Paper accepted for publication in *Psychological Research*
Vol. 0 (2019), pp.~1–22 ISSN XXX-XXXX-XXXX

Novel representations that support rule-based categorization are acquired on-the-fly during category learning

Fabian A. Soto
Department of Psychology
Florida International University

F. Gregory Ashby
Department of Psychological & Brain Sciences
University of California, Santa Barbara

Humans learn categorization rules that are aligned with separable dimensions through a rule-based learning system, which makes learning faster and easier to generalize than categorization rules that require integration of information from different dimensions. Recent research suggests that learning to categorize objects along a completely novel dimension changes its perceptual representation, making it more separable and discriminable. Here we asked whether such newly learned dimensions could support rule-based category learning. One group received extensive categorization training and a second group did not receive such training. Later, both groups were trained in a task that made use of the category-relevant dimension, and then tested in an analogical transfer task (Experiment 1) and a button-switch interference task (Experiment 2). We expected that only the group with extensive pre-training (with well-learned dimensional representations) would show evidence of rule-based behavior in these tasks. Surprisingly, both groups performed as expected from rule-based learning. A third experiment tested whether a single session (less than one hour) of training in a categorization task would facilitate learning in a task requiring executive function. There was a substantial learning advantage for a group with brief pre-training with the relevant dimension. We hypothesize that extensive experience with separable dimensions is not required for rule-based category learning; rather, the rule-based system may learn representations "on the fly" that allow rule application. We discuss what kind of neurocomputational model might explain these data best.

Keywords: categorization, rule learning, perceptual dimension, separability

People often face the challenging task of learning new object categories based on visual properties. Children must quickly master many such categories, and similar category learning is required of adults who are learning a new skill, such as X-ray diagnostics, or a new hobby, such as bird watching. In most cases, the object categories that we encounter cannot be recognized on the basis of a few psychologically differentiated dimensions, such as object height or color. Rather, categorization of new objects requires integrating information from a large number of features that perceptually interact with one another. Yet, most categorization studies use stimuli varying along a handful of dimensions known to be easily extracted and psychologically differentiated. Important concepts in the categorization literature, such as the influence of selective attention (Goldstone, 1994b; Nosofsky, 1986), task difficulty (Zaki & Kleinschmidt, 2014), category structure (Smith, 2014), and learning strategy (Ashby et al., 1998; Ashby & Valentin, 2005), are only meaningful when interpreted in relation to such pre-existing dimensional representations.

What happens when dimensions are not available to describe stimuli in a categorization task? One possibility is that novel stimulus features and dimensions are acquired through perceptual learning (Dosher & Lu, 2017; Goldstone et al., 2009) accrued during categorization training. For example, a large body of work suggests that categorization training modifies the perceptual representation of the stimuli involved (for reviews, see Goldstone et al., 2009; Goldstone & Hendrickson, 2010), including the extraction and representation of novel psychologically differentiated stimulus dimensions, or *dimension differentiation* (Folstein et al., 2012; Goldstone & Steyvers, 2001; Soto & Ashby, 2015). Such di-

Preparation of this article was supported by NIH grant 2R01MH063760 and by the US Army Research Office through the Institute for Collaborative Biotechnologies under Grant W911NF-07-1-0072. The U.S. government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government. Correspondence should be addressed to Fabian A. Soto, Department of Psychology, Florida International University, 11200 SW 8th St, AHC4 460, Miami, FL 33199; Phone: 305-348-8423; Email: fasoto@fiu.edu.

mension differentiation involves increased discriminability (Folstein et al., 2012; Goldstone & Steyvers, 2001; Van Gulick & Gauthier, 2014) and perceptual separability (Soto & Ashby, 2015) of the category-relevant dimension. In addition, recent evidence from the cognitive neuroscience literature indicates that this form of learning is accompanied by changes in visual cortex representations (Ester et al., 2017; Folstein et al., 2013).

These results highlight the continuous and dynamic interaction between cognition and perception (for recent reviews, see Collins & Olson, 2014; Goldstone et al., 2015). However, while the influence of categorization on perceptual representation has been widely studied, there has been far less interest in understanding whether novel dimensions acquired through dimension differentiation produce the same kind of behavior as traditional psychologically differentiated dimensions in categorization tasks. For example, there is much evidence showing that categorization rules that are aligned to separable dimensions have a number of properties suggestive of "rule learning": compared to categorization rules that are not aligned to separable dimensions (i.e., "information-integration" categorization tasks, see Ashby & Gott, 1988), they are learned faster (Smith et al., 2010; Smith & Ell, 2015), learning does not depend on immediate feedback (Maddox et al., 2003; Maddox & Ing, 2005) and does not require any feedback at all under certain conditions (Ashby et al., 1999), and they produce knowledge that can be easily transferred to new responses (low response-specificity: Ashby et al., 2003) and stimuli (low stimulus-specificity: Casale et al., 2012).

If categorization training produces learning of novel separable dimensions, and these dimensions in turn facilitate rule-based category learning, then this two-way interaction significantly increases the brain's ability to adapt quickly to environmental challenges (Goldstone et al., 2015). Rule-based category learning is fast and flexible in comparison to other learning mechanisms (e.g., Ashby et al., 2003; Casale et al., 2012; Smith & Ell, 2015), allowing people to acquire knowledge more quickly about object categories and their associations with behaviorally-significant events, and to generalize more precisely such knowledge to novel category exemplars. Accordingly, people attempt to apply rules in unfamiliar categorization tasks, even when this is an unadaptive strategy (Ashby et al., 1999). However, rule learning requires stimulus representations that can support the proposal and testing of dimensional rules (Ashby et al., 1998; Hélie et al., 2015); that is, it requires a representation in terms of separable dimensions. If such a representation can be learned during a categorization task, then rule-based category learning can be applied in most circumstances.

In a series of behavioral experiments, we explored the question of whether newly-learned dimensions support the kind of rule-based category learning observed with traditional separable dimensions. We focused on three well-known properties of rule learning in categorization: learning is fast, has low stimulus-specificity, and has low response-specificity. As in previous research (e.g., Folstein et al., 2012; Goldstone & Steyvers, 2001; Soto & Ashby, 2015), we created multidimensional stimuli by morphing unfamiliar faces (see Figure 1 and description in Methods section of Experiment 1). Specifically, we first took two parent faces and morphed them to several degrees, to create a single face dimension. Then, we took two dimensions created this way and morphed each of their levels to obtain a two-dimensional space.

A wealth of evidence suggests that dimensions created this way are integral (Blunden et al., 2015; Folstein et al., 2012; Goldstone & Steyvers, 2001; Soto & Ashby, 2015), meaning that before any training they cannot be extracted from the stimuli and selectively attended. In particular, the specific morphed face dimensions illustrated in Figure 1, and used in the present study, have been previously shown to be perceptually integral ahead of training according to a variety of tests (see Soto & Ashby, 2015). Evidence of the integrality of those specific dimensions comes from a strong Garner interference effect (Garner, 1974), a failure of marginal accuracy invariance, a failure of response time invariance, and a model-based analysis of data using general recognition theory (for reviews of these tests, see Ashby & Soto, 2015; Soto et al., 2017). That is only the evidence showing integrality *in the specific stimuli used here*. In addition, previous research suggests that morphed dimensions *in general* are integral. Goldstone & Steyvers (2001) found a Garner interference effect using their morphed face stimuli, and Blunden et al. (2015) found that multidimensional scaling modeling of similarity ratings obtained from those stimuli were fitted better by an euclidean metric (known to be related to integral dimensions) than by a city-block metric (known to be related to separable dimensions; see Garner, 1974; Soto & Wasserman, 2010b). Folstein et al. (2012) found that orthogonal and diagonal boundaries are learned equally well with stimuli varying along morphed car dimensions, which is usually found with integral but not separable dimensions. In sum, every study published in the literature has shown that morphed dimensions like those illustrated in Figure 1 are integral rather than separable. This body of converging evidence means that the assumption of integrality in morphed dimensions in general, and in the stimuli used here in particular, is strongly supported.

As rule-based category learning is thought to depend on selective attention to separable dimensions (Ashby et al., 1998; Hélie et al., 2015), unknown morphed dimensions cannot support rule-based category learn-

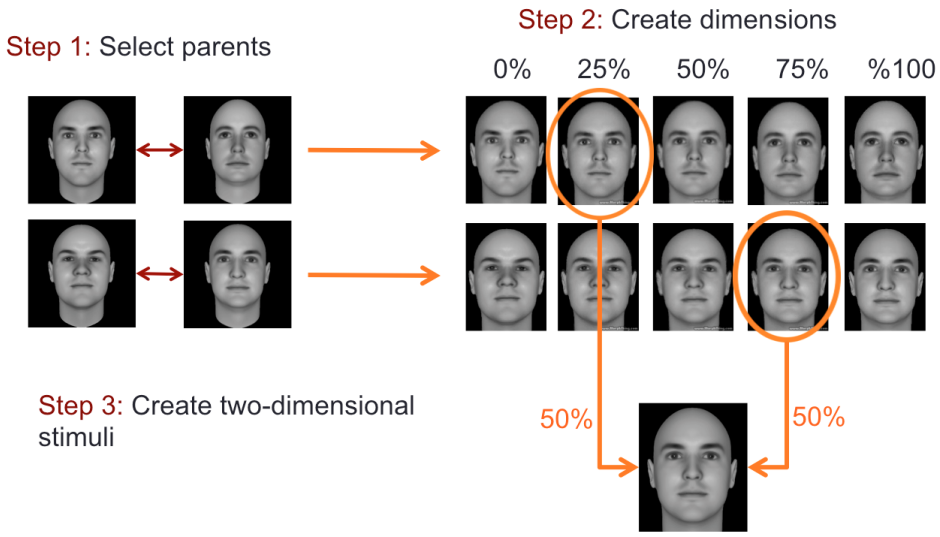

Figure 1. Schematic representation of the procedure used to create a two-dimensional space of morphed faces. Reprinted from *Cognition*, Vol 139, Fabian A. Soto and F. Gregory Ashby, "Categorization training increases the perceptual separability of novel dimensions", Pages 105-129, Copyright 2015, with permission from Elsevier.

ing. Instead, morphed faces seem to change in a variety of different shape dimensions at the same time. Correct categorization of stimuli on the basis of a morphed face dimension requires integrating all these sources of information before a categorization decision is made. Such pre-decisional processing is the landmark of information-integration categorization tasks (Ashby & Gott, 1988), which are thought to be solved through a procedural learning mechanism (Ashby et al., 1998, 2011; Ashby & Valentin, 2005).

For these reasons, we expected that people who are completely naive to the dimensions would show no evidence of rule-based category learning in a task in which the categorization rule is aligned to one of the morphed dimensions. On the other hand, extensive categorization training with stimuli varying on morphed dimensions makes them more psychologically privileged (Folstein et al., 2012; Goldstone & Steyvers, 2001) and increases their separability (Soto & Ashby, 2015). As dimensional differentiation and separability increase, so does the ability to extract the relevant dimensions from the stimuli and pay selective attention to them. For this reason, we expected that people who had extensive training in a categorization task using a particular morphed dimension would show evidence of rule-based category learning in a new task in which the categorization rule is aligned to the known dimension.

We tested rule-based category learning through an analogical transfer test measuring flexible generalization across irrelevant stimulus dimensions (Casale et al., 2012) in Experiment 1, and through a button-switch interference test measuring response-specificity of category learning (Ashby et al., 2003) in Experiment 2. Unexpectedly, we found that people with or without previous extensive training could show evidence of rule-based category learning, showing levels of generalization and behavioral flexibility that were similar to those supported by previously known face dimensions like gender or emotional expression. These levels of generalization and behavioral flexibility were higher than those observed in a task requiring integration of information from previously known dimensions. This suggests that extensive categorization training is not necessary for rule-based category learning. Instead, representations that support the use of rule-based categorization seem to be learned on-the-fly during categorization training with stimuli that lack a previous dimensional structure. Because this hypothesis was post-hoc and we only had negative evidence supporting it, in Experiment 3 we tested whether limited categorization pre-training would facilitate subsequent learning of a complex categorization task thought to require executive function. The results supported the hypothesis of fast learning of representations that support rule-based category learning.

