Designing probabilistic category learning experiments: The probabilistic prototype distortion task

Nicolás Marchant (nicolasmarchant@alumnos.uai.cl)
Center for Social and Cognitive Neuroscience, School of Psychology, Universidad Adolfo Ibáñez.
Av. Presidente Errázuriz, 3328, Las Condes, Santiago de Chile.

Sergio E. Chaigneau (sergio.chaigneau@uai.cl)
Center for Social and Cognitive Neuroscience, School of Psychology, Universidad Adolfo Ibáñez.
Av. Presidente Errázuriz, 3328, Las Condes, Santiago de Chile.

Abstract

Many category learning experiments use supervised learning (i.e., trial-by-trial feedback). Most of those procedures use deterministic feedback, teaching participants to classify exemplars into consistent categories (i.e., the stimulus *i* is always classified in category *k*). Though some researchers suggest that natural learning conditions are more likely to be inconsistent, the literature using probabilistic feedback in category learning experiments is sparse. Our analysis of the literature suggests that part of the reason for this sparsity is a relative lack of flexibility of current paradigms and procedures for designing probabilistic feedback experiments. The work we report here offers a novel paradigm (the Probabilistic Prototype Distortion task) which allows researchers greater flexibility when creating experiments with different p(category|feature) probabilities, and also allows parametrically manipulating the amount of randomness in an experimental task. In the current work, we offer a detailed procedure, implementation, experimental results and discussion of this novel procedure. Our results suggest that by designing experiments with our procedures, the experimental setup allows subjects to achieve the desired classification performance.

Keywords: Category learning; Probabilistic feedback; Perceptual categorization; Experimental designs

Introduction

In the current work we present a method to develop category learning experiments with probabilistic feedback, allowing a researcher to flexibly stipulate category to feature association strengths, and at the same time easily controlling category to whole exemplar (i.e., feature combination) association. In the reported experiment, we use our method to train subjects, and show that participants learn as predicted, suggesting that the method is apt to be used on almost any category learning task with probabilistic feedback.

In category learning, a common procedure is to create stimuli by combining binary valued features. In general, with *n* features, it is possible to create 2^n stimuli spanning all possible feature combinations (see, Ashby & Valentin, 2018). Each stimulus or exemplar is thus a particular combination of features in one of their two possible states (Posner & Keele, 1968; Reed, 1972). In category learning experiments, participants are typically trained with all possible stimuli (but sometimes only with a subset; Medin & Schaffer, 1978). They may learn that certain combinations are members of category A and others are not (A not-A task), or that certain combinations are members of category A and others of category B (A or B task). After training, subjects are tested for classification performance (though other dependent variables are also possible). In most cases, subjects are provided with corrective feedback during training.

The most common type of feedback in category learning experiments is deterministic feedback (DF; Ashby & Ell, 2001; Nosofsky, Palmeri, & McKinley, 1994). In DF, each feature combination is always member of one of the categories. Consequently, subjects' performance increases during training towards some asymptotic performance level. Probabilistic feedback (PF) has been used much less frequently (Ashby & Gott, 1988; Gluck & Bower, 1988; Knowlton, Squire & Gluck, 1994; Little & Lewandowsky, 2009; Meeter, Radics, Gluck, & Hopkins, 2008). In PF, each feature combination belongs to a category with a probability less than 1.0.

The justification for the current work is twofold. First, it is likely that natural learning conditions integrate feedback from different sources, hence PF may be more representative of natural learning environments (Little & Lewandowsky, 2009; Meeter, et al., 2008; Lagnado, Newell, Kahan, & Shanks, 2006). DF assumes a perfect teacher, and natural conditions are probably more like PF, where the teaching signal may be inconsistent (e.g., an unprecise teacher). Second, at least some empirical findings in category learning may be conditional on DF. For example, inter-feature correlations have for a long time been considered an important part of conceptual representations (Hoffman & Rehder, 2010; Ell, Smith, Peralta & Hélie, 2017). It is generally accepted that subjects in category learning experiments tend not to learn inter-feature correlations. For inter-feature correlations to be learned, inference tasks have to be used (Chin-Parker & Ross, 2002; Yamauchi, Love & Markman, 2002). However, when PF has been used with classification procedures, evidence has been found that subjects do learn inter-feature correlations (Little & Lewandowsky, 2009). Consequently, it is possible that categorization phenomena being uncovered by using DF fail to generalize under PF conditions. Given these concerns, PF should be used more broadly.

