identification probabilities than the bad identifiers (i.e., S6, see Figure 2A and 2D), sections where the good identifiers showed lower identification responses than the bad identifiers (i.e., S1:S4 and S8:S11), or no difference (i.e., S5 and S7) at the aggregate. We then compared US expectancy ratings between subgroups separately for each of these four sections. It is worth highlighting that our approach in this analysis was to use subgroup differences on stimulus identification to examine whether the same subgroup differences could be found on the expectancy measure. In both datasets, we found that (a) US expectancy ratings to the CS+ was higher in the *good* compared with *bad identifiers* subgroup; (b) did not differ between subgroups for the test stimuli most proximal to the CS+, and (c) was lower for more distant test stimuli on either side in the *good* compared with *bad identifiers* subgroup (see Table 2). These

[1] The use of stimulus as categorical rather than scale predictor in a second order Polynomial was preferred based on the Bayesian information criterion (BIC) with lower values indicating better model fit

Figure 2
Experiments Using a Simple Conditioning Protocol, Dataset 1 (Top) and 3 (Bottom)



Note. (A) Probability of being identified as CS+ across the stimulus range per cluster. Black and red markers indicate correct and incorrect responses, respectively. (B and E) Averaged US expectancy ratings per stimulus per cluster (thick lines) and individual gradients (thin lines). (C and F) The relative number of participants reporting each generalization rule per cluster. (D) Averaged identification ratings per cluster (thick lines) and individuals (thin lines). The dashed line indicates which stimulus was used as CS+. Error bars represent standard errors of the mean. US = unconditioned stimulus; S = stimulus. See the online article for the color version of this figure.

contrasts indicate that differences in US expectancy ratings between the different clusters closely parallel differences in identification performance for matched sections along the stimulus dimension.

Rules

We examined whether the proportion of participants deriving linear and similarity rules differed between clusters. Cluster allocation significantly predicted which response rule people selected for their US expectancy ratings at the end of the task (dataset 1, Figure 2C). *Bad identifiers* were not more likely to select either of the rules ($\beta i = -.48$, $SE = .30$, $z = -1.59$, $p = .112$) while *good identifiers* compared with the *bad identifiers* were more likely to select the similarity rule then the linear rule ($\beta c = 1.85$, $SE = .38$, $z = 4.91$, $p < .001$). The same pattern of results was observed in dataset 3 ($\beta i = -.07$, $SE = .27$, $z = -.27$, $p = .786$; $\beta c = 1.57$, $SE = .36$, $z = 4.322$, $p < .001$). In line with these results, we found large Bayes

Table 2
Custom Contrasts for US Expectancy Ratings

	Dataset 1				Dataset 3				Dataset 2			
	Bad vs. good				Bad vs. good				Bad vs. good			
Stimulus	*df*	*t*	*p*	P_c	*df*	*t*	*p*	P_c	*df*	*t*	*p*	P_c
S6 (CS+)	1,712	−2.16	.031*	.062	1,454	−3.28	.001***	.002**	1,814	−2.80	.005**	.010**
S5 and S7	1,205	−1.29	.199	.99	822	−1.81	.071	.071	1,204	−1.13	.260	.260
S1:S4	565	6.97	<.001***	<.001***	394	5.24	<.001***	<.001***	559	6.15	<.001***	<.001***
S8:S11	565	8.92	<.001***	<.001***	394	9.72	<.001***	<.001***	559	3.18	.002**	.006**

Note. US = unconditioned stimulus. Uncorrected and corrected *p* values using the Holm-Bonferroni correction are reported.
* $p < .05$. ** $p < .01$. *** $p < .001$.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

factors favoring a relationship between cluster and rule in both datasets (dataset 1: BF = 111125. Cramer's V = .39, 95% confidence interval, CI [.26, 1]; dataset 3: BF = 4879. Cramer's V = .33, 95% CI [.20, 1]).