Experiment 1

Casale et al. (2012) reported that, across a variety of conditions, learning of categorization tasks that require extracting information from a single separable dimension leads to strong generalization to novel stimuli, even when

those new stimuli are relatively dissimilar to the original training stimuli. On the other hand, learning of categorization tasks that require the integration of information from two separable dimensions leads to poor transfer to novel stimuli having about the same dissimilarity relation to training stimuli. Casale et al. explained their results in terms of multiple systems of category learning. The unidimensional task is learned by explicitly testing different dimensional rules, which allows easy transfer of the chosen rule to new circumstances. This "rule-based" generalization of category learning was termed *analogical transfer.* On the other hand, the information-integration task is learned by associating relatively small regions of perceptual space with responses (Ashby & Waldron, 1999), which limits transfer of this association only to stimuli within such small regions, through "similarity-based" generalization.

In line with this interpretation, a body of evidence suggests that generalization of learning may depend on at least two different mechanisms in people (Livesey & McLaren, 2009; Natal et al., 2013; Perez et al., 2018; Shanks & Darby, 1998). In similarity-based generalization, transfer of learning is limited to stimuli that are perceptually similar to the training stimuli. In rule-based generalization, transfer of learning depends on whether or not a known rule can be applied to the new stimulus, regardless of its perceptual similarity to those stimuli experienced during learning of the rule. Different participants may show either similarity-based generalization or rule-based generalization in the same task and for the same stimuli (Livesey & McLaren, 2009; Natal et al., 2013), and the likelihood of observing rule-based generalization can be manipulated through explicit instruction (Natal et al., 2013).

In the present experiment, we used a version of the Casale et al. (2012) test to determine whether newly-learned dimensions can support rule-based generalization. Participants were trained in a categorization task in which stimuli varied along two novel morphed face dimensions, with only one of these dimensions being relevant for correct categorization. After training, participants were tested with new faces that varied both in the category-relevant dimension and in a completely new morphed face dimension. Thus, as in the study by Casale et al., the testing stimuli were perceptually dissimilar to the training stimuli, but they could be classified according to the learned rule.

Using our morphing procedure (Figure 1), we created several two-dimensional face spaces from which we could obtain stimuli to present to participants in our experiments. Figure 2 shows a schematic representation of the stimuli and tasks used for each group. For the two experimental (ID) groups (see Figure 2, top panel), the two dimensions were completely novel, created from unfamiliar identities. Group ID-learned received three sessions of categorization pre-training, followed by one session of categorization training using the same task (stimulus coordinates and category boundary) as during pre-training, but new stimuli and response labels. The morphed face dimension parallel to the category boundary, and thus irrelevant to the categorization task, was changed from pre-training to training. The extensive categorization pre-training to which group ID-learned was exposed is known to increase separability of the category-relevant dimension, even when the trained dimension is combined with a novel irrelevant dimension (Soto & Ashby, 2015). On the other hand, group ID-new received the same categorization training, but without any pre-training sessions, which means that for this group the dimension relevant for the categorization task was completely novel. Generalization of category learning was tested in both groups using stimuli created from the category-relevant dimension and a completely new irrelevant dimension. We expected that the availability of a learned dimension in group ID-learned would facilitate rule-based learning during the categorization task, in turn facilitating generalization of learning to stimuli during testing. On the other hand, because for group ID-new the category-relevant dimension is completely new, we did not expect this group to use rule-based learning during the categorization task, which should result in a lower level of generalization than group ID-learned during testing.

All other groups were controls included to determine the level of generalization that could be expected from categorization tasks using familiar face dimensions (that is, dimensions that can be extracted from the stimuli and selectively attended without any pre-training). Thus, all control groups received a single session of categorization training followed by a generalization test. In the GEN/ID control group (see Figure 2, middle panel), one of the parent faces previously used to create the category-relevant dimension in the ID groups was replaced by a female , producing a gender categorization task. The category-irrelevant dimensions were created from the same parent faces as in the ID groups, but each combined with a female face to generate novel gender-neutral dimensions.

For the three remaining groups, stimuli varied in two known dimensions: gender and emotional expression (angry/sad; see Figure 2, bottom panel). In group GEN/EMO, the category-relevant dimension was gender and the category-irrelevant dimension was emotion, as represented by the vertical category boundary in Figure 2 (in red color). In group EMO/GEN, the category-relevant dimension was emotion and the category-irrelevant dimension was gender, as represented by the horizontal category boundary in Figure 2 (in blue color). In group

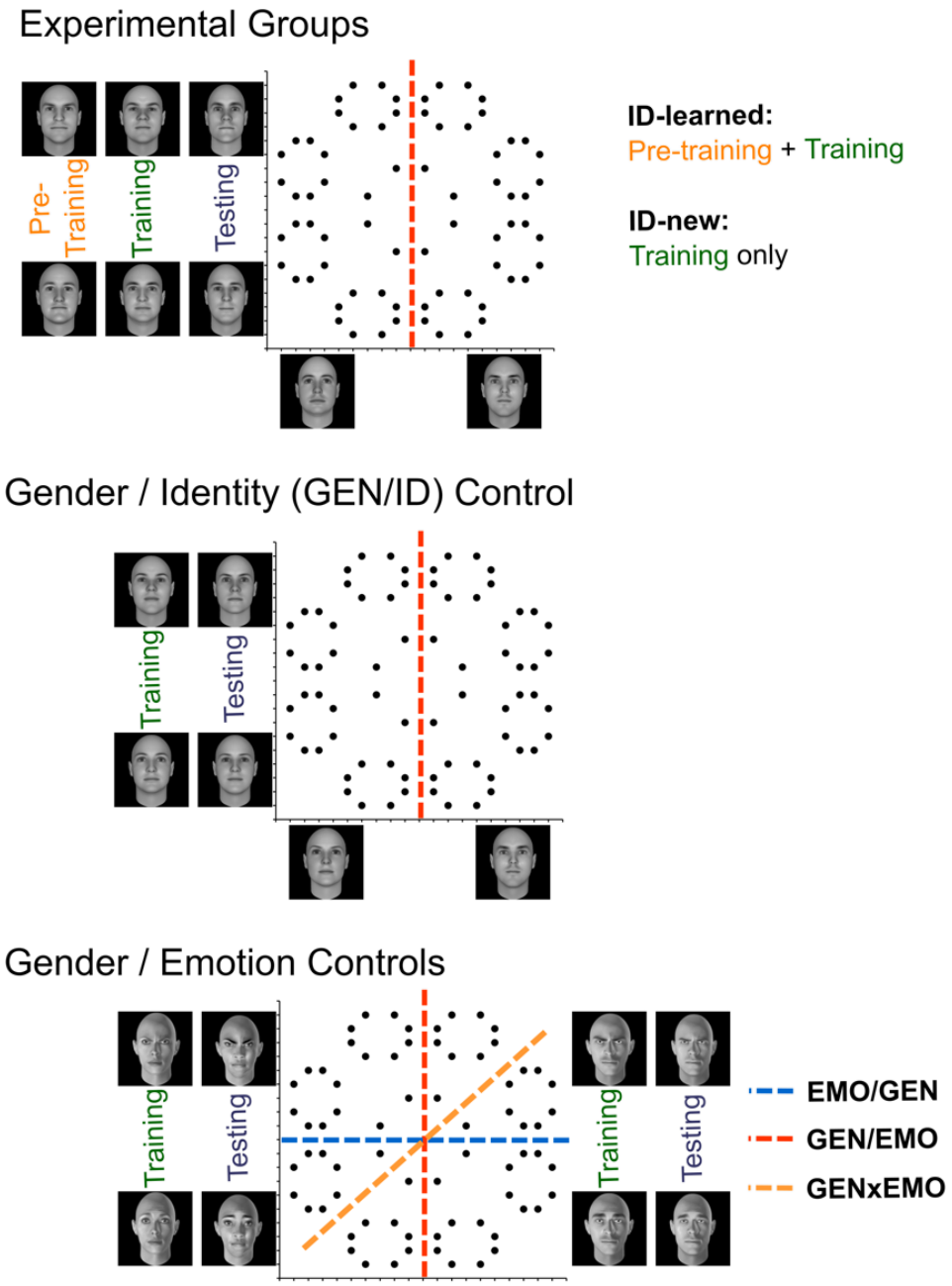

Figure 2. Schematic representation of the stimuli and tasks used in Experiments 1 and 2. The faces shown next to each dimension represent the parents for that specific dimension. When more than one pair of parents is shown, they have been labeled according to the phase of the experiment in which they were used to obtain stimuli (Pre-Training, Training, or Testing). The points inside the coordinate system represent stimuli obtained from a specific combination of levels for each dimension. The dotted lines separating such points in two classes represent the category boundary used for training. For more details on the stimuli and tasks used for each particular group, see the main text.

GENxEMO, learning of the categorization task required integrating information from both gender and emotion, as demonstrated by the diagonal category boundary in Figure 2 (in orange color). For all these groups, the parents used to create training stimuli and testing stimuli differed in identity, but had the same gender and emotional expression.

The groups GEN/ID, GEN/EMO and EMO/GEN provided an estimate of the generalization level that we should observe in a categorization task aligned to familiar face dimensions, a form of "ceiling" level for analogical transfer. On the other hand, group GENxEMO provided an estimate of how much generalization of learning should be observed in a categorization task requiring integration of information from two familiar face dimensions (see Casale et al., 2012), a form of "floor" level for analogical transfer. While those four control groups provided a means to assess different levels of analogical transfer, the most interesting results would come from group ID-learned and ID-new. ID-learned indicates how much analogical transfer is observed when the relevant category dimension has been learned through intensive previous categorization training. Such intensive training is typical of studies on dimension differentiation (e.g., Folstein et al., 2012; Goldstone & Steyvers, 2001; Soto & Ashby, 2015). If dimension learning through categorization training creates novel separable dimensions that can support rule-based generalization, then the performance of this group should be closer to the "ceiling" than to the "floor" of analogical transfer. Group ID-new was meant as a control of how much analogical transfer is observed after training with completely novel face dimensions. We expected that the brief experience with the relevant category dimension in this group would not be enough to produce a new representation capable of supporting analogical transfer. Contrary to our expectations, we found that generalization in this group was close to the estimated "ceiling" of analogical transfer, and indistinguishable from generalization in groups for whom the relevant dimension was familiar, either because of training (ID-learned) or because of prior knowledge (GEN-ID, GEN-EMO, EMO-GEN).

Materials and Methods

Participants. 172 undergraduates at the University of California Santa Barbara voluntarily participated in this experiment in exchange for class credit or a monetary compensation. There were 19 participants in group ID-learned, 36 participants in group ID-new, 35 participants in group GEN/ID, 25 participants in group GEN/EMO, 27 participants in group EMO/GEN, and 30 participants in group GENxEMO. Participants were assigned to groups in a semi-random way, with more participants assigned to groups that were expected to perform more poorly in the main training task (ID-new, GEN/ID and GENxEMO). We had pre-set performance criteria to include participants in the main analyses (see below), so more participants were needed to obtain relatively balanced group sizes despite differences in performance.

Stimuli. For all groups, morphs with different proportions of each parent face were generated in MATLAB using the factorial procedure of Goldstone and Steyvers (2001). The procedure is illustrated in Figure 1. In the first step, pairs of faces are chosen to be the parents for a dimension. The chosen images were converted to grayscale and their intensity histograms were equalized, to ensure that stimuli along the resulting morphing dimensions varied in shape features, but not in simpler features such as skin color and brightness. Different parent faces were chosen for different groups.

The stimuli shown to the ID-learned and ID-new groups were created from 8 parent images chosen from a database of 300 computer-generated caucasian faces described by Oosterhof and Todorov (2008), created using the Facegen Modeller program (http://facegen.com), Version 3.1. From the original database, 30 male faces were chosen that had similar eyebrow color and similar levels of facial fat. Four pairs of faces were chosen as parents from those 30 candidates, according to two criteria. These criteria were chosen to ensure that the final stimuli shown to groups ID-learned and ID-new did not have a clear dimensional structure; rather, the dimensional structure of these stimuli had to be learned through categorization training. The first, more formal criterion was that all pairs of faces should have relatively equivalent mean similarity, to ensure that the dimensions created from them had relatively similar salience. Following the original study by Goldstone and Steyvers (2001), dissimilarity ratings were obtained for the 30 faces in a pilot study with twelve participants. The study used the efficient method described by Goldstone (1994a) to measure dissimilarities, and individual ratings were normalized by dividing each of them by the largest rating from the participant. Thus, final dissimilarity values ranged from zero to one for all participants. Four pairs of faces with mean dissimilarities within 15% of each other were chosen as parents. The second, less formal criterion was that the parent pairs should not be discriminable along any easily verbalizable dimension, as judged by the experimenters. That is, we made sure that the pair of faces could not be discriminated on the basis of common face categories (e.g., sex, race, age, etc.) or non-facial features (e.g., head width, head size, facial fat, ear shape, etc.). The reader can corroborate that this is correct by looking at the parent faces displayed in the top panel of Figure 2.