We suspect that the relative lack of research in category learning that uses PF is due to difficulties that researchers might experience if attempting to design PF experiments. By looking at the Little and Lewandowsky (2009) experiments, it is apparent that they took their DF condition and turned it into a PF condition by simply changing the probabilities of each exemplar or feature combination. In contrast to using probabilities 1 and 0, in their PF conditions a feature combination belonging to category A with p(A) = .75, belonged to category B with p(B) = .25. The main problem with the Little and Lewandowsky procedures is that they allow little flexibility when designing experiments. Basically, probabilities are assigned to complete feature combinations, and individual feature probabilities are only derived. For experimental control, one would want a procedure that can flexibly combine feature probability choices with whole stimulus probabilities.

A procedure that could be used for giving PF is the Weather Prediction Task (WP; Knowlton, Squire, & Gluck, 1994; Gluck, Shohamy, & Myers, 2002). In the WP task, subjects are presented with combinations of playing cards similar to those of the Wisconsin Card Sorting Test (Monchi, Petrides, Petre, Worsley, & Dagher, 2001), and they have to learn to use them to predict the weather (i.e., rain or sun). During training, subjects are presented with combinations of cards (i.e., 1, 2, 3 or 4 cards). Problematically, to compute the probability of the outcome (i.e., rain) given each individual card, a very specific set of card combinations and of outcome probabilities have to be chosen. As in the Little and Lewandowsky procedures, here too there is little flexibility for researchers to select different cue probabilities. That many researchers continue to use the same probabilities used in the original Knowlton, Squire, and Gluck (1994) paper, attests to the paradigm's lack of flexibility in allowing the generation of stimuli with different probability patterns (Gluck, et al., 2002; Meeter, et al., 2008; Meeter, Myers, Shohamy, Hopkins, & Gluck, 2006).

The issue of how to design experiments with PF is what motivates the current work. In what follows, we present a novel way of designing PF category learning experiments, which can be flexibly used to determine p(category|feature), and to easily compute the probability of feature combinations (i.e., p(category|feature combination)). The method also allows introducing biases for a given category, and to create stimuli that gradually move from completely random to fully deterministic. Though we don't explore this issue in the current work, our novel procedures will allow researchers to study inter-feature relations if they wish to introduce them in their task. In the experiment we report, we show that the method allows participants to learn feature weights that are consistent with the relative informativeness of the features predicting category outcome. For reasons that will become clear shortly, we call this paradigm the *Probabilistic Prototype Distortion* task (PPD).

Method

Participants

Thirty-six undergraduate students (27 females) aged 18 to 37 (mean = 20.11, *SD* = 3.21), signed informed consent to participate in the experiment for course credit. Participants were randomly assigned to one of the three experimental conditions (AB, BC and CA, twelve participants in each condition). The experiment lasted approximately 30 minutes.

Design

We set up a 3 (condition: AB, BC and CA) x 3 (feature coefficient: A, B and C) mixed design experiment, with the last being a within-subjects factor. The *Probabilistic Prototype Distortion* task (PPD) is a mixture of classical prototype distortion tasks (Posner & Keele, 1968; Casale & Ashby, 2008) with probabilistic feedback. As discussed in the introductory section, there are some probabilistic classification procedures (e.g., Knowlton, Mangels, & Squire 1996; Gluck, Shohamy, & Myers, 2002; Little & Lewandowsky, 2009; Kruschke & Johansen, 1999), but they are hard to implement flexibly. To the best of our knowledge, there is no current procedure that allows flexibly combining feature probabilities (i.e., p(category|feature)) with overall feature combination probability (i.e., p(category|feature combination)).

The PPD assumes an "A-B" prototype procedure, though other procedures like "A-not A" or Inference are also possible. As any other prototype distortion task, two prototypical categories are created: A and B. As it will be clear next, our stimuli are composed of three features that have binary values (1 or -1). Category A prototype is defined by the "111" combination and category B prototype is defined by the "-1-1-1" feature combination. Using effect coding (see Table 1), we combined every possible feature-state to obtain a total of 8 exemplars ($2^3 = 8$).

Table 1: Effect coding for every binary valued feature (f_1, f_2 and f_3) and every exemplar combination.