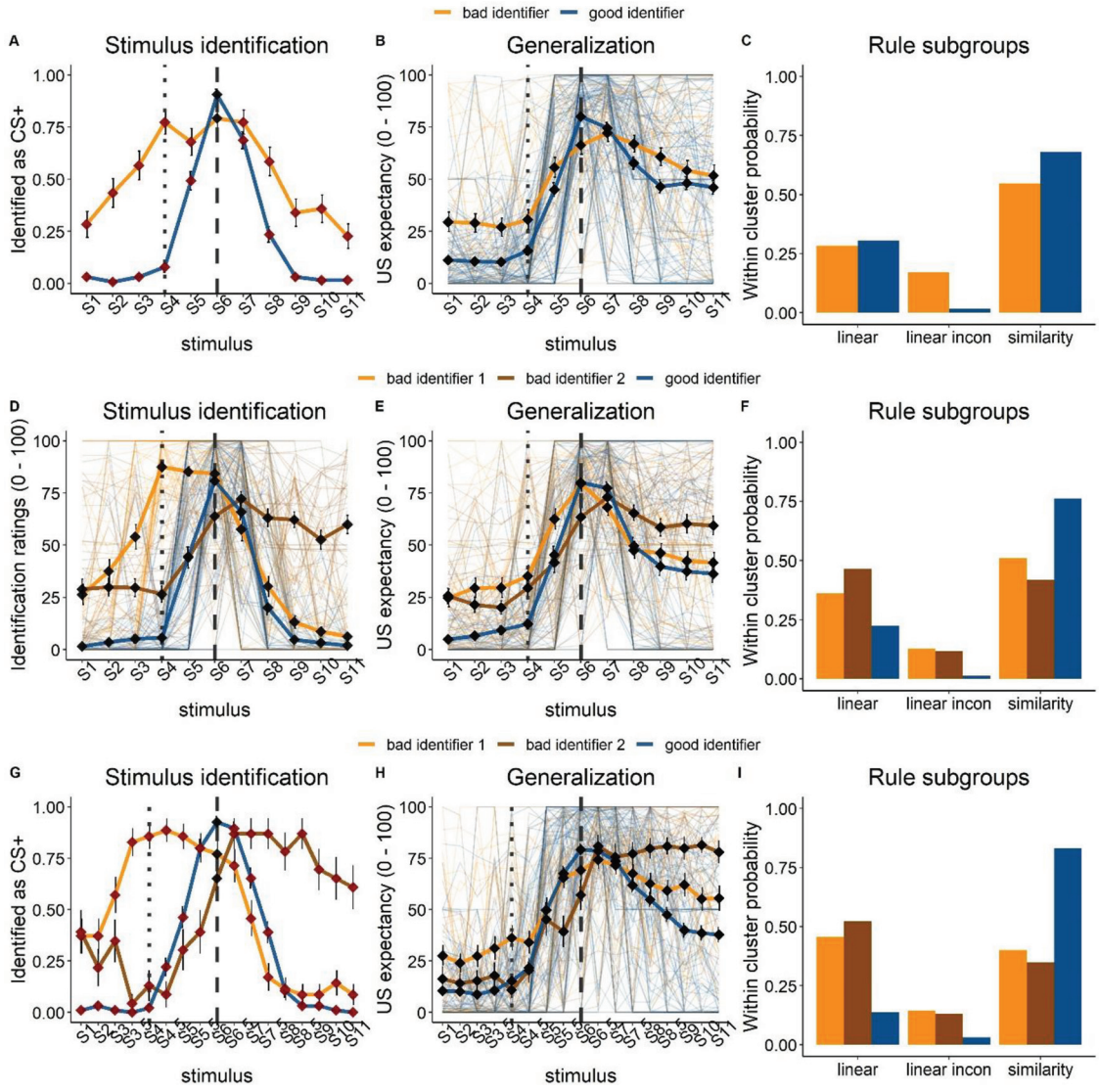
Differential Conditioning - Datasets 2, 4, and 5

Stimulus Identification

As visible in Figure 3, a two-cluster solution was preferred in dataset 2 (Figure 3A). In one cluster, response probabilities peaked at the CS+ (i.e., correct responses) and started to decline immediately as the physical proximity of the test stimulus to the CS+ decreased; 70.71% (N = 128) of the sample in dataset 2 were allocated to this cluster. The remaining 29.29% (N = 53, dataset 2) were allocated to the other cluster. This cluster gradient suggests poorer performance, although it seems to differ slightly from the *bad identifiers'* gradients observed in the datasets that used a simple conditioning protocol (see Figure 3A). Response probabilities reached a plateau from the CS− till the CS+, after which response probabilities started to decrease, albeit to a lesser extent compared with the *good identifiers*.