The stimuli shown to the GEN/ID group were cre-

ated from some of the same parents as groups ID-learned and ID-new, but morphed with 5 different female faces taken from the Oosterhof and Todorov (2008) database. As shown in the middle panel of Figure 2, one of these females replaced one parent in the category-relevant dimension, and the other four females were randomly assigned to be morphed with each of the parents of the category-irrelevant dimensions. The resulting category-relevant dimension varied along gender, whereas the category-irrelevant dimensions varied in identity while being gender-neutral.

The stimuli shown to groups EMO/GEN, GEN/EMO and GENxEMO were morphed from parent faces harvested from the internet, which were originally created using FACSGen (Roesch et al., 2011), a tool that generates FACS-correct expressions (i.e., using the Facial Action Coding System, see Ekman et al., 1978) in face models using the Facegen Modeller program (http://facegen.com). This ensured that all synthetic parent faces were created using the same software.

In the second step of the procedure illustrated in Figure 1, each dimension is created by generating morphs with different proportions from each pair of parents. Here, dimensions were created using a continuous sequence of 19 morphs for each pair of parents, with percentages of parent 2 equal to 0%, 6%, 14%, 20%, 24%, 30%, 32%, 38%, 42%, 50%, 58%, 62%, 68%, 70%, 76%, 80%, 86%, 94%, and 100%. The third and final step was to generate a two-dimensional space by factorially combining each of the faces in each dimension with each of the faces in the other dimension. As shown in the Figure 1, these two-dimensional morphs include 50% from each of the one-dimensional morphs. The specific stimuli from the two-dimensional space presented to participants are represented by the points in Figure 2. This circular configuration of points has been used in the past to show learning of new dimensions (Folstein et al., 2012) and has the advantage that the circular arrangement de-emphasizes the dimensional structure of the stimuli (Goldstone & Steyvers, 2001). That is, without the presence of a particular category bound, the stimulus configuration is the same even if one rotates the *x*- and *y*-axes by any multiple of 45-degrees. Imagine taking the space at the bottom of Figure 2 (the panel entitled "Gender / Emotion Controls"), and rotating the axes 45-degrees clockwise. Now, the orange bound would be vertical, and it would have the same spatial relation with exemplars from the two categories at each side as the red bound had before rotation. What this means is that the actual coordinate system used to describe the stimuli is relatively arbitrary, and the category bound (i.e., the feedback received during categorization training) is the only information that participants receive about important directions in the space. Indeed, dimension differentiation is observed after training with bounds in any direction of a morphed space (Folstein et al., 2013; Goldstone & Steyvers, 2001).

Procedure. The categorization task presented to participants was the same across groups and experimental phases, with slight variations. At the beginning of each session, instructions were displayed on the screen indicating that the participant's task was to categorize faces as accurately as possible into two different categories (clubs) based purely on physical appearance. The instructions also explained the structure of each trial and how to report a categorization response. Participants were warned that they would need to guess the correct answer early in training, but they would get more accurate as the experiment progressed. Sessions were divided in blocks of 72 trials each. Each stimulus (36 per category) was presented once in a block, with the order randomized within the block. There were voluntary breaks of 1 min between blocks, which the participant could finish by clicking on a button labeled "continue." Each trial started with the presentation of a white cross in the middle of a black screen for 500 ms. Immediately afterwards a face stimulus was presented in the middle of the black screen until the participant pressed one of the two response buttons in the keyboard or a time deadline of 2 s was reached, whichever happened first. During pre-training sessions, participants could press the keys B or Y in their keyboard, which were re-labeled "X" and "Y", respectively. During training sessions, participants could press the keys D or K in their keyboard, which were re-labeled "A" and "B", respectively. After a key press, the participant received feedback about the correct response. For correct responses, the word CORRECT was presented for 500 ms, in green font color in the middle of the screen, accompanied by a pleasant chime presented through the headphones. For incorrect responses or if the time deadline was reached, the word INCORRECT was presented for 500 ms, in red font color in the middle of the screen, accompanied by an unpleasant buzzer presented through the headphones. This was followed by a 1 s inter-trial interval, during which the monitor was completely black.

Participants in group ID-learned were exposed to 3 sessions of pre-training in the categorization task shown in the top panel of Figure 2. The sessions were run within a span of three days and no more than two sessions were run on the same day. Consecutive sessions were separated by at least 1 hour and at most 25 hours, with the exception of a single pair of sessions that was separated by 10 minutes. Each pre-training session consisted of 9 blocks of 72 trials each, for a total of 648 trials. Group ID-learned was the only one exposed to these pre-training sessions.

Participants in all groups were exposed to a single ses-

sion of categorization training, in the tasks illustrated in Figure 2. All participants, regardless of group, were instructed to categorize faces as accurately as possible into two categories (clubs) based purely on physical appearance, and had to learn the categories through feedback as indicated above. The session consisted of 8 blocks of 72 trials each, for a total of 576 trials. Immediately after this training, participants were exposed to an analogical transfer test (as in Casale et al., 2012). They received new instructions indicating that they would be shown new faces, and that their task would be to correctly guess whether those faces belonged to Club A or Club B. They were also informed that they would not receive feedback about the correctness of their responses. Participants were then exposed to a single block of 72 trials with new stimuli that resulted from the combination of the trained category-relevant dimension and a completely new category-irrelevant dimension (see Figure 2). The trial structure was the same as for pre-training and training sessions, but no feedback was provided.

Participants in groups ID-learned and ID-new completed one session of a Garner filtering task and one of an identification task, in addition to the pre-training and training sessions. Results from those extra sessions are reported elsewhere (Soto & Ashby, 2015).

Data Analysis. All analyses were performed using R v. 3.2.1 (R Core R Core Team, 2015) extended with the packages *reshape2* v. 1.4.1, *plyr* v. 1.8.3, and *ggplot2* v. 2.1.0, running in RStudio v. 0.99.486 (R Studio Team, 2015). A rejection criterion of $\alpha = 0.05$ was used in all statistical tests.

Participants were excluded from all analyses if they did not reach a performance level of 70% correct during categorization training. This criterion was set before performing the main analyses.

Backward learning curves (Smith & Ell, 2015) were created by finding the first block in which each participant achieved a proportion of correct responses equal or greater than 70%, setting that block's number to zero and numbering all earlier blocks from that participant's data accordingly (i.e., -1, -2, -3... for earlier blocks, and 1, 2, 3 for later blocks). This means that the learning curve for each individual participant remains the same, reflecting proportion of correct responses in a given training block, with the only difference being a change in the numbering of blocks. The individual curves are then aligned taking block zero (i.e., the first block in which all participants reached criterion) as a reference, and the group average and standard errors are computed by block. These learning curves are sometimes useful to elucidate whether learning during a categorization task is gradual, which has been associated with procedural category learning, or step-like, which has been associated with rule-based category learning (Smith & Ell, 2015). Because of they way backward learning curves are built, most of the time they show a sudden increase in performance around block zero (Smith & Ell, 2015), even if a gradual learning process underlies behavior. Thus, the most informative features of the curve are whether it is flat or gradual before and after that point.

A generalization decrement was computed by subtracting the proportion of correct responses during the test from the proportion of correct responses during the last categorization training block. Independent *t*-tests were performed to determine whether the mean generalization decrement was significantly higher than zero in each group. The reported *p*-values from these tests were corrected for multiple comparisons using the Bonferroni procedure.

To determine whether generalization decrements varied across groups, we performed a one-way ANOVA with group as factor and generalization decrement as dependent variable, followed by pre-planned pairwise comparisons between each of the experimental groups (ID-learned and ID-new) and each of the control groups (GEN/ID, EMO/GEN, GEN/EMO, GENxEMO), as well as between the two experimental groups, for a total of 9 comparisons. These comparisons were carried out using Fisher's LSD test and the Bonferroni correction for multiple comparisons. Any non-planned (i.e., post-hoc) comparisons reported here were performed using the Newman-Keuls procedure for multiple comparisons, which considers all possible 15 comparisons between pairs of means.

Results and Discussion

The number of participants excluded from the final analysis due to poor performance (<70%) in the categorization task were: 1 from group ID-learned (18 included), 10 from group ID-new (26 included), 14 from group GEN/ID (21 included), 4 from group GEN/EMO (26 included), 2 from group GENxEMO (23 included), and 2 from group EMO/GEN (25 included).

Figure 3 shows the main results of Experiment 1. The forward learning curves depicting mean percent correct during each block of the categorization task are shown in panel a, whereas the backward learning curves are shown in panel b. Note first that groups GEN/EMO and EMO/GEN, trained in a unidimensional categorization rule, reached a high and asymptotic level of performance very quickly, by block 3. The backward learning curves reveal a sudden increase in performance at block zero (the block in which participants reached a criterion of 70% correct), without further increases in performance after this point. In contrast, group GENxEMO, trained in a diagonal (i.e., information-integration) categorization rule, shows more gradual increases in performance that never reach comparably high values. The backward

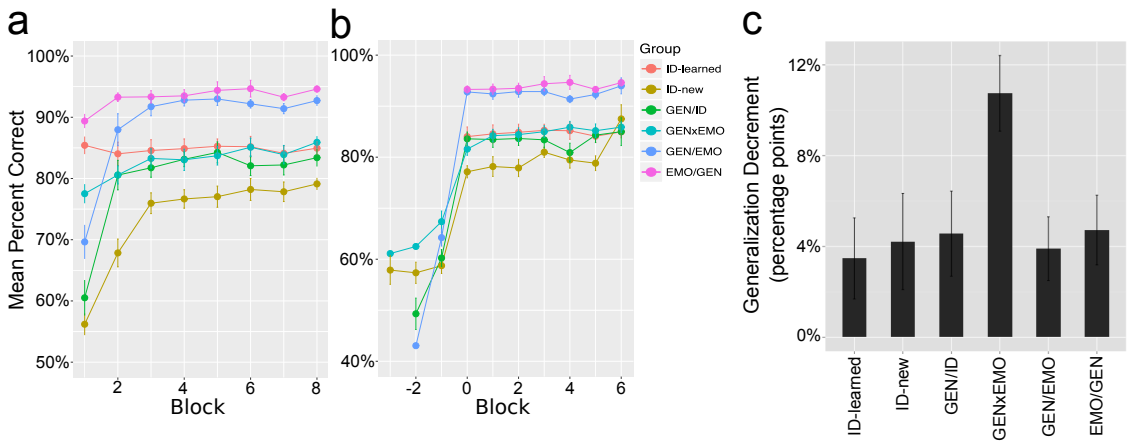

Figure 3. Results of Experiment 1. (a) Forward learning curves depicting mean percent correct during each block of the categorization task; (b) backward learning curves (see text for description) depicting mean percent correct during each block of the categorization task, where block numbers have been individually shifted so that zero represents the first block in which a participant reaches 70% correct; (c) mean generalization decrement observed as a result of a change in the category-irrelevant dimension during the analogical transfer test, with a lower value representing *better* generalization.

learning curve for this group reveals a slow increase in performance both before and after block zero, with a much smaller jump in performance at block zero. This pattern of results mirrors previous findings with simpler separable dimensions (e.g., Smith et al., 2010; Smith & Ell, 2015). Group GEN/ID showed results similar to those from the other groups trained with a unidimensional categorization rule, but with a much lower asymptotic performance.

The learning curves of group ID-learned show a pattern similar to those from the control groups with unidimensional categorization rules: asymptotic performance starting from block one. On the other hand, group ID-new shows the slowest learning and lowest performance level among all groups (Figure 3a) and a backward-learning curve without a clear pattern (Figure 3b): a flat curve before block zero is followed by a considerable jump in performance, which is a signature of rule-based learning, but a gradual improvement in performance is seen after that point, which is a signature of procedural learning.