Exemplar	*f*1	*f*2	*f*3
E1	1	1	1
E2	1	1	-1
E3	1	-1	1
E4	1	-1	-1
E5	-1	1	1
E6	-1	1	-1
E7	-1	-1	1
E8	-1	-1	-1

What is novel in our procedure is the use of the logistic regression equation to assign probabilities to exemplars. Using logistic regression, where each feature has an individual feature-weight, we can define an overall exemplar probability for each feature combination where p(category|feature combination) can be computed as shown in eq. (1).

$$p = {}^{1}\!/_{1 + e^{-c(\beta_0 + \beta_1 f_1 + \beta_2 f_2 + \beta_3 f_3)}} \qquad (1)$$

where, c is a sensibility parameter, B_0 is the constant on the equation and B_1, B_2, and B_3 are feature-weights. We created three different conditions (AB, BC and CA) that differentially weighted features in each specific feature combination. For example, in condition AB, feature A (f_1) is the most diagnostic feature (weight = 2.5), whereas feature B (f_2) contributes less to the classification probability (weight = 1), and finally feature C (f_3) has no diagnosticity at all (weight = 0). This same logic is applied to the other two conditions (BC and CA). Feature-weights for each condition are shown in Table 2. By using eq. (1) and the experimenter-specified feature weights, probabilities for each exemplar (i.e., feature combination) can be easily computed (see Table 3).

Table 2: Individual feature-weights (B_i) for each condition.

feature	Condition AB	Condition BC	Condition CA
(A)f1	2.5	0	1
(B)f2	1	2.5	0
(C)f3	0	1	2.5

Table 3: Classification probability to category A for each condition by every exemplar using eq. (1). Classification probability to category B is 1 - p(A). These probabilities can be used to guide how probabilistic feedback is provided (see text for details).

Exemplar	Condition AB p(A)	Condition BC p(A)	Condition CA p(A)
E1	0.9	0.9	0.9
E2	0.9	0.7	0.3
E3	0.7	0.3	0.9
E4	0.7	0.1	0.3
E5	0.3	0.9	0.7
E6	0.3	0.7	0.1
E7	0.1	0.3	0.7
E8	0.1	0.1	0.1

Finally, we used Luce's axiom (1962) to calculate individual feature diagnosticity (see eq. (2)). Accordingly, a feature-weight equal to 2.5 has a diagnosticity of .77, a feature-weight equal to 1 has a diagnosticity of .59 and feature-weight equal to 0 has a diagnosticity of .5.

$$p(A|f) = \frac{L_A(f)}{[L_A(f) + L_B(f)]} \qquad (2)$$

Importantly, exemplar probabilities in Table 3 specify how feedback is to be provided. A 0.9 probability in Table 3 tells the researcher that, e.g., exemplar E1 will be a member of category A on 90% of the trials and of category B in the remaining 10% of trials. This means that even if subjects have learned that exemplar E1 is most probably a member of category A, they will continue receiving corrective feedback on 10% of those trials in which exemplar E1 has to be classified. Additionally, this procedure allows that each individual feature is experienced in association with each category as specified in Table 2 and the corresponding feature diagnosticities. To set up any similar experiment, all the researcher needs to do is to specify the feature weights.

Because the probabilistic nature of the task, perfect performance is impossible. Optimal performance is near 77%. This means that optimal classification performance would be possible if subjects only use the feature weighted equal to 2.5 and completely ignore the other two. As will become clear in what follows, our data analysis methods allowed us to test whether subjects in our experiment resorted to such a rule or learned the specified feature weights instead.

Materials and procedures

We created two prototypical *ceremonial symbols*, similar to those used in Hoffman and Rehder (2010) and in Rehder, Colner, and Hoffman (2009). Ceremonial symbols are composed of three circles (each covering 9.62 cm^2 on the screen). Each of them encloses one of the two possible symbols (see Fig. 1). Each category prototype had a particular symbol combination. Circles remained in the same location throughout the task, and the binary-valued symbols were exclusive to each circle. The experiment was built using PsychoPy v3 and mounted online through the Pavlovia environment (Peirce, Gray, Simpson, MacAskill, Höchenberger, Sogo, Kastman, & Lindelov, 2019).