Figure 3

Experiments With a Differential Conditioning Protocol, Datasets Two (Top), Four (Middle), and Five (Bottom)



Note. (A and G) Probability of being identified as CS+ across the stimulus range per cluster. Black and red markers indicate correct and incorrect responses, respectively. (B, E, and H) Averaged US expectancy ratings per stimulus per cluster (thick lines) and individual gradients (thin lines). (C, F, and I) Relative rule frequency per cluster. (D) Averaged identification ratings per cluster (thick lines) and individuals (thin lines). The dotted and dashed lines indicate which stimulus was used as CS+ and as CS−. Error bars represent standard errors of the mean. US = unconditioned stimulus, S = stimulus. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

In dataset 4 and 5, we found that a three-cluster solution was preferred (Figure 3D and 3G). In the first subgroup of *bad identifiers* (dataset 4: $N = 47$, 27.65%; dataset 5: $N = 35$, 22.87), identification ratings (or response probabilities) reached a plateau from approximately the CS− till the CS+ and decreased as test stimuli approached the end of the spectrum. In the second subgroup of *bad identifiers* (dataset 4: $N = 43$, 25.29%; dataset 5: $N = 23$, 15.03%), identification ratings (or response probabilities) did not peak at the CS+ but at S7 and remain relatively high for the test stimuli on this side of the CS+ compared with test stimuli on the CS− side of the CS+. The final cluster was the *good identifiers* (dataset 4: $N = 80$, 47.06%; dataset 5: $N = 95$, 62.09%) with identification ratings (or response probabilities) peaking at the CS+ (i.e., correct responses) and declining immediately after that in both directions away of the CS+.

US Expectancy

We found differences in US expectancy ratings across the stimulus range with distinct gradients between clusters in all three datasets (see Table 3 for omnibus effect and Figures 3B, 3E, and 3H). Similar to the simple conditioning datasets, we performed custom contrasts and found that cluster differences in US expectancy ratings closely paralleled cluster differences in CS identification performance across regions of the stimulus dimension. In comparison with the bad identifiers, the good identifiers had higher US expectancy at the CS+, no difference in US expectancy ratings for the adjacent test stimuli, and lower ratings for most of the remaining stimuli on either side of the CS+ (see Table 2 for dataset 2). We found a similar result when comparing the clusters with the three-cluster solution datasets (4 and 5; see Table 4 for contrasts). Higher US expectancies were observed for stimuli where one of the clusters had a higher probability of (mis)identifying the stimulus as the CS+ (for binary judgements) or higher identification ratings (for continuous judgements).

Rules

In dataset 2, *bad identifiers* had similar probabilities of reporting the inconsistent linear rule compared with the linear rule ($\beta i = -.51$, $SE = .42$, $z = -1.21$, $p = .226$), while *good identifiers* compared with the bad identifiers were significantly less likely to report the inconsistent linear rule ($\beta c = -2.46$, $SE = .84$, $z = -2.93$, $p = .003$). The bad identifiers were more likely to report the similarity rule than the linear rule ($\beta i = .66$, $SE = .32$, $z = 2.07$, $p = .038$), as were the *good identifiers* with odds not differing from the odds in the *bad identifiers* subgroup ($\beta c = .14$, $SE = .37$, $z = .38$, $p = .70$; Figure 3C). In line with these results, we found a Bayes factor favoring a relationship between cluster and rule (BF = 77, Cramer's $V = .29$, 95% CI [.16, 1]. However, when the inconsistent linear rule was excluded, the results favored the null hypothesis (BF = .334, Cramer's $V = .02$, 95% CI [.00, 1]).

In dataset 4, *good identifiers* were significantly less likely to report the inconsistent linear rule compared with the linear rule ($\beta i = -2.88$, $SE = 1.03$, $z = -2.813$, $p = .005$) while both clusters of *bad identifiers compared with the former* did not had different odds ($\beta c_2 = 1.50$, $SE = 1.14$, $z = 1.32$, $p = .189$; $\beta c_3 = 1.84$, $SE = 1.13$, $z = 1.63$, $p = .102$). Furthermore, *good identifiers* were more likely to report the similarity rule than the linear rule ($\beta i = 1.22$, $SE = .27$, $z = 4.55$, $p < .001$), while both clusters of *bad identifiers* had lower odds of reporting the similarity rule compared with the *good identifiers* ($\beta c_2 = -1.33$, $SE = .42$, $z = -3.15$, $p = .002$; $\beta c_3 = -.88$, $SE = .42$, $z = -2.11$, $p = .035$). Again we found a large Bayes factor favoring a relationship between cluster and rule (BF = 389.67. Cramer's $V = .24$, 95% CI [.12, 1] that remained after exclusion of the inconsistent linear rules: BF = 41.72, Cramer's $V = .02$, 95% CI [.00, 1]).