The most important results from this experiment are shown in Figure 3c, which displays the mean generalization decrement observed in the analogical transfer test. Here a lower value represents *better* generalization of category learning to the testing stimuli. The generalization decrement observed for both experimental groups (ID-learned and ID-new) looks low and similar to that observed in control groups dealing with unidimensional categorization rules (GEN/ID, GEN/EMO and EMO/GEN). On the other hand, they seem much smaller than the generalization decrement observed in the control group dealing with a diagonal categorization rule (GENxEMO). In line with these observations, one-sample t-tests revealed a generalization decrement significantly different from zero in group GENxEMO, $t(22) = 6.45$, $p < .001$, and in group EMO/GEN, $t(24) = 3.08$, $p < .05$, but not in other groups. The ANOVA revealed a significant effect of group on the generalization decrement, $F(5, 128) = 2.37$, $p < 0.05$. Pairwise comparisons revealed that the generalization decrement of group GENxEMO was significantly different from that observed in the experimental groups (ID-learned and ID-new), but no other comparisons were significant.

Note also that the control groups GENxEMO, GEN/EMO and EMO/GEN, which are identical in all aspects except for the categorization task that they had to perform (unidimensional vs. diagonal), show the generalization results that would be expected from previous research, with high performance decrements during testing for group GENxEMO, but low performance decrements in the other two groups. Post-hoc comparisons using the Newman-Keuls correction for multiple comparisons revealed a significant difference between groups GENxEMO and EMO/GEN. Although the comparison between GENxEMO and GEN/EMO was not significant ($p = 0.059$, corrected), it seems likely that this was due to low statistical power in the post-hoc comparison. This replicates the results that Casale et al. (2012) found using gratings varying in spatial frequency and orientation, but with face stimuli varying in the dimensions of gender and emotional expression. In both cases, analogical transfer is stronger for groups trained in a unidimensional categorization task than for a group trained in an information-

integration categorization task.

In sum, both group ID-learned and ID-new showed excellent transfer, close to that found for groups that had to discriminate gender (GEN/ID and GEN/EMO) or emotional expression (EMO/GEN), but different from that observed for the group that had to integrate information from both gender and emotional expression during training (GENxEMO). This suggests that extensive training in group ID-learned was not necessary for learning of new dimensions.

Experiment 2

Another known property of rule-based learning is response flexibility: category learning can be easily transferred to a novel mapping between stimuli and responses. In contrast, procedural learning is characterized by response specificity, and does not easily transfer to a mapping between stimuli and motor responses that is different from that observed during training. For example, Ashby, Ell and Waldron (2003) trained participants in two types of categorization task. Unidimensional categorization tasks required extracting information from a single separable dimension, as in the task used for groups GEN/ID, GEN/EMO and EMO/GEN during Experiment 1. In contrast, an information-integration task required the integration of information from two separable dimensions, as in the task used for group GENxEMO during Experiment 1. After training, participants were asked to switch the response buttons used to report each category. Participants trained in the information-integration task showed a drop in performance during the button-switch test, which was termed a *button-switch interference* effect. Participants in the unidimensional task did not show the same button-switch interference effect. Ashby et al. interpreted their results as evidence that information-integration tasks, but not unidimensional tasks, are acquired via procedural learning processes.

A button-switch interference effect can be found in unidimensional tasks under specific testing conditions (Maddox et al., 2010; 2007;Nosofsky et al. 2005), possibly reflecting the fact that all tasks have a procedural component, but response costs are smaller than in information-integration tasks (Maddox et al., 2010; Nosofsky et al., 2005) and experimental manipulations can dissociate between the interference effects found in unidimensional and information-integration tasks (e.g., Maddox et al. 2004b, 2010, 2007), suggesting that they are due to different underlying mechanisms.

The present experiment used the same groups and design as Experiment 1, but participants were exposed to a button-switch test similar to that used by Ashby et al. (2003) to determine whether newly-learned dimensions can support flexible re-assignment of categories to responses.

Materials and Methods

Participants. 149 undergraduates at the University of California Santa Barbara voluntarily participated in this experiment in exchange for class credit or a monetary compensation. There were 22 participants in group ID-learned, 36 participants in group ID-new, 23 participants in group GEN/ID, 21 participants in group GEN/EMO, 21 participants in group EMO/GEN, and 26 participants in group GENxEMO.

Stimuli. Stimuli were the same as those used in Experiment 1.

Procedure. All procedures were exactly the same as those used for Experiment 1, with the exception of the test session at the end of the experiment. Here, participants were presented with a button-switch test instead of an analogical transfer test. They received new instructions indicating that the names of the two clubs (A and B) would be switched during the rest of the experiment, so that for every face to which they responded with "A" they should now respond to "B", and vice-versa. Participants were exposed to a single block of 72 trials that was exactly the same as previous categorization training blocks, but with the assignment of responses to categories reversed. As in previous studies demonstrating a difference in the button-switch interference effect between unidimensional and information-integration tasks (e.g., Ashby et al., 2003; Maddox et al., 2007,1), participants were given a long response time deadline of 2 s, as short response time deadlines induce an interference effect even in unidimensional tasks (see Nosofsky et al., 2005).

Data Analysis. All data analyses were performed exactly as described for Experiment 1.

Results and Discussion

The number of participants excluded from the final analysis due to poor performance (<70%) in the categorization task were: 3 from group ID-learned (19 included), 11 from group ID-new (25 included), 5 from group GEN/ID (18 included), 0 from group GEN/EMO (21 included) 5 from group GENxEMO (21 included), and 0 from group EMO/GEN (21 included).

Figure 4 shows the main results of Experiment 2. The forward learning curves depicting mean percent correct during each block of the categorization task are shown in panel a, whereas the backward learning curves are shown in panel b. Note first that the pattern of results for the control groups GEN/EMO, EMO/GEN and EMOxGEN is almost identical to that found in Experiment 1, which again replicates previously-observed differences in learning curves for groups trained in unidimensional versus information-integration categorization tasks (e.g., Smith et al., 2010; Smith & Ell, 2015). Group GEN/ID showed

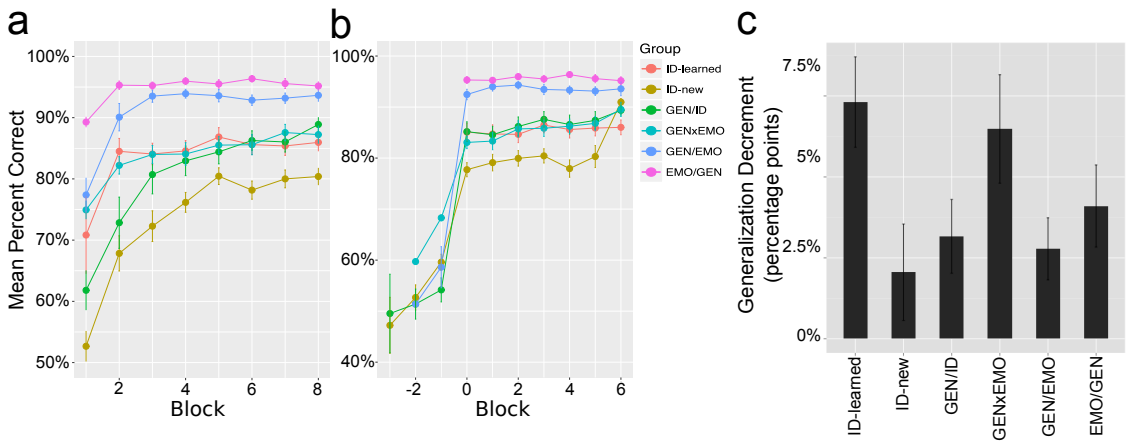

Figure 4. Results of Experiment 2. (a) Forward learning curves depicting mean percent correct during each block of the categorization task; (b) backward learning curves (see text for description) depicting mean percent correct during each block of the categorization task, where block numbers have been individually shifted so that zero represents the first block in which a participant reaches 70% correct; (c) mean generalization decrement observed as a result of a change in the mapping of categories to response keys during the button-switch interference test, with a lower value representing *better* generalization.

a more gradual increase in performance in the forward learning curve, but a step-wise backward learning curve. The pattern of results observed for groups ID-learned and ID-new are also very similar to that observed in Experiment 1, except that here it took until block 2 for group ID-learned to reach asymptotic performance in the forward learning curve.

The most important results from this experiment are shown in Figure 4c, which displays the mean generalization decrement observed in the button switch test. Here a higher value represents more interference produced by the change in the assignment of response buttons to categories during test; that is, a lower value represents *better* performance in the test. All groups showed some level of generalization decrement during test (i.e., a drop in performance due to button switch). However, for the control groups trained in a uni-dimensional categorization rule (GEN/ID, GEN/EMO and EMO/GEN), this decrement was relatively small. The experimental group without any categorization pre-training (ID-new) shows a similarly low decrement that, if anything, seems smaller than that observed in the unidimensional control groups. On the other hand, both the control group with a diagonal categorization rule (GENxEMO) and the experimental group exposed to categorization pre-training (ID-learned) show a higher generalization decrement during test. In line with these observations, one-sample t-tests revealed a generalization decrement significantly different from zero in groups GENxEMO, $t(20) = 3.87$, $p < .01$, ID-learned, $t(18) = 5.24$, $p < .001$, and EMO/GEN, $t(20) = 3.24$, $p < .05$, but the generalization decrement found in all other groups was not significant after correction for multiple comparisons. The ANOVA revealed a significant effect of group on the generalization decrement, $F(5, 119) = 2.42$, $p < 0.05$. Surprisingly, pairwise comparisons revealed that the only significant difference was between the two experimental groups. Group ID-learned did not differ significantly from any of the control groups, although it is likely that the comparison with group GENxEMO, with $p = 0.081$, did not reach significance simply due to the conservativeness of the Bonferroni correction.

Some of the results from this experiment are in line with those from the previous experiment, in that a group without any experience with the category-relevant dimension showed levels of transfer during test that were similar to those shown by groups trained in a unidimensional rule involving a familiar face dimension. This again suggests that extensive training in group ID-learned was either not necessary for learning of new dimensions, or that learning of new dimensions is not necessary for performing a button switch during categorization.

On the other hand, the results from group ID-learned are quite surprising, in that this group showed what seemed to be the strongest button-switch interference effect among all groups in the experiment. That is, a group exposed to training known to produce learning of novel separable dimensions, which in turn are thought to support rule-based category learning, showed an effect that is usually linked to procedural learning (Ashby et al., 2003).

One explanation for this counter-intuitive pattern of results is that some subjects in the ID-learned group may

have developed automatic responding due to their extensive categorization training. Response keys during pre-training and training phases had different labels (red-colored "X" and "Y" labels during pre-training; yellow-colored "A" and "B" labels during training) and spatial positions (keys B and Y during pre-training; keys D and K during training). The spatial positions were chosen to ensure a top-down arrangement during pre-training and a left-right arrangement during training. However, note that in a standard keyboard the key B is slightly to the left of key Y. This could mean that participants transferred not only their knowledge about the categories from pre-training to training, but also their knowledge of the assignment of categories to left and right response keys. Under such circumstances, they essentially received four sessions of training in which one category was assigned to the left response buttons and the other category was assigned to the right response buttons. Such extensive training would produce automatic responding, regardless of whether participants learned the task through a procedural or rule-based strategy, leading to a strong button-switch interference effect (see Helie et al., 2010).

In line with this idea, the random assignment of response keys to categories resulted in a majority of the participants included in the analysis (16 out of 19) having one category consistently assigned to "left" response buttons (i.e., B in pre-training and D in training) and the other category consistently assigned to "right" response buttons (i.e., Y in pre-training and K in training). In Experiment 2b, we confirmed that pre-training group ID-learned in a categorization task that avoids the development of automatic responding leads to a small button-switch interference, comparable to that observed in group ID-new.

Experiment 2b

In the previous experiment, we found that extensive pre-training with a categorization task in group ID-learned produced a button-switch interference effect that was significantly higher than that observed without such pre-training in group ID-new. One explanation of this counter-intuitive result is that consistent assignment of "left" response buttons to one category and "right" response buttons to another produced the development of automatic responding. Here, we repeated a smaller version of the previous experiment, including only groups ID-learned, ID-new and GEN/ID. The main goal was to show that, if group ID-learned is exposed to a task that avoids the development of automatic responding during the pre-training phase, then the difference in interference effect observed with group ID-new should disappear. We also included group GEN/ID to provide a benchmark for rule-based performance.