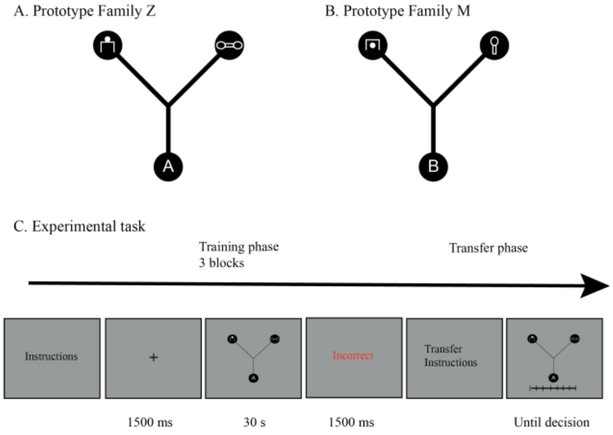

Figure 1. Complete experiment set-up. (A) Prototype ceremonial symbol for Family Z. (B) Prototype ceremonial symbol for Family M. Note that other exemplars are created by distorting these two prototypes. (C) Experimental procedure, which is composed of two phases: Training (240 trials) and Transfer (8 trials).

As with any prototype distortion task, the PPD involves two phases: Training and Transfer (Fig. 1C). During training, people had to classify if the presented exemplar belonged into one of the two possible categories: Family Z (Fig. 1A) and Family M (Fig. 1B). People had to press the keyboard letter Z if they believed that the exemplar belonged to Family Z or press letter M if they believed it belonged to Family M. During training, trial-by-trial feedback was provided. For correct responses, a green "correct" word appeared on screen for 1500 ms. For incorrect responses, a red "incorrect" word was presented. Subjects had only 30 seconds to press a button, otherwise a "too slow" message appeared on screen. Training phase was composed of 3 blocks with 80 trials each, for a total of 240 trials. After subjects completed training, they were moved to the transfer phase. In transfer, subjects had to rate the eight possible exemplars using a similarity scale. The scale ranged from 1 (more similar to Family Z) to 8 (more similar to family M), without a middle point. Each exemplar was rated only once. However other procedures like generalization and category membership ratings are also possible as means of ascertaining what is it that subjects learned during training.

Results

Training results: Results of the training phase revealed that participants in condition AB achieve a mean accuracy of .58 (SD=.06), .63 in condition BC (SD = .08) and .62 in condition CA (SD=.07). In none of the conditions did subjects approach optimal performance, nor did they show signs of using the most diagnostic feature as a rule for classification (see design section). This is likely to be the result of the PF procedure, as will be discussed below. A factorial 3 (conditions: AB, BC, CA) x 3 (blocks) design with the last being the repeated measure factor, revealed a main effect of block (F(2,66)=9.55, MSe=.05, p<.001, η_p^2=.22, power=.98), and a non-significant interaction between block and condition (p>.05). These results suggest that there was a learning effect across blocks in every condition. Contrast comparisons revealed a significant difference between block 1 and block 2 (F(1,33)=9.99, MSe=.12, p=.003, η_p^2=.23, power=.87) and a non-significant difference between block 2 and block 3 (p>.05). This suggests that in the three conditions learning occurred primarily from block 1 to 2, and it does not appear further changes in learning occurred from block 2 to 3 (see Fig. 2).

Additionally, we compared our classification predictions given by eq. (1) with training accuracy. We averaged accuracy in block 3 for every exemplar in each condition. As the analysis above reveals, most of the subjects had learned the classification criterion by block 3. Because we wanted to compare two linear trajectories (the predicted and the observed), we directly estimated the R^2 between the observed and predicted responses (based on the average observed vs. predicted probability for each exemplar). As Fig. 3 shows, there is a clear tendency for subjects in every condition to learn the exemplar classification probability. In general, fittings are fairly good (for condition AB R^2=.90, for condition BC R^2=.97, and for condition CA R^2=.89). We suspect that a higher number of training blocks would lead to a significant increase in model fittings. As will be discussed shortly, our results show that by providing PF as our experiment illustrates, subjects are able to learn to classify close to what eq. (1) stipulates. This suggests that our procedures could in fact be used to flexibly design PF experiments by fixing desired p(category|feature) probabilities and easily deriving other corresponding values of interest from that starting point.