A similar set of results was found in dataset 5. *Bad identifiers* had similar probabilities of reporting the inconsistent linear rule compared with the linear rule ($\beta i = -.18$, $SE = .30$, $z = -.60$, $p = .55$), while *good identifiers* compared with the former were significantly less likely to report the inconsistent linear rule ($\beta c = -1.54$, $SE = .57$, $z = -2.69$, $p = .007$; Figure 3I). Furthermore, *bad identifiers* were equally likely to report the similarity rule compared with the linear rule ($\beta i = .00$, $SE = .29$, $z = .01$, $p = .99$), while *good identifiers* were much more likely to report the similarity rule of the linear rule compared with *bad identifiers* ($\beta c = 1.16$, $SE = .367$, $z = 3.21$, $p = .001$). Similarly to dataset 4, in dataset 5 the Bayes Factor suggested a relationship between cluster and rule (BF = 272801, Cramer's $V = .33$, 95% CI [.22, 1]), that also remained after exclusion of the inconsistent linear rule (BF = 101622, Cramer's $V = .44$, 95% CI [.30, 1]).

Differential Conditioning - Dataset 6

Stimulus Identification (CS−)

In the final dataset, we identified two subgroups of participants—again best described as good and bad (CS−) identifiers (Figure 4A); 69.37% ($N = 77$) were classified into the *good identifiers* cluster, as their accuracy peaked at the CS− location and decreased across the spectrum. For the remaining 30.06% of the participants ($N = 34$) who were classified into the *bad identifiers* cluster, identification responses peaked around the CS+ instead of the CS and were much lower for the actual CS−.

Table 3
F-Tests of Omnibus Effects

	Dataset 2			Dataset 4			Dataset 5		
Fixed effect	*df*	*F*	*p*	*df*	*F*	*p*	*df*	*F*	*p*
Stimulus	10, 1,790	81.41	<.001***	10, 1,670	85.70	<.001***	16, 2,400	76.54	<.001***
Cluster	1, 179	18.82	<.001***	2, 167	14.49	<.001***	2, 150	12.93	<.001***
Cluster × Stimulus	10, 1,790	4.43	<.001***	10, 1,670	5.74	<.001***	32, 2,400	6.75	<.001***

Note. US = unconditioned stimulus. Uncorrected and corrected *p* values using the Holm-Bonferroni correction are reported.
*** $p < .001$.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 4
Custom Contrasts of Dataset 4 and 5 for US Expectancy

	S6 (CS+)				S1:S4				S8:S11			
Contrast	*df*	*t*	*p*	P_c	*df*	*t*	*p*	P_c	*df*	*t*	*p*	P_c
Dataset 4												
Bad 1 vs. good	1,656	0.10	.92	.92	492	7.05	<.001***	<.001***	492	1.19	.234	.468
Bad 2 vs. good	1,656	−3.00	.003**	.015*	492	5.08	<.001***	<.001***	492	6.36	<.001***	<.001***
Bad 1 vs. bad 2	1,656	−2.60	.009**	.036*	492	1.59	.113	.339	492	4.66	<.001***	<.001***
Dataset 5												
Bad 1 vs. good	2,147	−1.77	.08	.24	459	6.00	<.001***	<.001***	459	4.99	<.001***	<.001***
Bad 2 vs. good	2,147	−3.31	.001***	.005**	459	1.07	.283	.283	459	10.15	<.001***	<.001***
Bad 1 vs. bad 2	2,147	−1.56	.120	.240	459	−3.49	.005**	.020*	459	5.11	<.001***	<.001***

Note. US = unconditioned stimulus. Uncorrected and corrected *p* values using the Holm-Bonferroni correction are reported.
* $p < .05$. ** $p < .01$. *** $p < .001$.