To avoid development of automatic responding, group ID-learned was pre-trained in a task that involved an inconsistent assignment of responses to categories. In each trial, participants were shown a stimulus and then asked "Is this an X?" or "Is this a Y?". Participants then had to respond "yes" or "no" by pressing a labeled response key. Thus, while participants had to learn an association of stimuli to category labels, each category was not associated with any specific motor response (Maddox et al., 2004b). The session involving categorization training and button-switch testing followed the exact same procedures as in the previous experiment for all three groups.

Materials and Methods

Participants. 71 undergraduates at the University of California Santa Barbara voluntarily participated in this experiment in exchange for class credit or a monetary compensation. There were 28 participants in group ID-learned, 20 participants in group ID-new, and 23 participants in group GEN/ID.

Stimuli. Stimuli were the same as those used in Experiment 1.

Procedure. All procedures were exactly the same as those used for Experiment 2, except for the task presented to participants in group ID-learned during pre-training. The only difference between this task and that described in the Procedure section of Experiment 1 is that the face stimulus was presented together with one of two possible questions: "Does this person belong to the GREEN club?" or "Does this person belong to the YELLOW club?". The question was displayed above the face image, with all words in white text except for "GREEN club", which were displayed in green color, and "YELLOW club", where were displayed in yellow color. The question displayed was randomly chosen for each trial. The keys "Y" and "B" in the keyboard were re-labeled "Y" and "N", respectively. Participants were instructed to use those keys to respond "Yes" or "No" to the question displayed in each trial. After a key press, the participant received feedback about the correct response, as described in the procedures for Experiment 1.

Data Analysis. All data analyses were performed exactly as described for Experiment 1.

Results and Discussion

The number of participants excluded from the final analysis due to poor performance (<70%) in the categorization task were: 3 from group ID-learned (25 included), 6 from group ID-new (14 included), and 2 from group GEN/ID (21 included).

Figure 5 shows the main results of Experiment 2b. The forward learning curves depicting mean percent correct during each block of the categorization task are

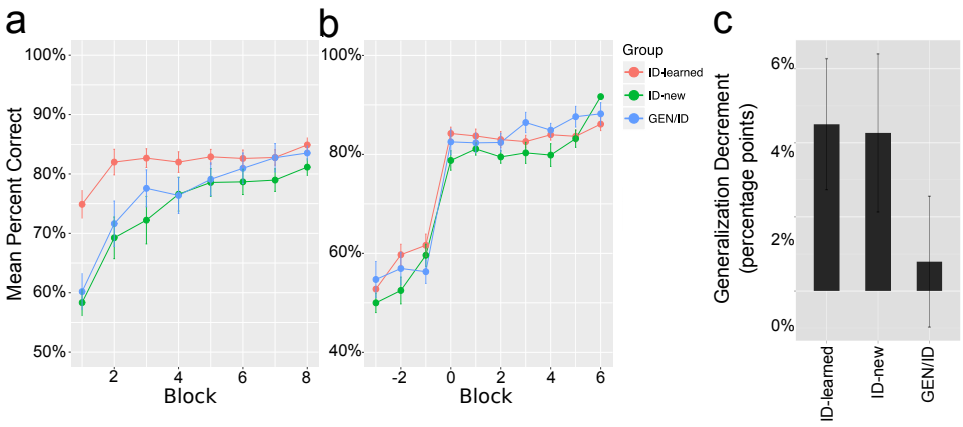

Figure 5. Results of Experiment 2b. (a) Forward learning curves depicting mean percent correct during each block of the categorization task; (b) backward learning curves (see text for description) depicting mean percent correct during each block of the categorization task, where block numbers have been individually shifted so that zero represents the first block in which a participant reaches 70% correct; (c) mean generalization decrement observed as a result of a change in the mapping of categories to response keys during the button-switch interference test, with a lower value representing *better* generalization.

shown in panel a, whereas the backward learning curves are shown in panel b. Results follow the same pattern as in previous experiments.

Figure 5c shows the mean generalization decrement observed in the button-switch test, with a higher value representing more interference resulting from the change in assignment of response buttons to categories during test. It can be seen that all groups show similarly low levels of button switch interference. In line with this observation, one-sample *t*-tests revealed non-significant generalization decrement in all groups (ID-learned: $t(24) = 2.55$, $p > .05$; ID-new: $t(13) = 2.00$, $p > .1$; GEN/ID: $t(20) = .45$, $p > .1$) and the ANOVA revealed no significant differences between groups, $F(2, 57) = 1.28$, $p > 0.1$. These results are in line with our hypothesis that participants in group ID-learned from Experiment 2 showed a large button-switch interference effect only due to extensive training (4 sessions) with a consistent mapping of categories to motor response location (left vs. right). Here, the use of a pre-training categorization task that does not consistently map between categories and motor responses produced a small generalization decrement in group ID-learned, which was comparable to that observed in group ID-new (unlike in the previous experiment).

To summarize the results of Experiments 2 and 2b: behavior in a button-switch interference test was similar to that observed in the analogical transfer test, in that a group without extensive pre-training in a categorization task (group ID-new) showed excellent generalization performance during both tests, close to that found for groups that had to discriminate gender (GEN/ID and GEN/EMO) or emotional expression (EMO/GEN), but different from that observed for the group that had to integrate information from both gender and emotional expression during training (GENxEMO). This suggests that extensive training in group ID-learned was not necessary for learning of new dimensions. In fact, the results from Experiment 2a suggested that extensive training in group ID-learned *impaired* performance in the test compared to a group without extensive training. However, Experiment 2b suggested that this impairment is only observed when such extensive training involves a consistent mapping of categories to motor responses, producing automatic responding. In terms of learning of representations that foster rule learning, extensive experience with the categories does not seem to improve transfer performance beyond that observed after brief experience.

Experiment 3

Morphed face stimuli lack the separable-dimension structure required for learning of unidimensional rules and are instead better described as integral dimensions (Blunden et al., 2015; Goldstone & Steyvers, 2001; Soto & Ashby, 2015), each comprising a variety of shape changes that must be integrated at a pre-decisional stage during categorization (Ashby & Gott, 1988). That is, although participants in the ID-new condition from our previous experiments had to learn a "unidimensional" rule (see Figure 2), the category-relevant and category-irrelevant dimensions are not initially perceived as separate dimensions that can be selectively attended. Rather, information from a variety of face features must be integrated in order to master the task. Because this is

technically an information-integration task, we expected that participants without experience with morphed face dimensions would approach a novel categorization task using a procedural learning strategy[1]. On the other hand, previous research suggests that extensive categorization training produces a separable-dimension structure in morphed face stimuli (Blunden et al., 2015; Goldstone & Steyvers, 2001; Soto & Ashby, 2015). For this reason, we expected that participants exposed to such extensive categorization training would show performance in transfer tests indicative of rule learning. Surprisingly, participants with or without extensive categorization pretraining showed performance indicative of rule learning in the previous experiments. These results suggest that learning of representations that foster the use of rule-based learning happens "on the fly," when people are first exposed to a categorization task involving novel morphed dimensions.

This hypothesis is post-hoc, and we only have negative evidence supporting it, in the form of no significant differences between groups ID-learned and ID-new in Experiments 1 and 2b. In the present experiment, we sought to obtain positive evidence for this hypothesis. One way to do this would be to give limited pretraining in a categorization task to one group of participants (group Pretrain) and then have this group, and a Control group without any prior experience, learn a categorization task thought to require the executive functions that implement rule-based category learning. If a single pre-training session is enough to produce representations that support rule-based learning, then the second complex categorization task should be solved much faster by group Pretrain than by the Control group.

In previous experiments, a single one-hour session seemed to be enough experience for group ID-new to show performance indicative of rule learning, so this is the experience that group Pretrain received here. Both groups were tested using a delayed-response yes-no task (DRYN), in which stimuli were briefly presented and followed by a 5-second retention interval; after this, participants were asked whether the face belonged to one of two possible clubs, and were prompted to respond "Yes" or "No". There are several reasons to believe that such a task would require a rule learning system rather than a procedural learning system. First, the procedural learning system is thought to operate by learning of simple associations between stimuli and response locations (Ashby et al., 2003) or labels (verbal or visual; see Spiering & Ashby, 2008). In the yes-no task, a consistent mapping between stimuli and responses is lost, and therefore such tasks are poorly learned by the procedural system (Maddox et al., 2004b). While the mapping between stimuli and category labels is still consistent, the task requires evaluating the stimulus' category label and match it against the label included in the yes-no question, a process that arguably requires executive function (Spiering & Ashby, 2008). Second, prompting a categorization response from participants only after a retention interval added a working memory requirement to the task. Working memory is thought to be an important mechanism behind rule-based category learning, but it seems to play a relatively minor role in procedural category learning. Experimental studies manipulating working memory though secondary tasks (either concurrent: Miles & Minda, 2011; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006; or sequential: Maddox et al., 2004a; Zeithamova & Maddox, 2007) have consistently shown that learning of rule-based categorization tasks is more strongly impaired by working memory load than learning of information-integration categorization tasks. [2]. While concurrent working memory tasks can impair learning of information-integration categorization tasks (Miles & Minda, 2011; Zeithamova & Maddox, 2006), this effect is weaker and less apparent than that observed in rule-based tasks (Maddox et al. 2004a; Miles & Minda 2011; Waldron & Ashby 2001; Zeithamova & Maddox 2006, 2007) and sometimes simply absent (Miles & Minda, 2011; Zeithamova & Maddox, 2007). In sum, our DRYN task has the double quality of being difficult to solve via procedural learning and requiring some of the executive processes that are a key part of rule-based learning, and for this reason we assumed that it would be solved through rule-based category learning. If representations that foster rule learning are learned quickly (within a single session), then such representations would be available to group Pretrain and not to the Control group, facilitating fast rule-based learning in the former but not in the latter.

Materials and Methods

Participants. 51 undergraduates at the University of California Santa Barbara voluntarily participated in this experiment in exchange for class credit or a mon-

[1]Technically, both the ID-new group and the GENxEMO group were exposed to information-integration tasks. However, the GENxEMO task involves integration of information from a pair of clearly differentiated dimensions (i.e., that can be selectively attended). On the other hand, the ID-new task involves integration of information from a variety of unknown dimensions, which may or may not be separable. Thus, the demands from the two tasks are different, but neither can be solved by an explicit rule based on selective attention to a known dimension. That is, both require a procedural strategy to be learned.

[2]Note, however, that *correlational studies* that have measured individual differences in working memory capacity in general do not support an association between such differences and performance in a specific task (Kalish et al., 2017; Lewandowsky et al., 2012).

etary compensation. There were 24 participants in group Pretrain and 27 participants in group Control.

Stimuli. The stimuli used during categorization pre-training were the same as those used for group ID-learned in Experiment 1 (see Figure 2). The stimuli used during the main DRYN categorization task were created just as described for the experimental groups of Experiment 1. However, the number of stimuli and their coordinates were modified to make the task easier to learn. The coordinates of the stimuli used are shown in Figure 6. As can be seen from the figure, the number of stimuli was reduced from 36 per category to 16 per category, and only stimuli far from the main category boundary were shown.

Procedure. Two groups of participants were included in this experiment: group Pretrain and group Control. Participants in group Pretrain were exposed to a single session of categorization training, about 40-50 minutes long, using the exact same stimuli and task as those described for categorization pre-training of group ID-learned in Experiment 1 (see Figure 2). Participants in group Control did not receive such categorization experience.

Both groups were tested using the DRYN task. In each trial, a stimulus was presented for 2s and followed by a 5s delay, after which participants were presented with one of two possible questions: "Does this person belong to the GREEN club?" or "Does this person belong to the YELLOW club?". The question was displayed by itself, centered in the screen, with all words in white text except for "GREEN club", which were displayed in green color, and "YELLOW club", where were displayed in yellow color. The question displayed was randomly chosen for each trial. The keys "Y" and "B" in the keyboard were re-labeled "Y" and "N", respectively. Participants were instructed to use those keys to respond "Yes" or "No" to the question displayed in each trial. After a key press, the participant received feedback about the correct response. We expected this task to be considerably more difficult to learn than the categorization tasks used in previous experiments, so the task was made easier by including a smaller number of stimuli (16 per category, instead of 36) far from the category boundary (see Figure 6).