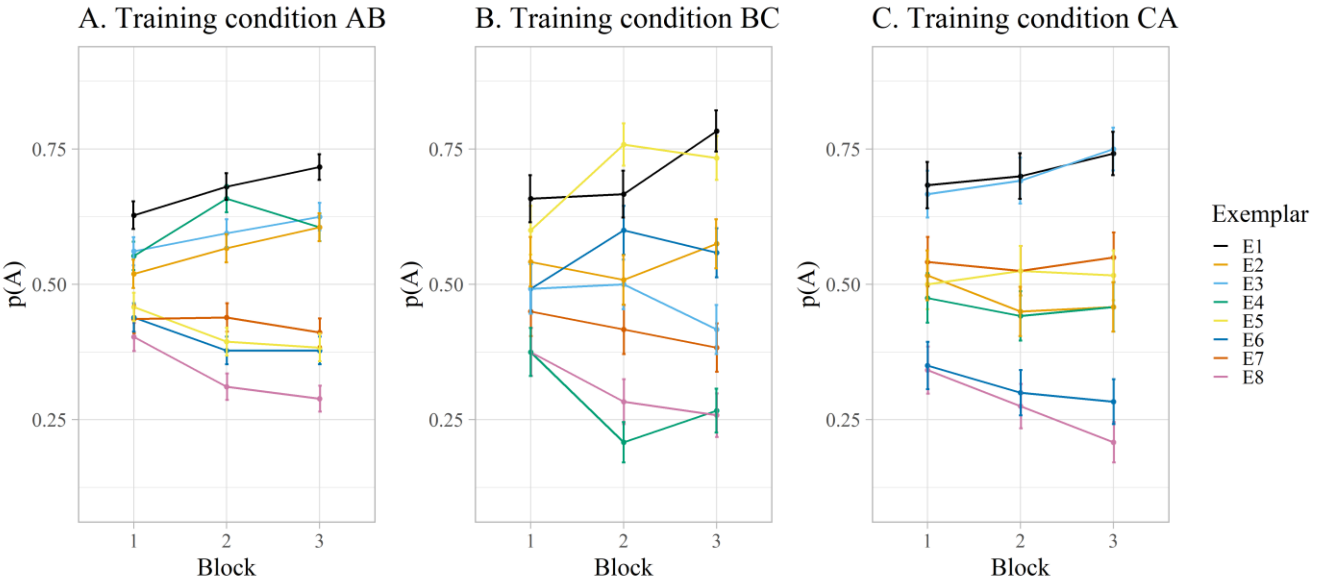

Figure 2. Classification probability of category A (Family Z) given each specific exemplar combination across blocks, showing that for each condition (AB, BC and CA) subjects approximately learned individual exemplars' p(A). A closer look to E8 shows that this exemplar is clearly classified in category B (1 - p(A)).