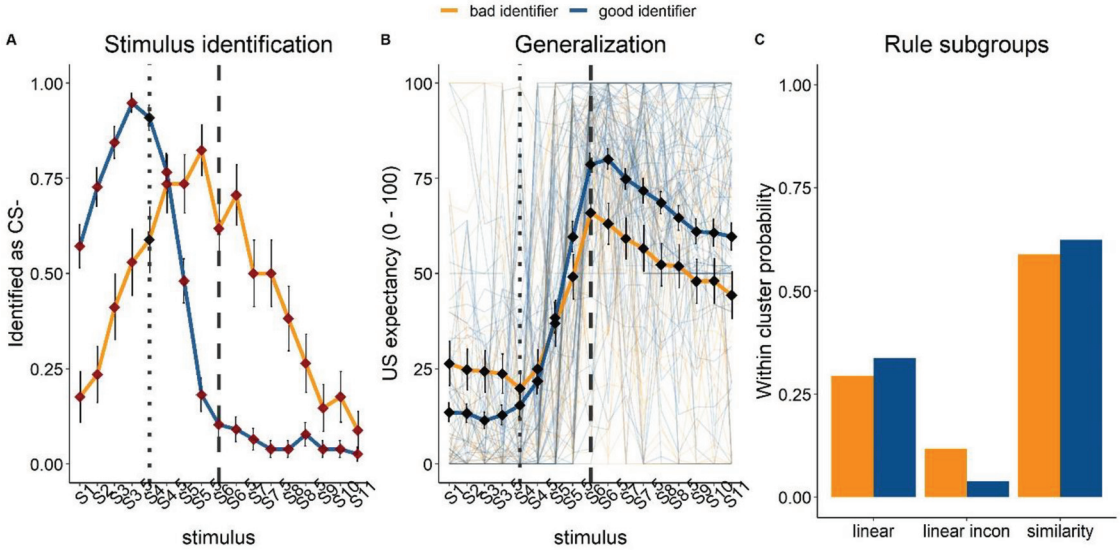
US Expectancy

US expectancy ratings differed according to stimulus (stimulus effect: $F(16, 1{,}744) = 54.73$, $p < .001$) and cluster (cluster effect: $F(1, 109) = 4.97$, $p = .028$; Cluster × Stimulus effect: $F(16, 1{,}744) = 4.27$, $p < .001$). Based on the point (S4.5) where identification probabilities crossed between both clusters, we created three custom contrasts to test for differences in US expectancy ratings between clusters. As visible in Figure 4B, US expectancy ratings were higher in the bad compared with the *good identifiers* for stimuli left of S4.5 ([S1:S4]: $t(363) = 3.36$, $p < .001$, $p_c = .002$), significantly lower for stimuli right of S4.5 ([S5:S11]: $t(156) = -5.17$, $p < .001$, $p_c < .001$) and did not differ at S4.5 ([S4.5]: $t(1{,}623) = .54$, $p = .59$, $p_c = .59$). Again, these contrasts indicate that differences in US expectancy ratings between the different clusters parallel cluster differences in identification performance.

Rules

Consistent with the other datasets, we found that *good identifiers* were more likely to report the similarity rule ($\beta i = .55$, $SE = .24$, $z = 2.362$, $p = .018$), and less likely to select the inconsistent linear rule ($\beta i = -2.23$, $SE = .61$, $z = -3.67$, $p < .001$), compared with the linear rule. The odds for the *bad identifiers* relative to the odds for the *good identifiers* did not differ for the similarity rule ($\beta c = -.01$, $SE = .44$, $z = -.03$, $p = .98$) while they were significantly higher for the inconsistent linear rule ($\beta c = 1.63$, $SE = .79$, $z = 2.06$, $p = .040$). The Bayes factor did not provide strong evidence for the null hypotheses nor a relationship between cluster and rule (BF = .436, Cramer's $V = .15$, 95% CI [0, 1]) nor after

Figure 4
Experiment With CS− Identification as Stimulus Identification Task



Note. Dataset 6: (A) Probability of being identified as CS− across the stimulus range per cluster. Black and red markers indicate correct incorrect responses, respectively. (B) Averaged US expectancy ratings per stimulus per cluster (thick lines) and individual gradients (thin lines). (C) Relative rule frequency per cluster. Error bars represent standard errors of the mean. US = unconditioned stimulus, S = stimulus. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

exclusion of inconsistent linear rule: BF = .413, Cramer's $V = 0$, 95% CI [0, 1]).

Discussion

In this study, we tested how individual differences in stimulus identification related to differences in patterns of stimulus generalization and reported generalization rules. Using cluster analysis in a set of six large datasets, we found highly consistent results despite differences in training procedure (simple or differential conditioning), identification format (binary or continuous judgments), and associative history of the to-be-identified stimulus (CS+ or CS−). The cluster analysis revealed distinct subgroups concerning CS identification in all datasets, representing participants who were good versus bad at identifying the CS+ (or CS−) from the other test stimuli. More important, we found a close relationship between CS identification performance and US expectancy gradients. Participants who showed more accurate CS identification were more likely to report the similarity rule than a linear rule, while participants who were less accurate at identifying the CS showed a reduction in this bias.

We observed distinct patterns of generalized responding between the different clusters in all datasets, closely paralleling their identification performance across the stimulus range. More precisely, based on differences in CS identification accuracy between the subgroups, we were able to predict not only points across the stimulus spectrum where US expectancy ratings differed but also the direction of the difference. For stimuli where one of the subgroups had a higher probability of (mis)identifying the stimulus as the CS+, higher generalized responding was observed. This led to higher US expectancy ratings for the *bad identifiers* at many of the test stimuli, while this led to higher US expectancy ratings at the CS+ for the *good identifiers*. Similarly, for stimuli where the subgroups did not differ in CS identification accuracies, they also did not differ on US expectancy. This was a robust relationship. We found the same close relationship when participants were asked to identify the CS− (with this time lower responding between subgroups when stimuli were more likely to be identified as CS−), and also when clustering was based on continuous CS identification ratings instead of binary judgments. Thus, although cluster analysis is sample- and task-dependent, the consistency of the results across the six datasets suggests that the reported relationships between individual differences in identification and generalization on a single dimension are somewhat reliable with the current task parameters.