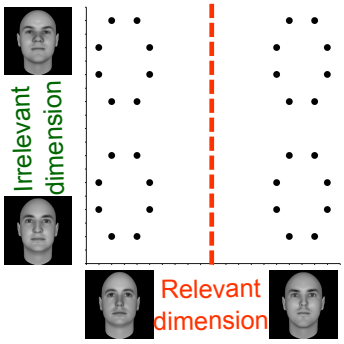

Figure 6. Schematic representation of the stimuli used during categorization training in Experiment 3. The faces shown next to each dimension represent the parents for that specific dimension. The points inside the coordinate system represent stimuli obtained from a specific combination of levels for each dimension. The dotted line represents the category boundary used for training. For more details on the stimuli and task used, see the main text.

Data Analysis. All analyses were performed using R v. 3.2.1 (R Core R Core Team, 2015) extended with the packages *reshape2* v. 1.4.1, *plyr* v. 1.8.3, and *ggplot2* v. 2.1.0, running in RStudio v. 0.99.486 (R Studio Team, 2015). A rejection criterion of $\alpha = 0.05$ was used in all statistical tests.

Participants were excluded from all analyses if they did not reach a performance level of 60% correct by the last block of categorization training. This criterion was set before performing the main analyses.

The main analysis was performed on forward learning curves. Proportion of correct responses was calculated for each participant in each training block, and these values were entered to a 2 (Group: Control or Pretrain) ×7 (Block number) mixed-design ANOVA. No follow-up tests were planned, as the main effects and interaction from the ANOVA would provide all the information necessary to test our main hypothesis.

Backward learning curves were created as described in Experiment 1.

Results and Discussion

The number of participants excluded from the final analysis due to poor performance (<60%) in the categorization task were: 5 from group Pretrain (19 included) and 8 from group Control (19 included).

Figure 7 shows the main results of Experiment 3, with forward learning curves shown in panel a and backward learning curves shown in panel b. From 7a we see that a single session of categorization pre-training in group Pretrain improved performance in the task considerably when compared to that of group Control. Group Pretrain shows extremely fast learning, with high performance around 80% correct from the first block that reaches an asymptotic level close to 90% correct in the second block. In comparison, group Control learned more slowly, reaching asymptotic performance around block four and never quite reaching performance levels as high as those observed in group Pretrain. In line with these observations, the ANOVA revealed both a significant main effect of Group, $F(1, 36) = 5.75$, $p < 0.05$, and a significant interaction between Group and Block, $F(6, 216) = 4.19$, $p < 0.001$. The main effect of Block was also significant,

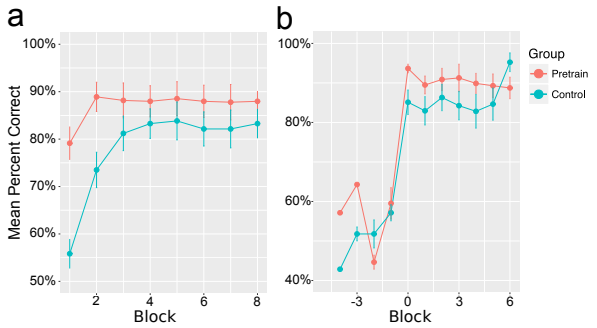

Figure 7. Results of Experiment 3. (a) Forward learning curves depicting mean percent correct during each block of the categorization task; (b) backward learning curves (see text for description) depicting mean percent correct during each block of the categorization task, where block numbers have been individually shifted so that zero represents the first block in which a participant reaches 70% correct.

$F(6, 216) = 14.8$, $p<0.001$, reflecting the fact that performance improved across blocks in both groups.

In sum, the results from the present experiment show that a single pre-training session is enough to speed-up learning and improve performance in a categorization task thought to require executive function. These results are in line with our hypothesis that learning of representations that foster the use of rule-based learning mechanisms happens relatively quickly ("on the fly"), upon first exposure to a categorization task involving morphed face dimensions.

The results with backward learning curves, shown in Figure 7b, suggest step-like learning in both groups. Both groups showed a large jump in performance at block zero, followed by performance that remained stable at asymptotic levels afterwards. Such step-like backward learning curves are suggestive of rule learning through hypothesis testing (Smith & Ell, 2015). This confirms that people apparently use rules to learn categorization tasks involving stimuli that vary along integral dimensions, at least for the kind of morphing dimension used here.

General Discussion

Here we explored the question of whether newly-learned dimensions support the kind of rule-based category learning commonly observed with traditional separable dimensions. Previous research shows that novel morphed dimensions are integral (Blunden et al., 2015; Goldstone & Steyvers, 2001; Soto & Ashby, 2015), but extensive categorization training with stimuli varying in such dimensions makes them more psychologically privileged (Folstein et al., 2012; Goldstone & Steyvers, 2001) and increases their separability (Soto & Ashby, 2015). For these reasons, we expected that people who are completely naive to the dimensions would show no evidence of rule-based category learning, but people who had extensive pre-training in a categorization task would show evidence of rule-based category learning in a new categorization task using the same relevant dimension. Against our expectations, the overall pattern of results from our experiments suggests that people without any experience with a set of morphed dimensions learn new categorization tasks involving such dimensions using a rule-based learning strategy. People with and without categorization pre-training showed similar levels of analogical transfer (Experiment 1) and button-switch interference (Experiments 2 and 2b) after training in a new categorization task. The behavior of both groups was similar to that observed from control groups trained in a uni-dimensional task aligned to familiar face dimensions (who are thought to use a rule-based category learning strategy) and dissimilar to that observed from control groups trained with an integration-information task (who are thought to use a procedural category learning strategy). In addition, people with and without categorization pre-training showed step-like learning of a complex task requiring executive function, suggesting rule learning.

These results suggest that the prior existence of psychologically privileged and separable dimensions at the outset of categorization training is not a requirement for rule-based learning. Instead, it seems as if people have a predisposition to learn categorization tasks using rule-based strategies, even when stimuli are not represented in a way that would facilitate such learning. The best current explanation for the overall pattern of results is that the appropriate representations for rule application are learned on-the-fly during categorization tasks. That is, our results do not necessarily indicate that rules can be learned without the need for dimensional structure in the stimuli. Instead, they suggest that representations that support rule-based category learning, which are likely to be new differentiated dimensions, are quickly learned and immediately used for performance of a categorization task involving novel morph dimensions.

There is some evidence in the previous literature that is in line with fast learning of new dimensions. Although most previous studies in this area involved training that was substantially longer than that given here to group ID-new (e.g., 25 repetitions in Experiment 1 of Goldstone & Steyvers, 2001; 27 repetitions in Soto & Ashby, 2015; 22-24 repetitions in Folstein et al., 2012; 28-56 repetitions in Van Gulick & Gauthier, 2014; 18-75 repetitions in Op de Beeck et al., 2003), Goldstone & Steyvers (2001, Experiments 2a and 3) showed evidence of dimensional differentiation after short training, similar to that received by group ID-new (around

eight stimulus repetitions). However, the Goldstone and Steyvers' experiments tested only dimension differentiation using speed of learning of a new task. In contrast, our Experiments 1 and 2 tested immediate rule transfer in the absence of any new feedback learning. That is, our experiments suggest that *the original category learning* is rule-based, meaning that participants in our ID-new group learned not only to differentiate dimensions after only eight repetitions of each training stimulus, but also used the newly differentiated dimensions for rule-based categorization performance. This is why we emphasize here that representations that support rule-learning seem to be learned "on-the-fly."

In addition, this is the first reported evidence that the dissociations between unidimensional and information-integration categorization tasks, previously well-documented with stimuli varying along simple dimensions–such as spatial frequency and orientation of gratings, or rectangle size and density of dots inside them–can be also found with naturalistic stimuli varying along complex dimensions, such as face gender and emotional expression.

The results from the present study also shed light on the correct interpretation of the various dissociations that have been previously found between categorization tasks that align and do not align with separable dimensions (i.e., "rule-based" and "information-integration" tasks). The overall pattern of results observed in Experiments 1 and 2 (see Figures 3, 4 and 5) suggests that psychologically privileged dimensions (perhaps also separable, although empirical evidence is lacking for face gender and expression) do not *facilitate* learning of rules, but rather *prevent* such learning when the categorization task is not aligned with them. Compared to other groups, the group exposed to an information-integration task involving familiar dimensions (GENxEMO) showed both weaker analogical transfer (see Figure 3c) and a stronger button-switch interference (see Figure 4c) than other groups. Procedural learning of information-integration tasks might be deployed as a way to use a currently-available set of dimensions to represent stimuli in a categorization task, rather than extracting a completely new dimension for that specific task. This way, procedural learning might work as a way to protect stimulus representations that have proven useful in the past, by avoiding interference between them and new category knowledge. As procedural learning does not involve any change in stimulus representation (only stimulus-response associations), it is an adequate strategy to avoid such interference.

The present results, together with a number of previous results, suggest that people have a strong predisposition towards learning new categories through the discovery and application of rules. Participants tend to use a rule-based strategy early in learning even in a task in which such a strategy is incorrect (e.g., Markman et al., 2006). In unsupervised categorization experiments with separable dimensions, people use mostly unidimensional rules (Ashby et al., 1999; Handel & Imai, 1972; Handel et al., 1980; Medin et al., 1987). Here we have found that not only is this predisposition very strong during categorization of stimuli varying along known dimensions, but is also applied when stimuli are completely novel, in the sense that no previous dimensions exist allowing the explicit proposal of rules.

What use are separable dimensions?

Rational theories of generalization suggest that separable dimensions are directions in stimulus space along which natural categories (i.e., "consequential regions") vary (Shepard 1987; Soto et al. 2014, 2015). In line with this idea, people use them preferentially and spontaneously for categorization (Ashby et al., 1999; Handel & Imai, 1972; Handel et al., 1980; Markman et al., 2006; Medin et al., 1987) and learn to extract them during categorization training with novel stimuli (Soto & Ashby, 2015). However, here we found that performance in tests of generalization (analogical transfer and button-switch interference) is similar when people are trained using stimuli with or without a dimensional structure. This opens the questions: What are separable dimensions good for? Why does the human brain learn to extract them in categorization tasks, and why are previously-available separable dimensions privileged during categorization tasks?

Our results suggest that known separable dimensions do not facilitate generalization compared to completely novel dimensions. On the other hand, they do seem to facilitate learning. In all our experiments, we found evidence that previous experience in a categorization task speeds new learning and leads to stronger performance. In most cases, asymptotic performance in a new categorization task was achieved within a single block (i.e., a single presentation of each unique stimulus) by participants who had exposure to categorization pre-training. They showed faster learning and higher performance than participants without such exposure.

Not all integral dimensions are created equal

It is important to note that the results presented here should not be expected to hold with all integral dimensions. In particular, they are likely to *not* hold with some traditional integral dimensions, such as brightness and saturation.

As highlighted by Soto et al. (2015), there is only one way in which two dimensions can be separable–when they are preferred directions in stimulus space that can

be selectively attended–but there are at least two ways in which two dimensions can be integral. According to the correlation hypothesis (Shepard, 1987; 1991), integral dimensions are those with values that correlate in natural classes. For example, the length and width of animals are correlated; longer mammals tend to also be wider. This requires integral dimensions to be privileged directions in stimulus space, with integrality arising from the way in which natural categories vary along such dimensions. According to the direction hypothesis (Austerweil & Griffiths, 2010; Soto et al., 2015), two dimensions are integral when natural classes are equally likely to extend in any direction in space, and therefore there are no privileged directions in stimulus space.

Although Soto et al. (2015) found evidence in line with the direction hypothesis, Kemler-Nelson (1993; see also Jones & Goldstone, 2013) concluded from a literature review that traditional integral dimensions are real psychological dimensions, despite usually being processed in a holistic way. To integrate both sets of findings, Soto et al. (2015) proposed a rational bayesian model allowing intermediate modes of processing between purely separable dimensions, in which attention can be only aligned to the dimensional axes, and purely integral dimensions, in which attention can be deployed along any arbitrary dimension in space. In this model, different directions in space are treated as hypotheses that can be weighted more or less during categorization and generalization tasks, thus representing degrees to which such directions are privileged.

Morphing dimensions seem to lie at the extreme of integrality, where all directions are weighted equally and therefore none is privileged. These dimensions not only interact during processing (Goldstone & Steyvers, 2001; Soto & Ashby, 2015), but also they are not privileged directions in stimulus space (Folstein et al., 2012; Goldstone & Steyvers, 2001). In addition, the fact that no direction is privileged explains why any direction can acquire such status through categorization training (Folstein et al., 2012; Goldstone & Steyvers, 2001). We expect that the effects of categorization training usually found with morphing dimensions, such as those presented here, are likely to hold with any other dimensions that show such extreme integrality.