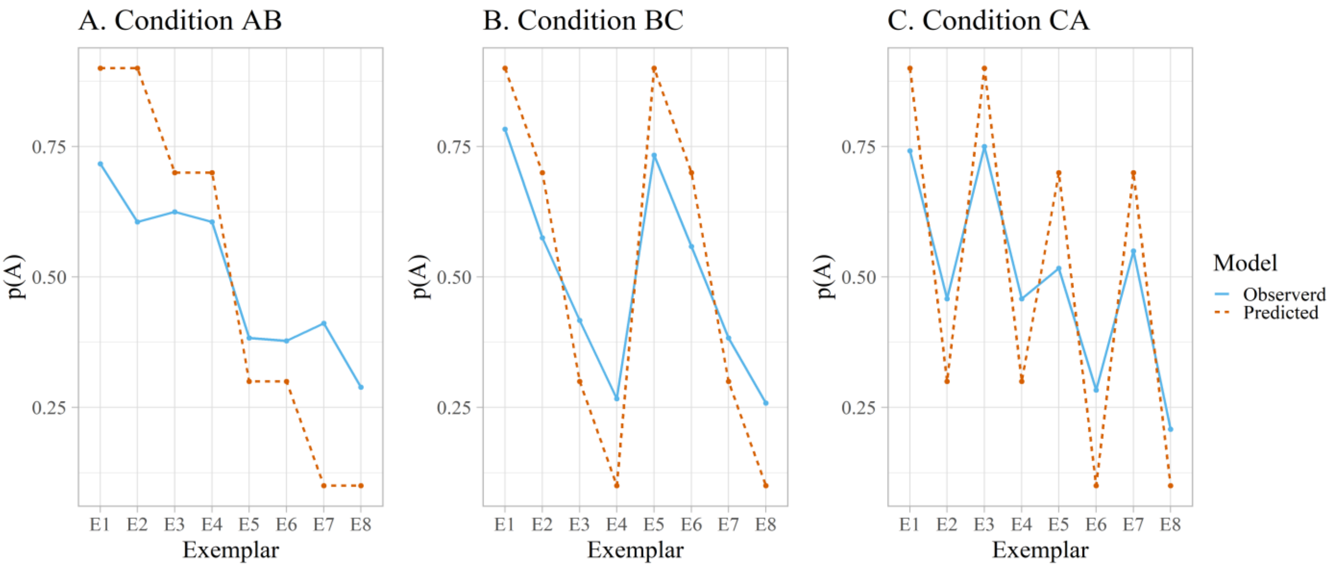

Figure 3. Probability of classification in category A for a given feature combination in each condition (AB, BC and CA). Dashed line shows predicted probabilities computed by eq. (1) and in Table 3. Continuous line shows average subject data on Block 3.

Analyzing transfer data

For transfer data we implemented individualized multiple regressions (similar to Rehder & Hastie, 2001). This method has been used in studies of conceptual representation (Puebla & Chaigneau, 2014; Rehder, 2003) and allows obtaining the relative feature weight implied in a set of category membership or similarity ratings. To implement individualized multiple regression on transfer data, we do as follows: First, similarity ratings were collected, as described in the Materials and Procedure section. Second, we used as predictors the effect coding shown in Table 1. Using these values as predictors in a regression equation allows us to predict a participant's similarity ratings. Essentially, this is possible because during transfer participants rate all possible feature combinations, and because these features are by design independent from each other (i.e., across exemplars, features are uncorrelated). Because individualized regression equations yield coefficients for single properties, these regression coefficients can then be used as individual data points reflecting, across participants, the contribution of each predictor variable to the ratings. Furthermore, the distribution of coefficients across participants can then be submitted to significance tests.

Transfer results: For the transfer phase we implemented the individualized regression method explained above. A repeated measure 3 (condition: AB, BC, CA) x 3 (feature-coefficient: A, B, C) mixed ANOVA, with the last being the repeated measure factor revealed a non-significant main effect of feature-coefficient (*F*(2,66)=2.91, *MS*e= 3.1, *p*=.06, η_p^2=.08, power=.55), but a significant interaction (*F*(4,66)=10.36, *MS*e=11.02, *p*<.001, η_p^2=.39, power=.99). To follow up on this significant interaction we performed one-way ANOVAs with feature-coefficients as dependent variables. The ANOVAs revealed a significant difference for feature-coefficient A (*F*(2,33)=6.33, *MS*e=5.97, *p*=.005), a significant difference for feature-coefficient B (*F*(2,33)=4.50, *MS*e=4.77, *p*=.019) and a significant difference for feature-coefficient C (*F*(2,33)=12.17, *MS*e=11.96, *p*<.001). Post-hoc tests using Bonferroni correction revealed a significant difference for feature-coefficient A between conditions AB and BC (*p*=.005), a significant difference for feature-coefficient B between conditions BC and CA (*p*=.024), and a significant difference for feature-coefficient C between conditions CA and AB (*p*<.001) and between CA and BC (*p*=.001). These post-hoc tests suggest that subjects in each condition were indeed focusing more on the relevant feature (f_{weight}=2.5), less on the less-relevant feature (f_{weight}=1) and completely ignoring the irrelevant feature (f_{weight}=0). In other words, subjects in each condition were learning the feature to outcome associations (see Fig. 4).

To corroborate this last hypothesis, we collapsed our data by condition and reordered each feature-coefficient by its weight. Then, we ranked our coefficients by their individual weights. We submitted our ranked weights to a one-way ANOVA with coefficient value as the dependent variable. We found a significant difference in ranked weights (*F*(2,107)=21.40, *MS*e=21.73, *p*<.001). Planned comparisons revealed a significant difference between weight 2.5 and the average between weight 1 and weight 0 (*t*(105)=-5.64, *p*<.001). And a non-significant difference between weight 1 and weight 0 (*t*(105)=1.47, *p*=.146).