The current results complement findings from the category learning literature suggesting a close relationship between perception and categorization (e.g., Carvalho & Goldstone, 2015; Goldstone, 1994; Goldstone & Barsalou, 1998). These studies showed that manipulating factors to influence perception changed what participants learned about categories of exemplars, and vice versa. In the current study, we found that individual differences in participants' ability to identify the stimuli from training predicted the type of explicit rule or strategy they reported for their generalization. The most consistent finding was that while the *good identifiers* showed a clear bias in reporting similarity rules over linear rules, this bias toward similarity rules was reduced in the *bad identifiers*, who most often had similar odds for both rules (see online supplemental materials for analyses per cluster). Thus, we have preliminary evidence that individual differences in stimulus identification are related to individual differences in inductive rules. It should be noted that our analysis approach cannot establish the direction of causation between these two variables. It could be that good stimulus identification increases the probability of reporting a rule where detailed awareness of stimulus features is important (i.e., a similarity rule). On the other hand, bad stimulus identification might foster a rule that relies on more noticeable or relational features (i.e., the difference in color between stimuli or classifying the stimuli as green or blue). Alternatively, it may be that what participants learn during training (stimulus features or relations, which ultimately determine the generalization rule) leads to better or worse encoding of the stimulus features. For instance, if participants notice the relational difference between the CS+ and CS− and form a relational response rule, they may stop attending to the exact physical features of the stimuli as that information is no longer relevant. If, however, participants solve the training task by attempting to memorize the exact color of the CS+, this should lead them to report a similarity generalization rule and also show good performance on the subsequent identification test. Establishing the direction of effect between these components of generalization will be an exciting avenue for future research. Additional studies are also needed to investigate to what extent our observations apply to different stimulus sets as well as more complex stimulus dimensions (e.g., real-world or multidimensional stimuli). We note, however, that one advantage of using simple stimuli varying on a single dimension is that it restricts the number of different rules that participants derive (Lee et al., 2018). Thus, while we cannot be sure of the generality of our results for other dimensions and stimuli, the choice of datasets (and stimuli) for this initial study was appropriate given our aims.

Given the assessment of generalization (CR) and stimulus identification in separate test phases in the current experiment, our findings demonstrate the generality of previous research that found a close relationship between identification and CR when assessed on the same trial. Together with previous work (Struyf et al., 2017; Zaman, Struyf, et al., 2019; Zaman, Ceulemans, et al., 2019; Zaman, Chalkia, et al., 2021), the current findings provide compelling evidence that differences in stimulus identification are important to consider in explaining generalization. Considering the role of stimulus identification in generalization models would enable them to better account for interindividual differences in generalization patterns. For instance, different response gradients between individuals may emerge despite an identical underlying similarity-based generalization process due to differences in stimulus misidentification. Individual (and group) parameter estimates reflecting the extent of such generalization process (i.e., *SD* of a Gaussian) are at risk of being incorrect when stimulus identifications are not incorporated within the model. A subject with many misidentifications on test trials (leading to enhanced responding) will have a flatter response gradient, resulting in overestimating the degree of generalization. Furthermore, including differences in stimulus perception within rule-based models may allow these models to better account for interindividual differences and explain why different subjects develop different response rules despite being exposed to the same task.

Some limitations inherent to the current datasets and analyses should be acknowledged. Cluster analysis is a data-driven method, meaning that its output (i.e., number of identified clusters and

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

within-cluster data patterns) is entirely dependent on the input. However, the assumption is that despite potential minor differences between samples, distinct data patterns (stemming from latent subgroups) within datasets should be identifiable from these analyses. In our analyses, the various procedural differences (e.g., response format in identification test, number of test stimuli) across datasets meant that the input to the cluster analyses differed both in terms of the number of stimuli and their possible values (binary or continuous). These differences potentially contributed to the minor differences in results between datasets. However, due to these methodological differences, a direct comparison between clusters from the different datasets should be made cautiously. In other words, the "good" and "bad" identifiers are potentially not equivalent between datasets (indeed, there are differences in the absolute identification performance between datasets). Nevertheless, it is the differences *between* clusters within a dataset that we consider of primary interest, and these cluster differences were highly consistent across datasets. Thus, we believe that clustering is a useful method to characterize reliable individual differences in identification performance.