On the other hand, traditional integral dimensions like brightness and saturation seem to lie somewhere between the two extremes of integrality and separability. One possibility is that the two integral dimensions are slightly more privileged than other directions (i.e., weighted more heavily), but all directions are weighted to some extent. This is in line with the hypothesis put forward by Smith & Kemler (1978) that integral dimensions are perceived holistically but also sustain a less preferred mode of processing in terms of component parts. This explains why integral dimensions seem to interact with one another during processing, but also show evidence of being privileged directions in stimulus space (Foard & Kemler-Nelson, 1984; Grau & Kemler-Nelson, 1988; Jones & Goldstone, 2013; Kemler-Nelson, 1993; Melara et al., 1993).

Another possibility is that dimensions might appear integral because privileged directions in stimulus space do exist, but they do not align with the dimensional axes (i.e., other directions are weighted most heavily) or they align with only one of them (i.e., only one of the axes is weighted most heavily, as in the "dominance metric", see Soto & Wasserman, 2010b). For example, Ell et al. (2012) studied unsupervised categorization with the integral dimensions of brightness and saturation, and they found that some participants categorized stimuli as if they could pay selective attention only to brightness, but not to saturation, and others behaved as if they were extracting a diagonal "grayness" dimension, going from dim and saturated stimuli to bright and desaturated stimuli. In line with the idea of a heavily-weighted diagonal dimension, categorization training using brightness and saturation fails to produce evidence of dimension differentiation (Goldstone, 1994b).

All these and more possibilities can be accommodated by the rational bayesian model of Soto et al. (2015; see also Soto et al., 2014), implemented as different patterns of pre-existent preferences for directions in stimulus space. More research will be necessary to fully characterize how different types of integral dimensions differ from one another and how this might affect learning and generalization in categorization tasks.

What kind of mechanism can account for these results?

Any mechanistic explanation of the results presented here must have the following features: (1) it must involve a mechanism to quickly learn to extract category-relevant information and ignore category-irrelevant information; (2) it must involve a way to quickly re-map responses (to reduce button-switch interference) and labels (to foster fast learning of YN task) to the representation of each category; (3) it must involve a way to facilitate encoding and/or maintenance of category representation in visual working memory. Several computational mechanisms in the literature have those features. Here we focus on neurocomputational mechanisms, which include both algorithmic and implementational details and thus generate more predictions to differentiate them in future research (Ashby & Helie, 2011).

One possibility involves the learning of novel category representations, which would be intermediate between visual representations and motor choices. If such representations are accessible to executive processes–allowing

retention in working memory, selective attention, and fast mapping to responses and labels–then all the results observed here can be explained. Importantly, behavioral evidence suggests that learning of intermediate representations is possible in both procedural and rule-based systems (Maddox et al., 2010), and neurocomputational models exist that implement both mechanisms.

Regarding learning in the procedural system, Cantwell et al. (2015) proposed a two-stage model of procedural categorization capable of learning intermediate category representations in the striatum through error-driven learning mechanisms. In this model, initial category learning is implemented in the visual cortico-striatal loop, which learns unified representations for similar groups of stimuli that have been assigned to the same category. These unified category representations are passed to the pre-SMA, which then feeds them to the motor cortico-striatal loop that learns to associate them with responses. The results from the analogical transfer test cannot be explained by this model in its original form, because learning in the caudate involves associations between specific stimuli (i.e., individual faces) and group membership. Presentation of new stimuli during the generalization test would not activate the learned intermediate representations. However, a small modification of the model would allow better generalization. If one assumes that visual cortical neurons represent each stimulus as a distributed pattern, with some neurons being activated by many stimuli in the same category (e.g., a neuron representing a facial feature shared by many members of the category), then an error-driven learning rule like that used by Cantwell et al. (2015) would selectively link such relatively category-specific neurons to intermediate representations (Soto et al., 2012; Soto & Wasserman, 2010a,1). Because new stimuli would also activate such category-specific visual neurons, they would in turn activate the learned intermediate representations. The specific settings used by Cantwell et al. (2015) in their simulations also seem unable to account for the results of Experiment 2, as their two-stage model showed a strong button-switch interference effect. Although the newly learned representations in this model can be quickly associated with new motor responses, it still takes some time for the model to learn the reversed assignment of categories to responses. However, it is possible that parameter settings different from those used by Cantwell et al. (2015) can reproduce the results of Experiment 2. Similarly, the results of Experiment 3 are also difficult to explain with this model, unless one assumes that the newly-learned representations transferred to pre-SMA can be flexibly applied to novel tasks and kept in working memory through the persistent activity typical of neurons in the PFC (for a review, see Riley & Constantinidis, 2016). In summary, the Cantwell et al. (2015) model is a candidate for learning of intermediate representations, but explaining the results found here would require additional assumptions.

Regarding learning of intermediate representations in the rule-based system, at least two possibilities exist. One possibility is that novel category representations are kept and updated in the hippocampus, under the influence of task demands and goals implemented in the prefrontal cortex, as proposed by Love and colleagues in their SUSTAIN model (Love & Gureckis, 2007; Love et al., 2004; Mack et al., 2016). In this model, the hippocampus keeps representations of stimulus clusters, and updates such representations as the result of surprising events (Love & Gureckis, 2007) and task demands (Mack et al., 2016). If a task requires more attention to a particular stimulus dimension, the PFC directs attention to that dimension and the hippocampal cluster representations are updated accordingly (Mack et al., 2016). One feature of this theory is that it assumes that updating can be done relatively quickly, given the known role of the hippocampus in fast learning. SUSTAIN seems able to explain all the results reported here. The theory can explain the results of Experiment 1 as the result of allocating attention to properties of stimuli that are relevant to the categorization task. Fast learning of clusters during categorization training can explain the lack of a button-switch interference effect in the groups with limited (ID-new) or extensive (ID-learned) categorization experience during Experiment 2, and the advantage of pre-training on learning of a rule-based categorization task in Experiment 3.

A second possibility is that novel category representations are kept and updated in lateral PFC, as suggested by a wealth of results from monkey electrophysiology experiments (e.g., Cromer et al. 2010; Freedman et al. 2003; Roy et al. 2010, 2014). Unfortunately, the neurocomputational mechanisms that guide learning of category representations in the lateral PFC are not well understood. To the best of our knowledge, no working computational model for this process has been proposed yet. In monkeys, categorical representations are observed only after extensive training, and it is believed that they form with the help from slow reward-driven learning processes that take place in the basal ganglia (Buschman & Miller, 2014). Such slow learning does not seem compatible with the fast learning observed in our experiments.

Engel et al. (2015) proposed a model in which intermediate representations are not learned *de novo*, but previously-available intermediate representations, in the form of neurons that are sensitive both to certain stimulus features and response choices, are selected and sharpened through learning driven by reward prediction er-

rors. In this model, previous neural selectivity serves as a "scaffold" for learning of categorical representations. Reward learning influences cortical representations directly through assumed dopaminergic neuromodulation of visual cortical neurons, which serves to modify the tuning of individual neurons in a way that facilitates the discrimination of categories. Thus, this model is somewhere between learning of *de novo* intermediate category representations, and enhancement of already existing representations.

Finally, it is possible that no intermediate representations are learned at all. For example, we could assume that an attentional mechanism is able to bias visual representations themselves, so that only category-relevant information is preferentially processed. A variety of mechanisms could achieve this. For example, attention could be biased towards visual features that predict positive feedback or reward (Anderson, 2016). Neurophysiological (Yamamoto et al., 2013) and neuroimaging (Anderson et al., 2014) data suggest that the tail of the caudate could be the site where learning of such reward-driven attentional biases is implemented. A recently-proposed neurocomputational model (Hays & Soto, 2017) suggests that learning of associations between visual representations and rewards in the caudate may influence those same visual representations via closed loops involving visual cortex and the basal ganglia. An important difference between this model and that of Engel et al. (2015) is the way in which reward-driven learning influences cortical representations. In the Hays and Soto model, reward learning influences cortical representations only indirectly through the output of cortico-striatal loops, which serves only to enhance already-existing visual representations.

Because this learning mechanism involves only enhanced processing of already-existing representations, it does not require slow training. The resulting attentional biases would allow selective processing of only visual representations that are informative for the categorization task (Soto & Wasserman, 2010a), which explains good analogical transfer. Attentional enhancement could also speed up the learning processes involved in reassignment of responses during the button switch interference test, and fast learning of the yes-no conditional discrimination task. However, it is not clear whether attentional mechanisms could explain the fact that in both of these tasks performance is essentially at ceiling within the first block. As with the Cantwell et al. (2015) model, an explanation based only on attentional learning might require the additional assumption that attention facilitates the use of other executive functions (e.g., encoding into visual working memory).

The data available related to these different hypotheses is limited, and it does not strongly favor one of them over the other. For example, an fMRI study on the neural correlates of dimension differentiation during categorization found changes in the representation of the category-relevant dimension in early visual cortex, in several PFC areas and in hippocampus (Folstein et al., 2013). Only further behavioral and neurobiological research can distinguish among these possibilities. Still, at this time SUSTAIN (Love & Gureckis, 2007; Love et al., 2004; Mack et al., 2016) seems better equipped than alternatives to explain the overall pattern of results observed here without substantial additions or modifications.

Limitations of this study

Morphed face dimensions can be shown to be integral by a variety of tests (Blunden et al., 2015; Goldstone & Steyvers, 2001; Soto & Ashby, 2015), and extensive training increases dimensional separability (Soto & Ashby, 2015). Because our current study did not include tests of dimensional separability, it is not clear whether the fast learning found here for groups without such extensive pre-training is accompanied by separability learning. It is possible that the representations learned in the present experiments to support rule-based categorization are not separable. It is thus unclear for now whether dimensional separability is a condition for rule-based learning.

The present experiments used faces as stimuli, which are particularly important objects for humans. Perhaps the adaptive importance of faces "prepares" people and other primates for fast learning of category representations that can support rule learning. Testing the generality of our results will require replicating them using morphed objects other than faces (e.g., cars or novel shapes: Folstein et al. 2012; Op de Beeck et al. 2003).

Finally, as explained in the previous section, there are many possible mechanistic explanations for the results obtained here, and our data do not allow us to discriminate among them. This task cannot be accomplished without substantial additional behavioral and neuroscientific research.

Conclusion

Here we show evidence that newly-learned dimensions support the kind of rule-based category learning commonly observed with traditional separable dimensions. In addition, we found that the prior existence of psychologically privileged and separable dimensions at the outset of categorization training is not a requirement for rule-based learning. Rule-based categorization performance was not only found with stimuli having a prior dimensional structure (i.e., face gender and emotion) or that acquired such structure through extensive categorization training. Rather, representations that support the

use of rule-based categorization seemed to be learned on-the-fly during brief categorization training with stimuli that lacked any previous dimensional structure. This confirms that people have a strong predisposition towards learning new categories through the discovery and application of rules, and that rule-based category learning is a powerful adaptive system that is not limited by the availability of separable dimensions for its application.

References

Anderson, B. A. (2016). The attention habit: How reward learning shapes attentional selection. *Annals of the New York Academy of Sciences*, 1369(1), 24–39.

Anderson, B. A., Laurent, P. A., & Yantis, S. (2014). Value-driven attentional priority signals in human basal ganglia and visual cortex. *Brain Research*, 1587, 88–96.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, 105(3), 442–481.

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, 31(7), 1114–1125.

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.

Ashby, F. G. & Helie, S. (2011). A tutorial on computational cognitive neuroscience: Modeling the neurodynamics of cognition. *Journal of Mathematical Psychology*, 55(4), 273–289.

Ashby, F. G., Paul, E., & Maddox, W. T. (2011). COVIS. In E. M. Pothos & A. J. Wills (Eds.), *Formal Approaches in Categorization* (pp. 65–87). New York, NY: Cambridge University Press.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Attention, Perception, & Psychophysics*, 61(6), 1178–1199.

Ashby, F. G. & Soto, F. A. (2015). Multidimensional signal detection theory. In J. Busemeyer, J. T. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford Handbook of Computational and Mathematical Psychology* (pp. 13–34). New York, NY: Oxford University Press.

Ashby, F. G. & Valentin, V. V. (2005). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Categorization in Cognitive Science* (pp. 548–572). New York: Elsevier.

Ashby, F. G. & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, 6(3), 363–378.