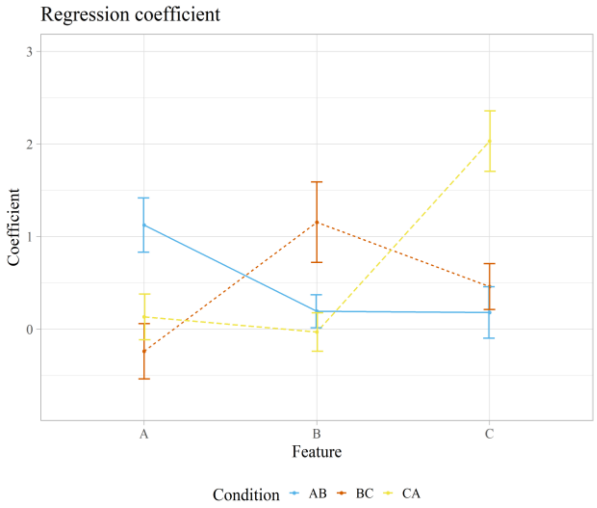

Figure 4. Regression coefficient weights obtained through individualized multiple regression method on transfer data across experimental conditions.

Discussion

Most category learning experiments use DF, while PF is underrepresented. However, not only is PF arguably more representative of natural learning conditions, but it is also possible that results obtained with DF may change when PF is used (e.g., Little & Lewandowsky, 2009).

In the current work, we have argued that a possible reason explaining why researchers don't make greater use of PF is that current PF paradigms are not flexible enough to accommodate different designs (e.g., the WP task; Knowlton, Squire, & Gluck, 1994). In contrast, in the current work we offer a way of designing PF category learning experiments that allows easily achieving any design with the desired probability characteristics. The method that the current experiment illustrates allows designing exemplars by setting from the start the relative contributions of every feature, and deriving feature combination probabilities (i.e., whole exemplar probabilities) from those relative contributions. Furthermore, by manipulating parameter c in eq. (1), a whole family of exemplar probabilities can be obtained, all of it consistent with the desired feature relative contributions.

In the current work we have illustrated the use of the PPD task and shown that the obtained classification probabilities can be used to guide PF such that subjects are able to consistently learn the desired classifications. In the category learning part of the task, subjects not only showed learning across blocks, but they also achieved classification performances very close to those implied by the task's design (Figs. 2 and 3). During transfer, subjects provided evidence of perceiving individual features' conceptual weights in the pattern intended by our design. Though only the most diagnostic feature showed a statistically higher regression coefficient, the pattern of means exhibited the predicted order (i.e., $f_1 > f_2 > f_3$). Current work in our laboratory is manipulating feature strengths so that subjects provide evidence of learning not only the correct pattern of means, but also the correct pattern of statistical differences implied in the design. Importantly, the PPD method can be used to parametrically vary different aspects of the task, such that many interesting issues can be explored. A few of them are offered here. We expect that the task will allow researchers to estimate how much randomness will make a classification task unlearnable, and how much determinism leads subjects to develop explicit categorization rules. Also, the task should allow researchers to obtain converging evidence with a different experimental setup, that PF promotes learning a category's internal structure, even though subjects learn the category by classification in contrast to learning it by making inferences. Furthermore, our experimental design could be applied to tasks with non-binary features. By discretizing continuous features into n levels (where $n > 2$), it is trivial to apply the same procedures described here, with the only limitation being the increase in the total number of exemplars necessary during training to cover all possible combinations.

In summary, the PPD task could allow researchers to flexibly design PF classification experiments beyond relatively fixed existing alternatives such as the WP task. We believe this is a contribution to mathematical modeling in cognition as well as providing new insights into learning and categorization.

Open science statement

All data is available for download at Open Science Framework (OSF): https://osf.io/x5269/

Acknowledgment

This research received financial support from graduate scholarship to the first author and ANID/Fondecyt grant 1190006 to the second author.