Our use of cluster analysis in this study was to identify individual differences in identification performance. As such, we cannot say what caused these subgroups to differ. It remains unclear what mechanism(s) are responsible for differences in CS identification as performance in the identification task involves multiple latent processes, including memory (i.e., encoding of CS features during training and retrieval during the test), stimulus perception, and attention. Future research should attempt to identify their relative contribution to identification performance and whether the differences observed are due to inherent differences in ability or incidental attention to different features in training.

We also wish to emphasize that due to the exploratory nature of cluster analysis, the interpretation of cluster output relied on post hoc inspection of the identification gradients in each cluster. Thus, the labels "good" and "bad identifiers" were derived from inspection of the width of the identification gradients and were intended to capture the key difference between clusters. However, labeling the clusters might not be as straightforward with increasing clusters numbers and more complex identification patterns. In this case, deriving an objective measure of identification performance would be useful. For example, performance could be defined as the degree to which performance deviates from optimal identification performance. An additional advantage of such a measure would be to allow for comparison between datasets, which is difficult with the current datasets (the few observations per participant precluded such approach).

Differences in identification performance may also reflect other factors of less theoretical interest. For example, since the data were collected online, it is possible that subgroups merely differed due to differences in the quality of the participants' display, affecting the ability to discriminate between the different hues. However, it is worth noting that other studies have identified similar (and other) subgroups in controlled experimental settings (Zaman, Struyf, et al., 2019; Zaman, Ceulemans, et al., 2019; Zaman, Struyf, et al., 2021). Another possibility is that the *bad identifiers* were inattentive or unmotivated responders who did not follow instructions and/or responded randomly throughout. Presumably, this type of response style would exhibit fast reaction times, as participants aim to progress through the task as quickly as possible. Although plausible, the results of additional analyses do not support this hypothesis (see online supplemental materials for these analyses): (a) overall response times were not faster in the *bad* compared with *good identifiers*, (b) US predictions were above chance level during the acquisition phase for most of the *bad identifiers*, and (c) additional reanalyses, with a stricter learning criterion did not change the reported pattern of results. Further, the close relationship between data patterns across the different phases (i.e., the shape of the CS identification or similarity gradient and the US expectancy gradient) would not be expected if the bad identifiers were truly responding randomly across the experiment. Similarly, in the differential conditioning datasets, the sigmoidal-shaped US expectancy gradient within the *bad identifiers* is difficult to reconcile with random responding. (d) Finally, it should be noted that the relatively flat US expectancy gradient in this group does not seem to emerge from an overall random-response style but from a mixture of heterogenous response patterns (e.g., some participants show linear gradients in either direction).

In summary, our study highlights the importance of considering individual differences in stimulus identification in understanding human generalization. We found that participants can be broadly classified into either good or bad stimulus identifiers (based on their identification pattern) that were predictive of the types of inductive rules that participants reported and the shape of the expectancy gradients. These relationships were robust in that they were found consistently among the six datasets varying in training contingencies and identification test format, remained using a stricter learning criterion than the original study (Lovibond et al., 2020; see online supplemental materials), and was not dependent on test presentation order (see online supplemental materials). Future research is needed to demonstrate the generalizability of our findings using different samples, stimulus sets, and task parameters and establish the direction of causation between identification and rules. Nevertheless, our results suggest that greater consideration should be given to interindividual variability in stimulus identification and induction in explaining response generalization.

References

Ashby, F. G. G., & Maddox, W. T. T. (1993). Relations between Prototype, Exemplar, and Decision Bound Models of Categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400. https://doi.org/10.1006/jmps.1993.1023

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.

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), 2815. https://doi.org/10.21105/joss.02815

Blackwell, E., de Leon, C. F. M., & Miller, G. E. (2006). Applying mixed regression models to the analysis of repeated-measures data in psychosomatic medicine. *Psychosomatic Medicine*, *68*(6), 870–878. https://doi.org/10.1097/01.psy.0000239144.91689.ca

Blough, D. S. (1975). Steady sate 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

Carvalho, P. F., & Goldstone, R. L. (2015). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? *Frontiers in Psychology*, *6*, 505. https://doi.org/10.3389/fpsyg.2015.00505

Ceulemans, E., & Kiers, H. A. L. (2006). Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical & Statistical Psychology*, *59*(Pt. 1), 133–150. https://doi.org/10.1348/000711005X64817

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968–986. https://doi.org/10.1037/0278-7393.23.4.968