Austerweil, J. L. & Griffiths, T. L. (2010). Learning hypothesis spaces and dimensions through concept learning. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Blunden, A. G., Wang, T., Griffiths, D. W., & Little, D. R. (2015). Logical-rules and the classification of integral dimensions: individual differences in the processing of arbitrary dimensions. *Front. Psychol*, 5, 1531.

Buschman, T. J. & Miller, E. K. (2014). Goal-direction and top-down control. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655), 20130471.

Cantwell, G., Crossley, M. J., & Ashby, F. G. (2015). Multiple stages of learning in perceptual categorization: evidence and neurocomputational theory. *Psychon Bull Rev*, 22(6), 1598–1613.

Casale, M. B., Roeder, J. L., & Ashby, F. G. (2012). Analogical transfer in perceptual categorization. *Memory & Cognition*, 40, 434–449.

Collins, J. A. & Olson, I. R. (2014). Knowledge is power: How conceptual knowledge transforms visual cognition. *Psychon Bull Rev*, 21(4), 843–860.

Cromer, J., Roy, J. E., & Miller, E. K. (2010). Representation of multiple, independent categories in the primate prefrontal cortex. *Neuron*, 66(5), 796–807.

Dosher, B. & Lu, Z. L. (2017). Visual perceptual learning and models. *Annual Review of Vision Science*, 3, 343–363.

Ekman, P., Friesen, W. V., & Hager, J. (1978). *The Facial Action Coding System (FACS): A technique for the measurement of facial action*. Palo Alto, CA: Consulting Psychologists Press.

Ell, S. W., Ashby, F. G., & Hutchinson, S. (2012). Unsupervised category learning with integral-dimension stimuli. *The Quarterly Journal of Experimental Psychology*, 65(8), 1537–1562.

Engel, T. A., Chaisangmongkon, W., Freedman, D. J., & Wang, X. J. (2015). Choice-correlated activity fluctuations underlie learning of neuronal category representation. *Nature Communications*, 6, 6454.

Ester, E. F., Sprague, T. C., & Serences, J. T. (2017). Category learning biases sensory representations in human visual cortex. *bioRxiv*, (pp. 170845).

Foard, C. F. & Kemler-Nelson, D. G. (1984). Holistic and analytic modes of processing: The multiple determinants of perceptual analysis. *Journal of Experimental Psychology: General*, 113(1), 94–111.

Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2012). Not all morph spaces stretch alike: How category learning affects object discrimination. *Jour-*

nal of Experimental Psychology: Learning, Memory, and Cognition, 38(4), 807–802.

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, 23(4), 814–823.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, 23(12), 5235–5246.

Garner, W. R. (1974). *The processing of information and structure*. New York: Lawrence Erlbaum Associates.

Goldstone, R. L. (1994a). An efficient method for obtaining similarity data. *Behavior Research Methods*, 26(4), 381–386.

Goldstone, R. L. (1994b). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123(2), 178–200.

Goldstone, R. L., de Leeuw, J. R., & Landy, D. H. (2015). Fitting perception in and to cognition. *Cognition*, 135, 24–29.

Goldstone, R. L., Gerganov, A., Landy, D., & Roberts, M. E. (2009). Learning to see and conceive. In L. Tommasi, M. A. Peterson, & L. Nadel (Eds.), *Cognitive biology: Evolutionary and developmental perspectives on mind, brain, and behavior* (pp. 163). Cambridge, MA: MIT Press.

Goldstone, R. L. & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1), 69–78.

Goldstone, R. L. & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, 130(1), 116.

Grau, J. W. & Kemler-Nelson, D. G. (1988). The distinction between integral and separable dimensions: Evidence for the integrality of pitch and loudness. *Journal of Experimental Psychology: General*, 117(4), 347–370.

Handel, S. & Imai, S. (1972). The free classification of analyzable and unanalyzable stimuli. *Perception & Psychophysics*, 12(1), 108–116.

Handel, S., Imai, S., & Spottswood, P. (1980). Dimensional, similarity, and configurai classification of integral and separable stimuli. *Perception & Psychophysics*, 28(3), 205–212.

Hays, J. & Soto, F. A. (2017). Modeling the mechanisms of reward learning that bias visual attention. *Journal of Vision*, 17(10), 1302–1302.

Hélie, S., Ell, S. W., Filoteo, J. V., & Maddox, W. T. (2015). Criterion learning in rule-based categorization: Simulation of neural mechanism and new data. *Brain and Cognition*, 95, 19–34.

Helie, S., Waldschmidt, J. G., & Ashby, F. G. (2010). Automaticity in rule-based and information-integration categorization. *Attention, Perception, & Psychophysics*, 72(4), 1013–1031.

Jones, M. & Goldstone, R. L. (2013). The structure of integral dimensions: Contrasting topological and Cartesian representations. *Journal of Experimental Psychology: Human Perception and Performance*, 39(1), 111–132.

Kalish, M. L., Newell, B. R., & Dunn, J. C. (2017). More is generally better: Higher working memory capacity does not impair perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(4), 503–514.

Kemler-Nelson, D. G. (1993). Processing integral dimensions: The whole view. *Journal of Experimental Psychology: Human Perception and Performance*, 19(5), 1105–1113.

Lewandowsky, S., Yang, L. X., Newell, B. R., & Kalish, M. L. (2012). Working memory does not dissociate between different perceptual categorization tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(4), 881–904.

Livesey, E. & McLaren, I. (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.

Love, B. C. & Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective, & Behavioral Neuroscience*, 7(2), 90.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111(2), 309–332.

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *PNAS*, 113(46), 13203–13208.

Maddox, W., Lauritzen, J., & Ing, A. (2007). Cognitive complexity effects in perceptual classification are dissociable. *Memory & Cognition*, 35(5), 885–894.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(4), 650.

Maddox, W. T., Ashby, F. G., Ing, A. D., & Pickering, A. D. (2004a). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition*, 32(4), 582–591.

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004b). Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic Bulletin & Review*, 11(5), 945–952.

Maddox, W. T., Glass, B. D., O'Brien, J. B., Filoteo, J. V., & Ashby, F. G. (2010). Category label and response location shifts in category learning. *Psychological Research*, 74(2), 219–236.

Maddox, W. T. & Ing, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, 31(1), 100–107.

Markman, A. B., Maddox, W. T., & Worthy, D. A. (2006). Choking and Excelling Under Pressure. *Psychol Sci*, 17(11), 944–948.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, 19(2), 242–279.

Melara, R. D., Marks, L. E., & Potts, B. C. (1993). Primacy of dimensions in color perception. *Journal of Experimental Psychology: Human Perception and Performance*, 19(5), 1082–1104.

Miles, S. J. & Minda, J. P. (2011). The effects of concurrent verbal and visual tasks on category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(3), 588–607.

Natal, S. D. C., McLaren, I. P. L., & Livesey, E. J. (2013). Generalization of feature- and rule-based learning in the categorization of dimensional stimuli: Evidence for dual processes under cognitive control. *Journal of Experimental Psychology: Animal Behavior Processes*, 39(2), 140–151.

Nosofsky, R., Stanton, R., & Zaki, S. (2005). Procedural interference in perceptual classification: Implicit learning or cognitive complexity? *Memory & Cognition*, 33(7), 1256–1271.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *J. Exp. Psychol.-Gen.*, 115(1), 39–57.

Oosterhof, N. N. & Todorov, A. (2008). The functional basis of face evaluation. *PNAS*, 105(32), 11087–11092.

Op de Beeck, H. P., Wagemans, J., & Vogels, R. (2003). The effect of category learning on the representation of shape: Dimensions can be biased but not differentiated. *Journal of Experimental Psychology: General*, 132(4), 491–511.

Perez, O. D., San Martin, R., & Soto, F. A. (2018). Exploring the effect of stimulus similarity on the summation effect in human causal learning. *Experimental Psychology*. status: Advance online publication.

R Core Team (2015). R: A language and environment for statistical computing.

Riley, M. R. & Constantinidis, C. (2016). Role of prefrontal persistent activity in working memory. *Front. Syst. Neurosci.*, 9, 181.

Roesch, E. B., Tamarit, L., Reveret, L., Grandjean, D., Sander, D., & Scherer, K. (2011). FACSGen: A tool to synthesize emotional facial expressions through systematic manipulation of facial action units. *J Nonverbal Behav*, 35(1), 1–16.

Roy, J. E., Buschman, T. J., & Miller, E. K. (2014). PFC Neurons Reflect Categorical Decisions about Ambiguous Stimuli. *Journal of Cognitive Neuroscience*, 26(6), 1283–1291.

Roy, J. E., Riesenhuber, M., Poggio, T., & Miller, E. K. (2010). Prefrontal cortex activity during flexible categorization. *Journal of Neuroscience*, 30(25), 8519–8528.

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

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, 237(4820), 1317–1323.

Shepard, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In J. Pomerantz & G. Lockhead (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner.* (pp. 53–71). Washington, DC: American Psychological Association.

Smith, J. D. (2014). Prototypes, exemplars, and the natural history of categorization. *Psychon Bull Rev*, 21(2), 312–331.

Smith, J. D., Beran, M. J., Crossley, M. J., Boomer, J., & Ashby, F. G. (2010). Implicit and explicit category learning by macaques (Macaca mulatta) and humans (Homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes*, 36(1), 54–65.

Smith, J. D. & Ell, S. W. (2015). One giant leap for categorizers: One small step for categorization theory. *PLOS ONE*, 10(9), e0137334.

Smith, L. B. & Kemler, D. G. (1978). Levels of experienced dimensionality in children and adults. *Cognitive Psychology*, 10(4), 502–532.

Soto, F. A. & Ashby, F. G. (2015). Categorization training increases the perceptual separability of novel dimensions. *Cognition*, 139, 105–129.

Soto, F. A., Gershman, S. J., & Niv, Y. (2014). Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychological Review*,

121(3), 526–558.

Soto, F. A., Quintana, G. R., Pérez-Acosta, A. M., Ponce, F. P., & Vogel, E. H. (2015). Why are some dimensions integral? Testing two hypotheses through causal learning experiments. *Cognition*, 143, 163–177.

Soto, F. A., Siow, J. Y. M., & Wasserman, E. A. (2012). View-invariance learning in object recognition by pigeons depends on error-driven associative learning processes. *Vision Research*, 62, 148–161.

Soto, F. A. & Wasserman, E. A. (2010a). Error-driven learning in visual categorization and object recognition: A common elements model. *Psychological Review*, 117(2), 349–381.

Soto, F. A. & Wasserman, E. A. (2010b). Integrality/separability of stimulus dimensions and multidimensional generalization in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 36(2), 194–205.

Soto, F. A. & Wasserman, E. A. (2010c). Missing the forest for the trees: Object discrimination learning blocks categorization learning. *Psychological Science*, 21(10), 1510–1517.

Soto, F. A., Zheng, E., Fonseca, J., & Ashby, F. G. (2017). Testing separability and independence of perceptual dimensions with general recognition theory: a tutorial and new R package (grtools). *Front. Psychol.*, 8, 696.

Spiering, B. J. & Ashby, F. G. (2008). Response processes in information–integration category learning. *Neurobiology of Learning and Memory*, 90(2), 330–338.

Team, R. (2015). Rstudio: Integrated development environment for r.

Van Gulick, A. E. & Gauthier, I. (2014). The perceptual effects of learning object categories that predict perceptual goals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(5), 1307–1320.

Waldron, E. M. & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin & Review*, 8(1), 168–176.

Yamamoto, S., Kim, H. F., & Hikosaka, O. (2013). Reward value-contingent changes of visual responses in the primate caudate tail associated with a visuomotor skill. *Journal of Neuroscience*, 33(27), 11227–11238.

Zaki, S. R. & Kleinschmidt, D. F. (2014). Procedural memory effects in categorization: Evidence for multiple systems or task complexity? *Mem Cogn*, 42(3), 508–524.

Zeithamova, D. & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, 34(2), 387–398.

Zeithamova, D. & Maddox, W. T. (2007). The role of visuospatial and verbal working memory in perceptual category learning. *Memory & Cognition*, 35(6), 1380–1398.

Compliance with Ethical Standards

The authors (Fabian Soto and F. Gregory Ashby) declare that they have no conflict of interest. This study was funded by NIH grant R01MH063760 and by the US Army Research Office through the Institute for Collaborative Biotechnologies under Grant W911NF-07-1-0072. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study. This article does not contain any studies with animals performed by any of the authors.