References

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. Trends in Cognitive Sciences, 5(5), 204–210. https://doi.org/10.1016/S1364-6613(00)01624-7

Ashby, F. G., & Gott, R. E. (1988). Decision Rules in the Perception and Categorization of Multidimensional Stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53. https://doi.org/10.1037/0278-7393.14.1.33

Ashby, F. G., & Valentin, V. V. (2018). The Categorization Experiment: Experimental Design and Data Analysis. Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, 1–41. https://doi.org/10.1002/9781119170174.epcn508

Chin-Parker, S., & Ross, B. H. (2002). The effect of category learning on sensitivity to within-category correlations. *Memory and Cognition*, *30*(3), 353–362. https://doi.org/10.3758/BF03194936

Ell, S. W., Smith, D. B., Peralta, G., & Hélie, S. (2017). The impact of category structure and training methodology on learning and generalizing within-category representations. *Attention, Perception, and Psychophysics*, *79*(6), 1777–1794. https://doi.org/10.3758/s13414-017-1345-2

Gluck, M. A., & Bower, G. H. (1988). From Conditioning to Category Learning : An Adaptive Network Model. *Journal of Experimental Psychology: General*, 117(3), 227–247. https://doi.org/https://doi.org/10.1037/0096-3445.117.3.227

Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do people solve the "weather prediction" task?: Individual variability in strategies for probabilistic category learning. *Learning and Memory*, *9*(6), 408–418. https://doi.org/10.1101/lm.45202

Hoffman, A. B., & Rehder, B. (2010). The Costs of Supervised Classification: The Effect of Learning Task on Conceptual Flexibility. *Journal of Experimental Psychology: General*, *139*(2), 319–340. https://doi.org/10.1037/a0019042

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*(5280), 1399–1402. https://doi.org/10.1126/science.273.5280.1399

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning Memory*, *1*(2), 106–120. https://doi.org/10.1101/lm.1.2.106

Kruschke, J. K., & Johansen, M. K. (1999). A Model of Probabilistic Category Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *25*(5), 1083–1119. https://doi.org/10.1037/0278-7393.25.5.1083

Lagnado, D. A., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. Journal of Experimental Psychology: General, 135(2), 162–183. https://doi.org/10.1037/0096-3445.135.2.162

Little, D. R., & Lewandowsky, S. (2009). Better Learning With More Error: Probabilistic Feedback Increases Sensitivity to Correlated Cues in Categorization. Journal of Experimental Psychology: Learning Memory and Cognition, 35(4), 1041–1061. https://doi.org/10.1037/a0015902

Luce, R. D. (1963). *Detection and recognition*. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), Handbook ofmathematical psychology (pp. 103– 189). New York: Wiley

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Meeter, M., Myers, C. E., Shohamy, D., Hopkins, R. O., & Gluck, M. A. (2006). Strategies in probabilistic categorization: Results from a new way of analyzing performance. *Learning and Memory*, 13(2), 230–239. https://doi.org/10.1101/lm.43006

Meeter, M., Radics, G., Myers, C. E., Gluck, M. A., & Hopkins, R. O. (2008). Probabilistic categorization: How do normal participants and amnesic patients do it? *Neuroscience and Biobehavioral Reviews*, *32*(2), 237–248. https://doi.org/10.1016/j.neubiorev.2007.11.001

Monchi, O., Petrides, M., Petre, V., Worsley, K., & Dagher, A. (2001). Wisconsin card sorting revisited: Distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *Journal of Neuroscience*, *21*(19), 7733–7741. https://doi.org/10.1523/jneurosci.21-19-07733.2001

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. Psychological Review, 101(1), 53–79. https://doi.org/10.1037/0033-295x.101.1.53

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods.* 10.3758/s13428-018-01193-y

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

Puebla, G., & Chaigneau, S. E. (2014). Inference and coherence in causal-based artifact categorization. *Cognition*, *130*(1), 50–65. https://doi.org/10.1016/j.cognition.2013.10.001

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407. https://doi.org/10.1016/0010-0285(72)90014-X

Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science*, *27*(5), 709–748. https://doi.org/10.1016/S0364-0213(03)00068-5

Rehder, B., Colner, R. M., & Hoffman, A. B. (2009). Feature inference learning and eyetracking. *Journal of Memory and Language*, *60*(3), 393–419. https://doi.org/10.1016/j.jml.2008.12.001

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, *130*(3), 323–360. https://doi.org/10.1037/0096-3445.130.3.323

Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning Nonlinearly Separable Categories by Inference and Classification. *Journal of Experimental Psychology: Learning Memory and Cognition*, *28*(3), 585–593. https://doi.org/10.1037/0278-7393.28.3.585