Douven, I., Wenmackers, S., Jraissati, 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., & Paz, R. (2015). Fear generalization and anxiety: Behavioral and neural mechanisms. *Biological Psychiatry*, *78*(5), 336–343. https://doi.org/10.1016/j.biopsych.2015.04.010

Ennis, D. M. (1988). Confusable and discriminable stimuli: Comment on Nosofsky (1986) and Shepard (1986). *Journal of Experimental Psychology: General*, *117*(4), 408–411. https://doi.org/10.1037/0096-3445.117.4.408

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

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140. https://doi.org/10.1037/0096-3445.127.2.107

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 of London. Series B, Biological Sciences*, *362*(1479), 449–454. https://doi.org/10.1098/rstb.2006.1972

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200. https://doi.org/10.1037/0096-3445.123.2.178

Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, *65*(2–3), 231–262. https://doi.org/10.1016/S0010-0277(97)00047-4

Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology*, *51*(1), 79–88. https://doi.org/10.1037/h0046219

Hoffmann, J. A.,von Helversen, B., & Rieskamp, J. (2016). Similar task features shape judgment and categorization processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(8), 1193–1217. https://doi.org/10.1037/xlm0000241

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

Honey, R. C., & Hall, G. (1989). Acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology. Animal Behavior Processes*, *15*(4), 338–346.

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

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory*. Appleton-Century-Crofts.

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

Kalish, H. I. (1958). The relationship between discriminability and generalization: A re-evaluation. *Journal of Experimental Psychology*, *55*(6), 637–644. https://doi.org/10.1037/h0048049

Kindt, M. (2014). A behavioural neuroscience perspective on the aetiology and treatment of anxiety disorders. *Behaviour Research and Therapy*, *62*, 24–36. https://doi.org/10.1016/j.brat.2014.08.012

Lashley, K. S., & Wade, M. (1946). The Pavlovian theory of generalization. *Psychological Review*, *53*(2), 72–87. https://doi.org/10.1037/h0059999

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, 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., Le Pelley, M. E., & Lovibond, P. F. (2022). Nonreactive testing: Evaluating the effect of withholding feedback in predictive learning. *Journal of Experimental Psychology: Animal Learning and Cognition*, *48*(1), 17–28. https://doi.org/10.1037/xan0000311

Lenth, R. (2020). *emmeans: Estimated marginal means, aka least-squares means* (R package version 1.5.0) [Computer software]. https://cran.r-project.org/package=emmeans

Lissek, S., & Grillon, C. (2010). Overgeneralization of conditioned fear in the anxiety disorders. *The Journal of Psychology*, *218*(2), 146–148. https://doi.org/10.1027/0044-3409/a000022

Little, J. L., & McDaniel, M. A. (2015). Individual differences in category learning: Memorization versus rule abstraction. *Memory & Cognition*, *43*(2), 283–297. https://doi.org/10.3758/s13421-014-0475-1

Livesey, E. J., & McLaren, I. P. L. (2009). Discrimination and generalization along a simple dimension: Peak shift and rule-governed responding. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(4), 554–565. https://doi.org/10.1037/a0015524

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

Mackintosh, N. J. (1974). The psychology of animal learning. In *Generalization* (pp. 484–539). Academic Press.

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

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). BayesFactor: Computation of Bayes factors for common designs (R package version 0.9.12-4.2) [Computer software]. https://cran.r-project.org/package=BayesFactor

Mostofsky, D. I. (1965). *Stimulus generalization*. Stanford University Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Prokasy, W. F., & Ball, J. F. (1963). Primary stimulus generalization. *Psychological Review*, *70*(4), 310–322. https://doi.org/10.1037/h0049354

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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Rouder, J. N., & Ratcliff, R. (2004). Comparing categorization models. *Journal of Experimental Psychology: General*, *133*(1), 63–82. https://doi.org/10.1037/0096-3445.133.1.63

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. (2021). A cognitive category-learning model of rule abstraction, attention learning, and contextual modulation. *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000321

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*(4), 405–415. https://doi.org/10.1037/0097-7403.24.4.405

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

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer. http://www.stats.ox.ac.uk/pub/MASS4

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

Wilderjans, T. F., Ceulemans, E., & Meers, K. (2013). CHull: A generic convex-hull-based model selection method. *Behavior Research Methods*, *45*(1), 1–15. https://doi.org/10.3758/s13428-012-0238-5

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

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*, 1–19. https://doi.org/10.3758/s13423-020-01780-1

Zaman, J., Struyf, D., Ceulemans, E., Vervliet, B., & Beckers, T. (2021). 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

Received March 5, 2021
Revision received April 5, 2022
Accepted April 7, 2022 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.